

Thinking Clearly in a Data-Driven Age

Ethan Bueno de Mesquita and Anthony Fowler

Harris School, University of Chicago

June 10, 2020

THIS IS A WORKING DRAFT. PLEASE DON'T DISTRIBUTE WITHOUT PERMISSION.

© 2019 Ethan Bueno de Mesquita and Anthony Fowler All Rights Reserved

Dedication Goes Here

Short Contents

Preface	xvi
1 Thinking in a Data Driven Age	1
I Establishing a Common Language	19
2 Correlation: What Is It and What Is It Good For?	20
3 Causation: What Is It and What Is It Good For?	58
II Does a Relationship Exist?	80
4 Correlation Requires Variation	81
5 Regression for Describing and Forecasting	112
6 Inferences about Relationships	145
7 Over-Comparing, Under-Reporting	172

8 Reversion to the Mean	212

III Is the Relationship Causal	238

9 Why Correlation Doesn't Imply Causation	239

10 Controlling for Confounders	293

11 Randomized Experiments	330

12 Regression Discontinuity Design	371

13 Difference-in-Differences	407

14 Assessing Mechanisms	435

IV From Information to Decisions	453

15 Turn Statistics into Substance	454

16 Measure Your Mission	495

17 The Perils and Promise of Quantitativication	531

Contents

Preface	xvi
Acknowledgements	xvi
1 Thinking in a Data Driven Age	1
What You'll Learn	1
Cautionary Tales	3
Abe's hasty diagnosis	3
Civil resistance	6
broken windows policing	9
Thinking and Data Are Complements, Not Substitutes	15
Readings	17
I Establishing a Common Language	19
2 Correlation: What Is It and What Is It Good For?	20
What You'll Learn	20
What is a correlation?	21

Fact or correlation?	27
What is a Correlation Good For?	31
Description	31
Forecasting	33
Causal inference	37
Measuring Correlations	39
Mean and Variance	39
Covariance	43
Correlation coefficient	44
Slope of the regression line	45
Populations and samples	46
Straight Talk about Linearity	47
Key Terms	52
Exercises	54
Notes	56
3 Causation: What Is It and What Is It Good For?	58
What you'll learn	58
What is causation?	59
Potential Outcomes and Counterfactuals	62
What is causation good for?	63
Conceptual issues	64
What is the cause?	64

The fundamental problem of causal inference	68
Causality and counterexamples	70
Causality and the law	72
Causality works in surprising ways	74
Causality can run backwards in time	74
Causality does not require physical connection	75
Causation need not imply correlation	76
Key Terms	77
Exercises	77
Notes	79
 II Does a Relationship Exist?	 80
 4 Correlation Requires Variation	 81
Selecting on the Dependent Variable	85
The 10,000 hour rule	86
Corrupting the youth	89
High school dropouts	93
Suicide attacks	95
The World is Conspiring to Make Us Select on the Dependent Variable	98
Back pain	98
Post-mortems	99
Life advice	107

Key Terms	108
Exercises	108
Notes	109
5 Regression for Describing and Forecasting	112
What you'll learn	112
Regression Basics: Crime and Temperature Revisited	113
Linear Regression, Non-Linear Data: Age and Voter Turnout	120
The Problem of Overfitting: Predicting Election Results	134
How Regression Is Presented	138
A Brief Intellectual History of Regression	139
Key Terms	142
Exercises	143
Notes	144
6 Inferences about Relationships	145
What you'll learn	145
Estimation	146
Why Do Estimates Differ From Estimands	148
Bias	150
Noise	151
What Makes for a Good Estimator?	152
Quantifying Precision	154
Statistical Inference and Hypothesis Testing	157

Confidence intervals	157
Hypothesis testing	159
Statistical significance	160
Statistical Inference about Relationships	161
What If We Have Data for the Whole Population?	164
Substantive vs. Statistical Significance	165
Key Terms	168
Exercises	170
Notes	171
7 Over-Comparing, Under-Reporting	172
What you'll learn	172
The Problem of Over-Comparing and Under-Reporting	180
p-hacking	183
p-screening	185
Are Most Scientific "Facts" False?	187
ESP	187
Get out the vote	190
p-hacking forensics	191
Potential Solutions	195
Change the significance threshold	195
Adjust p-values for multiple testing	196
Pre-registration	196

Replication	198
Test important, rather than cute, hypotheses	201
Don't obsess over statistical significance	202
Beyond Science	202
Superstars	203
Key Terms	207
Exercises	208
Notes	210
8 Reversion to the Mean	212
What you'll learn	212
Francis Galton and Regression to Mediocrity	214
Mean reversion is not a gravitational force	219
Seeking Help	224
Mean Reversion, the Placebo Effect, and Cosmic Habituation	227
The placebo effect	227
Cosmic habituation, redux	229
Where should we <i>not</i> expect reversion to the mean?	233
Key Words	235
Exercises	236
Notes	236

III Is the Relationship Causal	238
9 Why Correlation Doesn't Imply Causation	239
What You'll Learn	239
The Basic Problem	247
Sources of Bias	254
Confounders	254
Reverse causality	256
The 10,000 hour rule, revisited	258
Diet soda	263
How Different are Confounders and Reverse Causality?	264
Campaign spending	265
Signing the Bias	268
Contraception and HIV	273
Mechanisms vs Confounders	275
Bias and Noise	279
Key Terms	285
Exercises	287
Notes	291
10 Controlling for Confounders	293
What You'll Learn	293
How Does Regression Control?	300
Controlling and Causation	315

Reading a Regression Table	317
The Anatomy of a Regression	321
Controlling for Confounders versus Mechanisms	324
There is No Magic	325
Key Terms	327
Exercises	328
Notes	329
11 Randomized Experiments	330
What You'll Learn	330
Breastfeeding	331
Randomization and Causal Inference	335
Estimation and Inference in Experiments	340
Problems That Can Arise in Experiments (and Other Research Designs)	343
Noncompliance and instrumental variables	343
Chance imbalance	355
Lack of statistical power	358
Attrition	359
Interference	361
Natural Experiments	363
Key Words	366
Exercise	368
Notes	369

12 Regression Discontinuity Design	371
What You'll Learn	371
How to Implement an RD Design	378
Are Extremists or Moderates More Electable?	381
Continuity at the Threshold	385
Does continuity hold in election RD designs	391
Non-Compliance and the Fuzzy RD	393
Motivation and Success	401
Key Words	402
Exercises	403
Notes	406
13 Difference-in-Differences	407
What You'll Learn	407
Parallel Trends	409
2 Units and 2 Periods	412
N Units and 2 Periods	417
N Units and N Periods	423
Useful Diagnostics	424
Compulsory Voting in Australia	426
Is Obesity Contagious?	427
Key Terms	431
Exercises	432

Notes	433
14 Assessing Mechanisms	435
What You'll Learn	435
Causal Mediation Analysis	437
Intermediate Outcomes	439
Independent Theoretical Predictions	441
Testing Mechanisms by Design	444
Disentangling Mechanisms	445
Key Terms	449
Exercises	450
Readings	451
 IV From Information to Decisions	 453
15 Turn Statistics into Substance	454
What You'll Learn	454
What's the right scale?	455
Miles-per-gallon vs. gallons-per-mile	456
Percent vs. percentage point	459
From Statistics to Beliefs: Bayes' Rule	461
Bayes' rule	466
Information, beliefs, priors, and posteriors	467
Abe's celiac revisited	468

Finding terrorists in an airport	473
Bayes' rule and quantitative analysis	478
Expected Costs and Benefits	485
Amniocentesis and age	486
Key Words	489
Exercises	490
Readings	493
16 Measure Your Mission	495
What you'll learn	495
D You Measure the Right Variable?	496
Partial measures	496
Intermediate outcomes	499
Ill-defined missions	503
Do You Have the Right Sample?	507
External validity	508
Selected samples	510
Will Things Change Once You Act?	518
The shift in baseball	520
The war on drugs	522
Key Words	524
Exercises	525
Readings	528

17 The Perils and Promise of Quantitativication

531

Preface

A preface goes here.

Acknowledgements

We should highlight right away that some of this material was developed as a joint effort for an undergraduate class of Anthony's and an executive education class that Jake Shapiro, Liam Collins, Cathy Fetell, and Ethan designed together. Jake, Liam, and Cathy are due lots of credit and have our deep appreciation.

We'd like to thank Andy Hall, Scott Ashworth, Alex Fournaies, Austin Wright, Chris Berry, Jens Ludwig, Jeff Grogger, Kerwin Charles, Bruce Bueno de Mesquita, Pablo Montagnes, Mordecai Magencey, Gautam Nair, Jeff Ruff, Tom Naset, Vanitha Virudachalam, Matt Gabel, Kosuke Imai, Tom Budescu, Andrew Means, Andrew Little, Mike Spagat, Lindsey Cormack, Sephane Wolton, Dustin Tingley, and Matt Brems for providing incredibly helpful feedback.

We'd like to thank our students for catching numerous mistakes and errors in previous drafts. AK Alilonu, Denise Azadeh, Ellie Rutkey, and Al Shah caught a truly embarrassing number of them. Thank you!

Chapter 1

Thinking in a Data Driven Age

What You'll Learn

- Even well-trained people often make crucial errors with data.
- Learning to think clearly and conceptually about quantitative information is important for lots of reasons, even if you have no interest in a career as a data analyst.
- Thinking and data are complements, not substitutes.
- The skills you will learn in this book will help you use evidence to make better decisions in your personal and professional life and be a more thoughtful and well-informed citizen.

We live in a data-driven age. According to former Google CEO Eric Schmidt, the contemporary world creates as much new data every two days as had been created from the beginning of time through the year 2003. All this information is supposed to have the power to improve our lives, but to harness this power we must learn to think clearly about our data-driven world. Clear thinking is hard—especially when mixed up with all the technical details that typically surrounds data and data analysis.

Thinking clearly in a data-driven world is, first and foremost, about staying focused on ideas and questions, rather than technicalities. Unfortunately, the statistics and quantitative reasoning classes in which most people learn about data do exactly the opposite—that is, they focus on technical details. Students learn mathematical formulas, memorize the names of statistical procedures, and start crunching numbers without ever having been asked to think clearly and conceptually about what they are doing or why they are doing it. Such an approach can work for people to whom thinking mathematically comes naturally. But we believe it is counterproductive for the vast majority of us. When technicality pushes students to stop thinking and start memorizing, they miss the forest for the trees. And it's also no fun.

Our focus, by contrast, is on conceptual understanding. What features of the world are you comparing when you analyze data? What questions can different kinds of comparisons answer? Do you have the right question and comparison for the problem you are trying to solve? Why might an answer that sounds convincing actually be misleading?

It isn't that we don't think the technical details are important. But we believe that technique without conceptual understanding or clear thinking is a recipe for disaster. In our view, once you can think clearly about quantitative analysis, and once you understand why asking careful and precise questions is so important, technique will follow naturally. Moreover, this way is way more fun.

In this spirit, we've written this book to require no prior exposure to data analysis,

statistics, or quantitative methods. We've minimized (though certainly not eliminated) technical material in favor of plain English explanation wherever possible. Our hope is that this book will be used as an introduction to how to both think about and do quantitative analysis. We believe anyone can become a sophisticated consumer (and even producer) of quantitative information. It just takes some patience, perseverance, and a firm resolve to never allow technicality to be a substitute for clear thinking.

We are aware that most readers of this book will not become professional quantitative analysts. But we suspect many of you will have quantitative analysts working for or with you. And all of you will read studies, news reports, and briefings in which someone tries to convince you of a conclusion using quantitative analyses. This book will equip you with the clear thinking skills necessary to ask the right questions, be skeptical when appropriate, and distinguish between useful and misleading evidence.

Cautionary Tales

To whet your appetite for the hard work ahead, let's start with a few cautionary tales that highlight the importance of thinking clearly in a data-driven world.

Abe's hasty diagnosis

Ethan's first child, Abe, was born in July of 2006. As a baby, he screamed and cried almost non-stop at night for 5 months. Abe was otherwise happy and healthy, though a bit on the small side. When he was one year old the family moved to Chicago, without which move, you'd not be reading this book. (That last sentence contains a special kind of claim called a *counterfactual*. Counterfactuals are really important and you are going to learn all about them in Chapter 3.) His new pediatrician noticed that Abe was not only small for his age; he was growing more slowly than expected.

After some lab work, the doctors were pretty sure Abe had celiac disease—a digestive disease characterized by gluten intolerance. The good news: celiac disease is not life threatening, or even terribly serious if properly managed through diet. The bad news: in 2007, the gluten-free dietary options for kids were pretty miserable.

It turns out that Abe actually had two celiac-related blood tests. One came back positive (indicating that he had the disease), the other negative (indicating that he did not have the disease). According to the doctors, the positive test was over 80% accurate. “This is a strong diagnosis,” they said. The suggested course of action was to put Abe on a gluten free diet for a couple of months to see if his weight increased. If it did, they could either do a more definitive biopsy or simply keep Abe gluten free for the rest of his life.

Ethan asked for a look at the report on Abes bloodwork. The doctors stated that they didnt think that would be useful since Ethan wasnt a doctor. The doctors response was neither surprising nor hard to understand. People, especially experts and authority figures, don’t like acknowledging the limits of their knowledge. But Ethan wanted to make the right decision for his son, so he pushed hard for the information. One of the goals of this book is to give you some of the skills and confidence to be your own advocate in this way, when making use of information to make decisions in your own life.

Two numbers characterize the effectiveness of any diagnostic test. The first is its false negative rate, which is how frequently the test says a sick person is healthy. The second is its false positive rate, which is how often the test says a healthy person is sick. You need to know *both* the false positive rate and the false negative rate to interpret a diagnostic test’s results. So Abe’s doctors’ statement that the positive blood test was 80% accurate wasn’t very informative. Did that mean it had a 20% false negative rate? A 20% false positive rate? Do 80% of people who test positive have celiac disease?

Fortunately, a quick Google search turned up both the false positive and false negative rates for both of Abe’s testsc. Here’s what Ethan learned. The test on which Abe came up

positive for celiac disease has a false negative rate of about 20%. That is, if 100 people with celiac disease took the test, about 80 of them would correctly test positive and the other 20 would incorrectly test negative. This fact, we assume, is where the claim of 80% accuracy came from. The test, however, has a false positive rate of 50%! People who don't have celiac disease are just as likely to test positive as they are to test negative. In contrast, the test on which Abe came up negative for celiac disease had much lower false negative and false positive rates.

Before getting the test results, a reasonable estimate of the probability of Abe having celiac disease, given his "failure to thrive," was around 1 in 100. That is, about 1 out of every 100 kids who have a diagnosis of "failure to thrive" have celiac disease. Armed with the lab reports and the false positive and false negative rates, Ethan was able to calculate how likely Abe was to have celiac disease given his failure to thrive and the test results. Amazingly, the combination of testing positive on an inaccurate test and testing negative on an accurate test actually meant that the evidence suggested that Abe was much *less* likely than 1 in 100 to have celiac disease. In fact, as we will show you in Chapter 15, the best estimate of the likelihood of Abe having celiac, given the test results, was about 1 in 1,000. The blood tests that Abe's doctors were sure supported the celiac diagnosis actually strongly supported the opposite conclusion. Abe was almost certain not to have celiac disease.

Ethan called the doctors to explain what he'd learned and to suggest that moving his pasta-obsessed son to a gluten free diet, perhaps for life, was not the prudent next step. Their response, "A diagnosis like this can be hard to hear." Ethan found a new pediatrician.

Here's the upshot. Abe did not have celiac disease. The kid was just a bit small. Today he is a normal-sized kid with a ravenous appetite. But if his father didn't know how to think about information or lacked the confidence to challenge a mistaken expert, he'd have spent his childhood eating rice cakes. Rice cakes are gross, so he'd probably still be small.

Civil resistance

You may someday find yourself unhappy with your government. If things get bad enough, you might even decide to organize a protest against government actions or policies. At that moment, you will face many important decisions. Perhaps none is more important than whether to build a movement committed to non-violent resistance or one open to more violent forms of anti-government protest. In thinking through this quandry, you will surely want to consult your personal ethics. But you might also want to know what the evidence says about the costs and benefits. Which kind of organization is most likely to succeed in changing government behavior? Is one or the other approach more likely to land you in prison, the hospital, or the morgue?

There is some quantitative evidence that you might use to inform your decisions. First, comparing anti-government movements across the globe and over time, those that use less violence are more likely to achieve their goals. Governments more often make concessions to fully non-violent groups than to groups that use violence. And even comparing across groups that do use violence, governments more frequently make concessions to those groups that engage in violence against military and government targets, rather than against civilians. Second, the personal risks associated with violent protest are greater than those associated with non-violent protest. Governments repress violent uprisings more often than they do non-violent protests making concerns about prison, the hospital, and the morgue are more acute.

This evidence sounds quite convincing. Non-violent protest seems the obvious choice. It is apparently both more effective and less risky. And, indeed, based on this kind of data, the political scientists Erica Chenoweth and Evan Perkoski conclude that “planning, training, and preparation to maintain nonviolent discipline is key—especially (and paradoxically) when confronting brutal regimes.”

But let's reconsider the evidence. Start by asking yourself, in what kind of a setting is a group likely to engage in non-violent versus violent protest? A few thoughts occur to us. Perhaps people are more likely to engage in non-violent protest when they face a government that they think is particularly likely to heed the demands of its citizens. Or perhaps people are more likely to engage in non-violent protest when they have broad-based support among their fellow citizens, or when they are less angry at their government, perhaps because it is less brutal.

If any of these things are true, we should worry about the claim that maintaining non-violent discipline is key to building a successful anti-government movement. (Which isn't to say that we are advocating violence.) Let's see why.

Empirical studies comparing the frequency or size of government concessions in places that had violent versus non-violent protests find that, on average, governments are more likely to make concessions in places that had non-violent protests. The claimed implications on a particular interpretation of that difference. Namely, that the higher frequency of government concessions in non-violent places is *caused* by the use of non-violent tactics. Put differently, all else held equal, if a given movement were to switch from violent to non-violent methods, the government would become more likely to grant concessions. But is this interpretation really justified by the evidence?

Suppose it's the case that protest movements are more likely to turn to violence when they do not have broad-based support among their fellow citizens. Then, when we compare places that had violent versus non-violent protests, all else (other than protest tactics) is not held equal. Those places differ in at least two ways. First, they differ in terms of whether they had violent versus non-violent protests. Second, they differ in terms of how supportive the public is of the protest movement.

This second difference is a problem. You might imagine that public opinion has an independent effect on the government's willingness to grant concessions. That is, all else held

equal (including protest tactics), governments might be more willing to grant concessions to protest movements with broad-based public support. If this is the case, then we can't really know whether the fact that governments grant concessions more often to non-violent protest movements compared to violent protest movements is because of the difference in protest tactics (as the evidence above suggests) or whether it is because the non-violent movements also happen to be the movements with broad-based public support. This is the classic problem of mistaking correlation for causation.

It is worth noting a couple of things. First, if government concessions are in fact due to public opinion, then it could be the case that, were we actually able to hold all else equal in our comparison of violent to non-violent protests, we would find the opposite relationship. That is, that non-violence is not more effective than violence (who knows, maybe it is even counter-productive). As things stand, we just can't know.

Second, in this example, the conclusion that appears to follow if you don't force yourself to think clearly is one we would all like to be true. Who among us would not like to live in a world where violence is always counterproductive? But the whole point of using evidence to help us make decisions is to force us to confront the possibility that the world may not be as we believe or hope it is. Indeed, it is in precisely those situations where the evidence seems to say exactly what you would like it to say that it is particularly important to force yourself to think clearly.

Third, we've pointed to one issue in this example, but there are others. For instance, think about the other empirical claim we discussed: that violent protests are more likely to provoke the government into repressive crack-downs than are non-violent protests. Recall, we suggested that people might be more likely to engage in non-violent protest when they are less angry at their government, perhaps because it is less brutal. Ask yourself why, if this is true, we have a similar problem of interpretation. That is, why might the fact that there are more government crackdowns on average following violent protests than following

non-violent protests not mean that switching from violence to non-violence will reduce the risk of crack-downs. The argument will follow a similar logic to the one we just made regarding concessions. If you don't see how the argument works yet, that's okay. You will by the end of Chapter 9.

broken windows policing

In 1982, the criminologist George L. Kelling and the sociologist James Q. Wilson published an article in *The Atlantic* proposing a new theory of crime and policing that had enormous and long-lasting effects on crime policy in the United States and beyond.

Kelling and Wilson's theory is called "broken windows." It was inspired by a program in Newark, NJ that got police out of their cars and back walking a beat. According to Kelling and Wilson, the program had a profound effect on the communities in which it was implemented—increasing residents' sense of security and reducing their fear of violent crime.

What, Kelling and Wilson wondered, was it about having police on foot-patrol that led to this increased sense of security? Their answer, "what foot-patrol officers did was to elevate...the level of public order in these neighborhoods."

Reducing public disorder (e.g., panhandling, loitering, prostitution, destruction of property) is fundamental to good policing, in Kelling and Wilson's view, because disorder breeds criminality: "[A]t the community level, disorder and crime are usually inextricably linked, in a kind of developmental sequence. Social psychologists and police officers tend to agree that if a window in a building is broken and is left unrepaired, all the rest of the windows will soon be broken." Kelling and Wilson describe a vicious cycle in which small amounts of disorder lead gradually to communal break down and the rise of crime:

A piece of property is abandoned, weeds grow up, a window is smashed. Adults

stop scolding rowdy children; the children, emboldened, become more rowdy. Families move out, unattached adults move in. Teenagers gather in front of the corner store. The merchant asks them to move; they refuse. Fights occur. Litter accumulates. People start drinking in front of the grocery... At this point it is not inevitable that serious crime will flourish or violent attacks on strangers will occur. But many residents will think that crime, especially violent crime, is on the rise, and they will modify their behavior accordingly. They will use the streets less often, and when on the streets will stay apart from their fellows... Such an area is vulnerable to criminal invasion. Though it is not inevitable, it is more likely that here, rather than in places where people are confident they can regulate public behavior by informal controls, drugs will change hands, prostitutes will solicit, and cars will be stripped.

In addition to their observation of the Newark program, Kelling and Wilson cited the work of the psychologist Philip Zimbardo to bolster the case for the broken windows theory. Zimbardo conducted a study in which he abandoned two similar cars—both without license plates and with the hood left up—in two different settings. The first was abandoned in a disorderly part of the Bronx, the second in the affluent city of Palo Alto. The car in the Bronx was stripped and vandalized within a day, whereas the car in Palo Alto remained unmolested for over a week.

Subsequent studies have shown similar findings in more controlled environments. For instance, scientists from the University of Groningen in the Netherlands conducted randomized experiments in which they created a temptation to engage in some undesirable behavior such as littering, taking five dollars that didn't belong to you, or jumping a fence. They found that people were more likely to engage in such behavior in disorderly settings—for instance, when there was visible graffiti despite signs forbidding it—than in orderly settings.

This idea that policing focused on minimizing disorder can help to reduce violent crime had a big impact on police tactics. Most prominently, broken windows theory was the guiding philosophy in New York City in the 1990s. In a 1998 speech, then New York mayor Rudy Giuliani said,

We have made the ‘broken windows’ theory an integral part of our law enforcement strategy...There’s a continuum of disorder. Obviously, murder and graffiti are two vastly different crimes. But they are part of the same continuum, and a climate that tolerates one is more likely to tolerate the other...you concentrate on the little things, and send the clear message that this City cares about maintaining a sense of law and order, which is another way of saying people respecting the rights of other people, then the City as a whole will begin to become safer.¹

And, indeed, New York city police did increase their focus on the “little things” during the 1990s. According to a study by Hope Corman and Naci Mocan, misdemeanor arrests increased 70 percent during the 1990s. This broken windows approach to policing was accompanied by a dramatic reduction in crime in New York City. The same study reports that violent crime declined by more than 56 percent, double the reduction in the country as a whole.

To assess the extent to which broken windows was responsible for these good outcomes in New York City, Kelling and another collaborator, William Sousa, studied the relationship between violent crime and broken windows approaches across the city’s precincts. The extent to which the police focused on the “little things” differed from precinct to precinct. Thus, they reasoned, if minimizing disorder really causes a reduction in violent crime, then we should expect the largest reductions in crime in neighborhoods where the police were

¹<http://www.nyc.gov/html/records/rwg/html/98a/quality.html>

most focused on the broken windows approach.

For each of 75 precincts in New York City, Kelling and Sousa measured violent crime, misdemeanor arrests, various economic indicators, shifts in the young male population, and trends in crack cocaine consumption (the decline of the crack epidemic is often thought to be responsible for the reduction in crime nationwide). Encouragingly for the broken windows theory, they found that in places where misdemeanor arrests (the “little things”) were higher, violent crime was lower. They calculated that, “the average NYPD precinct... could expect to suffer one less violent crime for approximately every 28 additional misdemeanor arrests.”

This sounds pretty convincing. But let’s not be too quick to conclude that arresting people for misdemeanors is the answer to ending violent crime. Two other scholars, Bernard Harcourt and Jens Ludwig point out that the conclusions to be drawn from Kelling and Sousa’s data are not as straightforward as it appears at first blush. They encourage us to think a little more clearly about what might be going on.

Start by asking yourself which precincts were most likely to have been the focus of new policing tactics aimed at reducing violent crime. One would expect it to be those precincts most in need—that is, those that had recently suffered a big run-up in violent crime. And, indeed, this was the case.

Kelling and Sousa’s analysis focused on the relationship between misdemeanor arrests and violent crime in the 1990s, showing that precincts that had more misdemeanor arrests in the 1990s had a bigger reduction in violent crime. Harcourt and Ludwig show that the precincts that had more misdemeanor arrests in the 1990s were also the most violent precincts in the late 1980s—perhaps not surprisingly, the police implemented broken windows most intensively in the most high-crime precincts.

The problem that Harcourt and Ludwig point out is something called reversion to the

mean (which we'll talk about a lot more in Chapter 8). Here's the basic concern. In any given year, the amount of crime in a precinct is determined by lots of factors including policing, the local population, drugs, the economy, the weather, and so on. Many of those factors are unknown to us. Some of them are fleeting; they come and go across precincts year to year. As such, in any given precinct, we can think of there being some "baseline" level of crime, with some years randomly having more crime and some years randomly having less.

Imagine a precinct that had a really high level of violent crime in some year. Two things are likely to be true of that precinct. First, it is probably a precinct with a high baseline of violent crime. Second, it is also probably a precinct that had a bad year—i.e., for idiosyncratic reasons, the level of crime for that year was high relative to that precinct's baseline. The same, of course, is true in reverse for precincts that had a low level of crime in a given year. They probably have a low baseline of crime. But they also probably had a particularly good year. As we said, these random shifts in crime are fleeting. Some years are good in a precinct, some years are bad. On average, years are average. Thus, on average, precincts that had a particularly bad year this year will have a better year next year—not because bad luck has to run out, but because luck is determined by chance and if you had bad luck this year, on average your luck will be better next year.

Why is this a problem for Kelling and Sousa's conclusions? Because of reversion to the mean, we would expect the most violent precincts in the late 1980s to show a reduction in violent crime on average, even with no change in policing. Unfortunately for the study, it was those same high-crime precincts in the 1980s that were most likely to get broken windows policing in the 1990s. So, when we see a reduction in violent crime in the precincts that got broken windows policing, we don't know if it's the policing or reversion to the mean that's at work.

Harcourt and Ludwig go a step further to try to find more compelling evidence. Roughly

speaking, they look at how changes in misdemeanor arrests relate to changes in violent crime in precincts that started with similar levels of violent crime. By comparing precincts with similar starting levels of violent crime, they go some way toward eliminating the problem of reversion to the mean. Surprisingly, this simple change actually entirely flips the relationship! Rather than confirming Kelling and Sousa's finding that misdemeanor arrests are associated with a reduction in violent crime, Harcourt and Ludwig find that precincts that focused more on misdemeanor arrests actually appear to have experienced an *increase* in violent crime. Exactly the opposite of what we would expect if the broken windows theory is correct.

Now, this reversal doesn't settle the matter on the efficacy of broken windows policing. The relationship between misdemeanor arrests and violent crime that Harcourt and Ludwig find could be there for lots of reasons other than misdemeanor arrests causing an increase in violent crime. For instance, perhaps the neighborhoods with increasing misdemeanors are becoming less safe in general and would have experienced more violent crime regardless of policing strategies. What these results do show is that, as we think about how to deploy scarce police resources, the data, properly considered, don't offer the kind of unequivocal support for the broken windows idea that it might have seemed at first blush. And you can only see this if you have the ability to think clearly about some pretty tricky questions.

Importantly, however, you don't have to be a technical expert to ask the right questions. In which precincts did broken windows policing get implemented? What would the likely pattern of violence in those kinds of precincts have been even without a change in policing? These are conceptual questions that you can learn to ask without formal training in statistics or criminology.

Thinking and Data Are Complements, Not Substitutes

As these cautionary tales highlight, information and data alone can't tell us the answer to important questions like whether broken windows policing is, on net, a good policing policy. Data alone can't even tell us the answer to seemingly straightforward questions like whether Abe has celiac disease or which kinds of protests are most effective. Just because an argument contains seemingly sophisticated quantitative data analysis, that doesn't mean the argument is rigorous or right. To harness the power of information to make better decisions we must combine data analysis with clear thinking.

Our stories also illustrate how our intuitions can lead us astray. It takes lots of care and practice to train ourselves to think clearly about evidence. The doctors' intuition that Abe had celiac disease because of a test with 80% accuracy and the researchers' intuition that broken windows policing works because crime decreased in places where it was deployed seem sensible. But both intuitions are wrong, suggesting that we should be skeptical of our initial hunches. The good news is that clear thinking can become intuitive if you work at it.

Data and quantitative tools are not a substitute for clear thinking. In fact, quantitative skills without clear thinking is a dangerous thing. We suspect, as you read the coming chapters, you will be jarred by the extent to which unclear thinking affects even the most important decisions people make. Through the course of this book, we will see how misinterpreted information distorts life-and-death medical choices, national and international counterterrorism policies, business and philanthropic decisions made by some of the world's wealthiest people, how we set priorities for our children's education, and a host of other issues from the banal to the profound. Essentially no aspect of life is immune from critical mistakes in understanding and interpreting quantitative information.

In our experience, this is because unclear thinking about evidence is deeply ingrained

in human psychology. Certainly our own intuitions, left unchecked, are frequently subject to basic errors. Our guess is that yours are too. Most disturbingly, the experts on whose advice you depend—be they doctors, business consultants, journalists, teachers, financial advisors, scientists, or what have you—are often just as prone to making such errors as the rest of us. All too often, because they are experts, we trust their judgment without question and so do they. That is why it is so important to learn to think clearly about quantitative evidence for yourself. That is the only way to know how to ask the right questions that lead you, and those on whose advice you depend, to the most reliable and productive conclusions possible.

How could experts in so many fields make important errors so often? Expertise, in any area, comes from training, practice, and experience. No one expects to become an expert in engineering, finance, plumbing, or medicine without instruction and years of work. But, despite its fundamental and increasing importance for so much of life in our quantitative age, almost no one invests this kind of effort into learning to think clearly. And, as we've said, even when they do, they tend to be taught in a way that over-emphasizes the technical and under-emphasizes the conceptual, even though the fundamental problems are almost always about conceptual mistakes in thinking, rather than technical mistakes in calculation.

The lack of expertise in thinking presents us with two challenges. First, if so much expert advice and analysis is unreliable, how do you know what to believe? Second, how can you identify those expert opinions that in fact reflect clearly thinking?

This book provides a framework for addressing these challenges. Each of the coming chapters explains and illustrates, through a variety of examples, fundamental principles of clear thinking in a data-driven world. Part I establishes some shared language—clarifying what we mean by correlation and causation and what each is useful for. Part II discusses how we can tell whether a statistical relationship is genuine. Part III discusses how we can tell if that relationship reflects a causal phenomenon or not. And Part IV discusses

how we should and shouldn't incorporate quantitative information into our decision making.

Our hope is that reading this book will help you internalize the principles of clear thinking in a deep enough way that they start to become second nature. You will know you are on the right path when you find yourself noticing basic mistakes in how people think and talk about the meaning of information everywhere you turn—as you watch the news, read magazines, talk to business associates, visit the doctor, listen to the color commentary during athletic competitions, or participate in school, church, or other communal activities. You will, we suspect, find it difficult to believe how much nonsense you're regularly told by all kinds of experts. When this starts to happen, try to remain humble and constructive in your criticisms. But do feel free to share your copy of this book with those whose arguments you find are in particular need of it. Or better yet, encourage them to buy their own copy!

Readings

The essay on non-violence protest by Erica Chenoweth and Evan Perkoski that we quoted can be found at <https://politicalviolenceataglance.org/2018/05/08/states-are-far-less-likely-t>

The following book contains more research on the relationship between non-violence and efficacy:

Erica Chenoweth and Maria J. Stephan. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. Columbia University Press, 2011.

The following articles were discussed in this order on the topic of broken windows policing:

George L. Kelling and James Q. Wilson. 1982. *The Atlantic*. March Issue. <https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/>.

Philip G. Zimbardo. 1969. The Human Choice: Individuation, Reason, and Order versus Deindividuation, Impulse, and Chaos. *Nebraska Symposium on Motivation*, edited by W.J. Arnold and D. Levine. Lincoln: University of Nebraska Press.

Kees Keizer, Siegwart Lindenberg, and Linda Steg. 2008. The Spreading of Disorder. *Science* 322:1681-1685.

Hope Corman and H. Naci Mocan. 2005. Carrots, Sticks, and broken windows. *Journal of Law and Economics* 48(1):235-266.

George L. Kelling and William H. Sousa, Jr. 2001. Do Police Matter? An Analysis of the Impact of New York City's Police Reforms. Civic Report for the Center for Civic Innovation at the Manhattan Institute.

Bernard E. Harcourt and Jens Ludwig. 2006. broken windows: New Evidence from New York City and a Five-City Social Experiment. *University of Chicago Law Review* 73:271-320. *The published version has a misprinted sign in the key table. For the correction, see Errata, 74 U. Chi. L. Rev. 407 (2007).*

Part I

Establishing a Common Language

Chapter 2

Correlation: What Is It and What Is It Good For?

What You'll Learn

- Correlations tell us about the extent to which two features of the world tend to occur together.
- In order to measure correlations, we must have data with variation in both features of the world.
- Correlations *can* be useful for description, forecasting, and potentially causal inference. But we have to think clearly about when they're appropriate for each of these tasks.
- Correlations are about linear relationships, but that's not as limiting as you might think.

You may have heard the statement *correlation doesn't imply causation*. It's a good adage. However, in our experience, it's less useful than it might be because hardly anyone knows what correlation or causation are.

In Part I, we are going to spend some time establishing a shared vocabulary. Making sure that we are all using these and a few other key terms to mean the same thing is absolutely critical if we are to think clearly about them in the chapters to come.

This chapter is about correlation: what it is and what it's good for. Correlation is the primary tool through which quantitative analysts describe the world, forecast future events, and answer scientific questions. Careful analysts do not eschew or disregard correlations. But they must think clearly about which kinds of questions correlations can and cannot answer in different situations.

What is a correlation?

The *correlation* between two features of the world is the extent to which they tend to occur together. This definition tells us that a correlation is a relationship between two things (which we call *features of the world* or *variables*). If two features of the world tend to occur together, they are *positively correlated*. If the occurrence of one feature of the world is unrelated to the occurrence of another feature of the world, they are *uncorrelated*. And if, when one feature of the world occurs the other tends not to occur, they are *negatively correlated*.

What does it mean for two features of the world to tend to occur together? Let's start with an example of the simplest kind. Suppose we want to assess the correlation between two things (variables), and there are only two possible values for each one (we call these *binary* variables). For instance, whether it is after noon or before noon is a binary variable (by contrast, the time measured in hours, minutes, and seconds is not binary, it can take

many more than two values).

Political scientists and economists sometimes talk about the “resource curse” or the “paradox of plenty.” The idea is that countries with an abundance of natural resources are often less economically developed and less democratic than those with fewer natural resources. Natural resources might make a country less likely to invest in other forms of development, or they might make a country more subject to violence and exploitation.

To assess the extent of this resource curse, we might want to collect some data and measure the correlation between natural resources and some feature of the economic or political system. That process starts with collecting some data, which we’ve done. To measure natural resources we looked at which countries are major oil producers. We classify a country as a major oil producer if it exports more than 40,000 barrels per day per million people. And for the political system we looked at which countries are deemed to be autocracies versus democracies by the Polity IV Project. Table 2.1 indicates how many countries fit into each of the four possible categories: democracy and major oil producer, democracy and not major oil producer, autocracy and major oil producer, and autocracy and not major oil producer.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

Table 2.1: Oil Production and Type of Government.

We can figure out if these two variables—being a major oil producer and autocracy

versus democracy—are correlated by making a comparison. For instance, we could ask whether major oil producers are more likely to be autocracies than countries that aren't major oil producers. Or, similarly, we could ask whether autocracies are more likely to be major oil producers than democracies. If one of these statements is true, the other must be true as well. And these comparisons tell us whether these two features of the world—being a major oil producer and being an autocracy—tend to occur together.

In Table 2.1, oil production and autocracy are indeed positively correlated. 55 percent of major oil producers are autocracies ($\frac{11}{20} = .55$) while only about 20 percent of countries that aren't major oil producers are autocracies ($\frac{29}{147} \approx .20$). Equivalently, 27.5 percent of autocracies are major oil producers ($\frac{11}{40} = .275$), while only about 7 percent of democracies are ($\frac{9}{127} \approx .07$). In other words, major oil producers are more likely to be autocracies than are countries that aren't major oil producers, and then, necessarily, autocracies are more likely to be major oil producers than democracies.

As a descriptive matter, we find this positive correlation interesting. It is also potentially useful for prediction. Suppose there were some other countries outside our data whose system of government we were uncertain of. Knowing whether or not they were major oil producers would be helpful in predicting which kind of government they likely have.

Such knowledge could even potentially be useful for causal inference. Perhaps new oil reserves are discovered in a country and the State Department wants to predict what is likely to happen to the country's political system. This kind of data could potentially be informative for that question as well. Though, as we'll discuss in great detail in Chapter 9, we must be very careful about giving correlations this sort of causal interpretation.

We can assess correlations even when our data are such that it is hard to make a table of all the possible combinations like we did above. Suppose, for example, that we want to assess the relationship between crime and temperature in Chicago. We could assemble a spreadsheet where each row corresponds to a day, and each column corresponds with

a feature of each day. We often call the rows *observations* and the features listed in the columns *variables*. In this case, the observations are different days. One variable could be the average temperature on that day as measured at Midway Airport. Another could be the number of crimes reported in the entire city of Chicago on that day. Another still could indicate whether the Chicago Tribune ran a story about crime on its front page on that day. As you can see, variables can take values that are binary (front page story or not), discrete but not binary (number of crimes), or continuous (average temperature).

We've collected data like this for Chicago in 2018, and we'd like to assess the correlation between crime and temperature. But how can we assess the correlation between two non-binary variables?

One starting point is to make a simple graph, called a *scatter plot*. Figure 2.1 shows one for our 2018 Chicago data. In it, each point corresponds to an observation in our data—here, that means each point is a day in Chicago in 2018. The horizontal axis of our figure is the average temperature at Midway Airport on that day. The vertical axis is the number of crimes reported in the city that day. So the location of each point shows the average temperature and the amount of crime on a given day.

Just by looking at the figure, you can see that it sure appears that there is a positive correlation between temperature and crime. Points to the left of the graph on the horizontal axis (cold days) tend to also be pretty low on the vertical axis (low crime days) and days to right of the graph on the horizontal axis (warm days) tend to also be pretty high on the vertical axis (high crime days).

But how can we quantify this visual first impression? There are actually many different statistics that we can use to do so. One such statistic is called the *slope*. Suppose we found the line that best fits the data. By best fit, we mean, roughly, the line that minimizes how far the data points are from the line on average. (We will be more precise about this in Chapter 5.) The slope of the line of best fit is one way of describing the correlation between

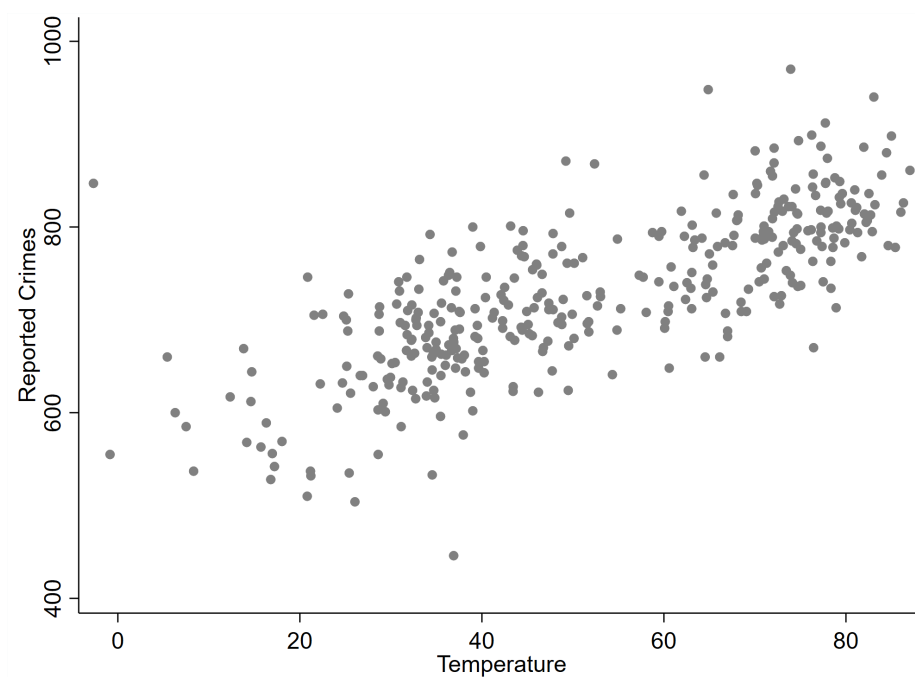


Figure 2.1: Crime and temperature in Chicago across days in 2018.

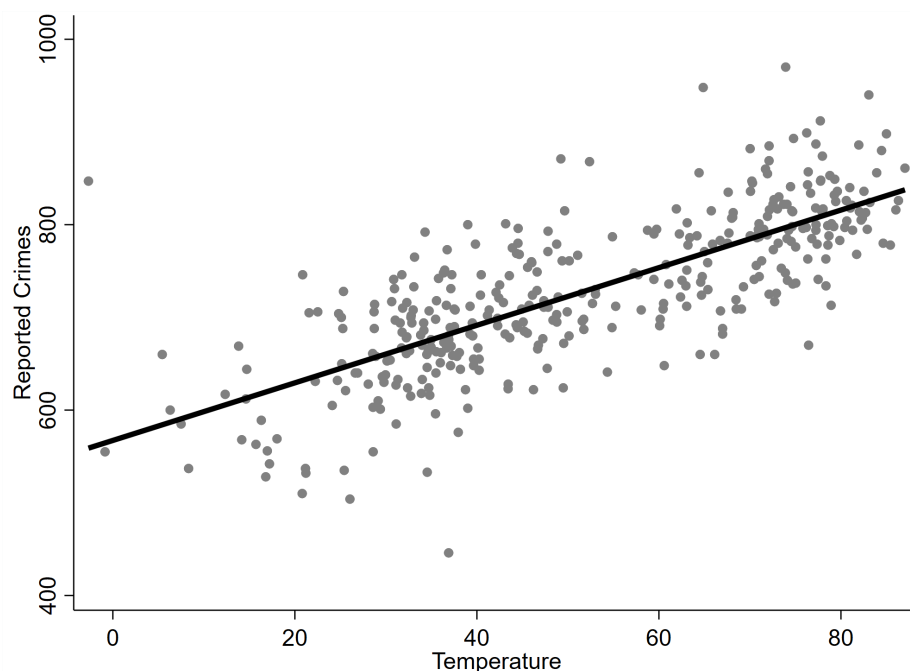


Figure 2.2: A line of best fit summarizing the relationship between crime and temperature.

these two continuous variables.

Figure 2.2 shows the scatter plot with that line added. The slope of the line tells us something about the relationship between those two variables. If the slope is negative, the correlation is negative. If the slope is zero, the temperature and crime are uncorrelated. If the slope is positive, the correlation is positive. And the steeper the slope, the stronger the correlation. Here we see that they are positively correlated—there tends to be more crime on warmer days. In particular, the slope is 3.1, so on average for every additional degree of temperature (in Fahrenheit), there are 3.1 more crimes.

Notice that how you interpret the slope depends on which variable is on the vertical axis and which one is on the horizontal axis. Had we drawn it the other way around (as in Figure 2.3), we would be describing the relationship between the same two variables. But

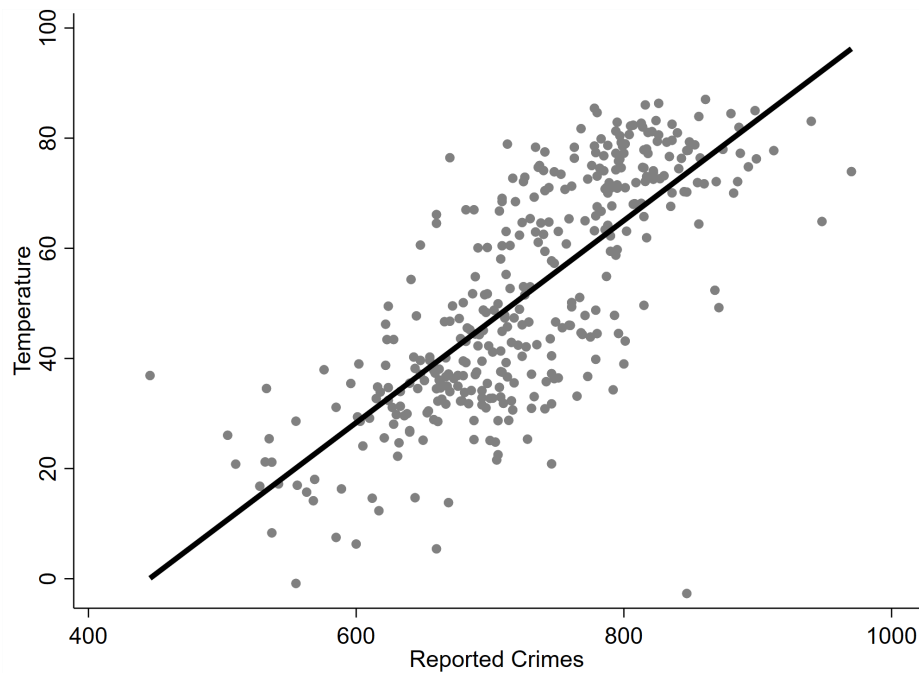


Figure 2.3: A line of best fit summarizing the relationship between temperature and crime.

this time, we would have learned that for every additional reported crime, on average, the temperature is 0.18 degrees higher. The sign of the slope (positive or negative) is the same regardless of which is on the horizontal vs. vertical axis because changing which variable is on which axis does not change whether they are positively or negatively correlated. But the actual number describing the slope and its substantive interpretation has changed.

Fact or correlation?

In order to establish whether a correlation exists, you must always make a comparison of some kind. For example, to learn about the correlation between temperature and crime, we need to compare hot and cold days and see whether the levels of crime differ, or alternatively, we can compare high- and low-crime days to see if their temperatures differ. This means

that to assess the correlation between two variables, we need to have variation in both variables. For example, if we collected data only on days when the average temperature was 0 degrees, we would have no way of assessing the correlation between temperature and crime. And the same is true if we only examined days with 500 reported crimes.

With this in mind, let's pause to check how clearly you are thinking about what a correlation is and how we learn about one. Don't worry if you aren't all the way there yet. Understanding whether a correlation exists turns out to be tricky. We are going to spend all of Chapter 4 on this topic. Nonetheless, it is helpful to do a preliminary check now. So let's give it a try.

Think about the following statements. Which ones describe a correlation and which ones do not?

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least 10,000 hours honing their craft.
4. Most politicians engulfed in a scandal win reelection.
5. Older people vote more than younger people.

How did you do? (It's a rhetorical question. How would you know how you did at this point?)

While each of these statements report a fact, not all of those facts describe a correlation—i.e., evidence on whether two features of the world tend to occur together. In particular, Statements 1, 3, and 4 do not describe correlations, while statements 2 and 5 do. Let's unpack this.

Statements 1, 3, and 4 are facts. They come from data. They sound scientific. And if we added specific numbers to these statements, we could call them *statistics*. But not all facts or statistics describe correlations. The key issue is that these statements do not describe whether or not two features of the world tend to occur together. And the reason is because they do not compare across different values of both features of the world.

To get a sense of this, focus on statement 4:

Most politicians engulfed in a scandal win reelection.

Two features of the world are discussed. The first is whether a politician is engulfed in a scandal. The second is whether the politician successfully wins reelection. So the correlation being hinted at is a positive correlation between being engulfed in a scandal and winning reelection. But we don't actually learn from this statement of fact whether those two features of the world tend to occur together. And the reason is that we have not compared the rate of winning reelection for those engulfed in scandal to the rate of reelection for those not engulfed in scandal.

We can assess this correlation, just not with the data described in statement 4. To assess the correlation, we'd need variation in both variables—being engulfed in a scandal and winning reelection. Just for fun, let's examine this correlation in some real data on incumbent members of the U.S. House of Representatives seeking reelection between 2006 and 2012. Scott Basinger from the University of Houston has systematically collected data on congressional scandals. Utilizing his data, let's see how many cases fall into four relevant cases: members engulfed in a scandal who were reelected, members engulfed in a scandal who were not reelected, scandal-free members who were reelected, and scandal-free members who were not reelected.

In Table 2.2, we see that statement 4 is indeed a true fact. 62 out of 70 (about 89%) members of Congress engulfed in scandal who sought reelection indeed won. But we also

	Not engulfed in scandal	Engulfed in scandal	Total
Not reelected	101	8	109
Reelected	1,192	62	1,254
Total	1,293	70	1,365

Table 2.2: Most members of Congress engulfed in scandal are reelected, but scandal and reelection are negatively correlated.

see that most members of Congress not engulfed in scandal won reelection. In fact, 1,192 out of 1,293 (about 92%) of these scandal-free members won reelection. By comparing the scandal-plagued members to the scandal-free members, we now see that there is actually a slight negative correlation between being engulfed in scandal and winning reelection.

We hope it is now clear why statement 4 does not convey enough information to know whether or not there is a correlation between scandal and reelection. The problem is that the statement is only about politicians engulfed in scandal. It tells us that more of those politicians win reelection than lose. But to figure out if there is a correlation between scandal and winning reelection, we need to compare the share of politicians engulfed in scandal who win reelection to the share of scandal-free politicians who win. Had only 85% of the scandal-free members of Congress won reelection, there would be a positive correlation between scandal and reelection in this case. Had 89% of them won, there would have been no correlation at all. But since we now know the true rate of reelection for scandal-free members was 92%, we see that there is a negative correlation. A similar analysis will show you that statements 1 and 3 also don't convey enough information, on their own, to assess a correlation.

Statements 2 and 5 do indeed describe correlations. Both statements make a comparison. Statement 2 tells us that cities with more crime have, on average, larger police forces than

cities with less crime. And statement 5 tells us that older people tend to vote at higher rates than younger people. In both cases, we are comparing differences in one variable (police force size or voting rates) across differences in the other variable (crime rates or age). This is the kind of information you need to establish a correlation.

As we said at the outset, don't worry if you got confused on this. Thinking clearly about what kind of information is necessary to establish a correlation, as opposed to just a fact, is tricky. We are going to spend Chapter 4 making sure you've really got it.

What is a Correlation Good For?

Now that we have a shared understanding of what a correlation is, let's talk about what a correlation is good for. We've noted that correlations are perhaps the most important tool of quantitative analysts. But why? Broadly speaking, it's because correlations tell us what we should predict about some feature of the world given what we know about other features of the world.

There are at least three uses for this kind of knowledge: (1) description, (2) forecasting, and (3) causal inference. Any time we make use of a correlation, we want to think clearly about which of these three tasks we're attempting and what has to be true about the world for a correlation to be useful for that task in our particular setting.

Description

Describing the relationships between features of the world is the most straightforward use for correlations.

Why might we want to describe the relationship between features of the world? Suppose you were interested in whether younger people are underrepresented at the polls in a particular election, relative to their size in the population. A description of the relationship

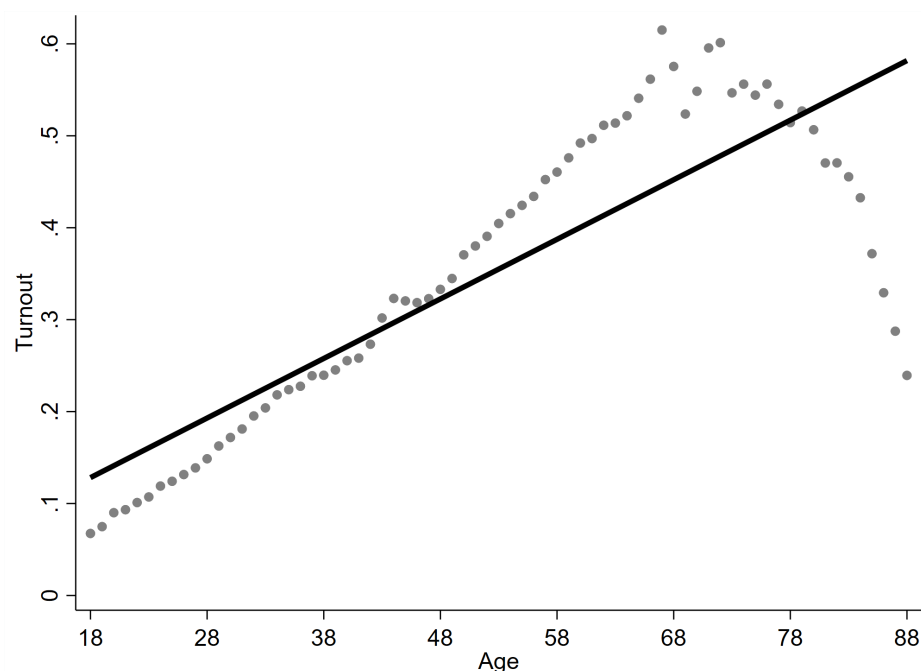


Figure 2.4: Voter turnout and age in the 2014 election.

between age and voting might be helpful. Figure 2.4 shows a scatter plot of data on age and average voter turnout for the 2014 U.S. Congressional election. In this figure, an observation is an age cohort. For each year of age, the figure shows the proportion of eligible voters who turned out to vote.

The figure also plots the line that best fits the data. This line has a slope of 0.006. In other words, on average, for every additional year of age, the chances that an individual turned out to vote in 2014 increases by 0.6 percentage points. So younger people do indeed appear to be under-represented, as they turnout at lower rates than older people.

This kind of descriptive analysis may be interesting in and of itself. It's important to know that younger people were less likely than older people to vote in 2014 and were therefore underrepresented in the electoral process. That relationship may inform how you think about the outcome of that election. Moreover, knowledge of this correlation might

motivate you to further investigate the causes and consequences of the phenomenon of younger people turning out at low rates.

Of course, uncovering this descriptive relationship doesn't necessarily mean that these younger people will continue to vote at low rates in future elections. So you can't necessarily use this knowledge to forecast future voter turnout. And it also doesn't mean that these younger people will become more likely to vote as they age. So you can't necessarily interpret this relationship causally. This descriptive analysis just tells us that older people were more likely to vote than younger people, on average, in the 2014 election. To push the interpretation further, you'd need to be willing to make stronger assumptions about the world, which we will now explore.

Forecasting

Another motivation for looking at correlations is forecasting or prediction. Forecasting involves using information from some sample population to make predictions about a different population.

Suppose you're running an electoral campaign, you have limited resources, and you're trying to figure out which of your supporters you should target with a knock on the door reminding them to turn out to vote. If you were already highly confident that an individual was going to vote in the absence of your intervention, you wouldn't want to waste your volunteers' time by knocking on that door. So accurate forecasting of voter turnout rates could improve the efficiency of your campaign.

Correlations like the one above regarding age and voter turnout could be useful for this kind of forecasting. Since age is strongly correlated with turnout, it might be a useful variable for forecasting who is and is not already likely to vote. For instance, if you were able to predict, based on age, that some group of voters is virtually certain to turn out

even without your campaign efforts, you might want to focus your mobilization resources on other voters.

To use the correlation between age and voter turnout for forecasting in this way, you don't need to know why they are correlated. But, unlike if you just want to describe the relationship between age and voter turnout in the 2014 election, if you want to forecast you do need to be willing to make some additional assumptions about the world.

As mentioned above, forecasting involves using information from one sample of data in order to make predictions about a different population. For instance, you might be using data on voters from past elections to make predictions about voters in future elections. Or you might be using the voters in one state to make predictions about voters in another state.

This raises two important concerns that you must think clearly about in order to use correlation for forecasting responsibly. The first is whether the relationship you found in your sample is indicative of a broader phenomenon or whether it is the result of chance variation in your data. Answering this question requires what we call *statistical inference*, which is the topic of Chapter 6. Second, even if you are convinced that you've found a real relationship in your sample, you'll want to think about whether your sample representative of the population that you are trying to make predictions about.

Let's go back to using age and voter turnout from one election to make predictions about another election. Doing so only makes sense if it is reasonable to assume that the relationship between these two variables isn't changing too quickly. That is, the correlation between age and voter turnout in, say the 2014 election is only useful for figuring out which voters to target in the 2016 election if it seems likely that the relationship between age and turnout in 2016 will be more or less the same as the relationship between age and turnout in 2014. Similarly, if you only had data on age and voter turnout in the 2014 election for 25 states, you might try to predict the correlation in the other 25 states. This would only

be sensible if you had reason to believe the states on which you did and did not have data were similar on relevant dimensions.

You'd also want to take care in making predictions beyond the range of available data. Our data tell us voter turnout rates for voters age 18–88. The line of best fit, of course, gives us predictions for any age. But we should be less confident extrapolating our predictions about voter turnout to 100-year-olds, since we don't actually observe any data for them, so, can't know whether the relationship described by the line is likely to hold for them or not, even for the 2014 election.

Relatedly, when using some statistic, like the slope of a line of best fit, to do prediction, one wants to think about whether the relationship is actually linear. If not, a linear summary of the relationship might be misleading. We'll discuss this in greater detail below.

It is worth noting that, in practical applications, it would be unusual to try to do forecasting simply using the correlation between two variables. One might, instead, try to predict voter turnout using its relationship with a host of variables like gender, race, income, education, or previous voter turnout. We'll discuss such multivariable and conditional correlations in Chapter 5.

Using data for forecasting and prediction is a rapidly growing area for analysts in policy, business, policing, sports, government, intelligence, and many other areas. For instance, suppose you're running your city's public health department. Every time you send a health inspector to a restaurant, it costs time and money. But restaurant violations of the health code do harm to your city's residents. Therefore, you would very much like to send inspectors to those restaurants that are most likely to be in violation of the health codes, so as not to waste time and money on inspections that don't end up improving public safety. The more accurately you can forecast which restaurants are in violation, the more effectively you can deploy your inspectors. You could imagine using data on restaurants that did and did not violate health codes in the past to try to predict such violations based on their correlation

with other observable features of a restaurant. Plausibly useful restaurant features might include Yelp reviews, information about hospital visits for food poisoning, location, prices, and so on.

This example points to another tricky issue. The very act of using correlations for prediction can sometimes make correlations that held in the past cease to hold in the future. For instance, suppose the health department observes a strong correlation between restaurants that are open 24 hours a day and health code violations. Based on that correlation, they might start sending health inspectors disproportionately to 24 hour restaurants. A savvy restaurant owner who becomes aware of the new policy might adapt to fool the health department, say closing from 2 AM to 3 AM every night. This small change in operating hours would presumably do nothing to clean up the restaurant. But the manager would have gamed the system, rendering predictions based on past data inaccurate for the future. We'll discuss this general problem of adaptation in greater detail in Chapter 16.

Forecasting would also be useful to a policy maker who would like to know the expected length of an economic downturn for budgetary purposes, a banker who wants to know the credit worthiness of potential borrowers, or an insurance company that wants to know how many car accidents a potential client is likely to get in this year. The managers of our beloved Chicago Bears would love to predict which college football players could be drafted to increase the team's chances of winning a Super Bowl, but unfortunately, data can't work miracles.

It is also worth thinking about the potential ethical implications of using predictions to guide behavior. Suppose it turns out that restaurants in neighborhoods with residents of a particular ethnic or racial identity happen to violate health codes more often. That's useful predictive information. But do we want our government to use such information to decide where to send health inspectors? Or are there ethical or social costs of targeting restaurants for inspection based on ethnicity or race that outweigh the benefits of more

accurate predictions? We will return to some of these ethical issues at the end of the book.

Causal inference

We might be interested in correlations is to learn about causal relationships. Many of the most interesting questions that quantitative analysts face are inherently causal. That is, they are about how changing some feature of the world would change some other feature of the world. Would improving college affordability reduce income inequality? Would increased policing curtail gang violence? Would a new marketing strategy boost profits? These are all causal questions. As we'll see throughout the book, using correlations to make inferences about causal relationships is common. But it is also fraught with opportunities for unclear thinking. (Understanding causality will be the subject of the next chapter.)

Using correlation for causal inference has all the potential issues we just discussed when thinking about using correlation for prediction. And, in addition, there are new issues. The key one is that correlation need not imply causation. That is, a correlation between two features of the world doesn't mean one of them causes the other.

Suppose you want to know the effect of high school math training on subsequent success in college. This is clearly an important question if you're a high school student, a parent or counselor of a high school student, or a policymaker setting educational standards. Will high school students be more likely to attend and complete college if they take advanced math in high school?

As it turns out, the correlation between taking advanced math and completing college is positive and quite strong—for instance, people who take calculus in high school are much more likely to complete college. And the correlation is even stronger for Algebra 2, Trigonometry, and pre-Calculus. But that doesn't mean that taking calculus causes students to complete college.

Of course, one possible source of this correlation is that calculus prepares students for college and causes them to become more likely to graduate. But that isn't the only possible source of this correlation. For instance, maybe, on average, kids who take calculus are more academically motivated than kids who don't. And maybe motivated kids are more likely to complete college regardless of whether or not they take calculus in high school. If that is the case, we would see a positive correlation between taking calculus and completing college even if calculus has no effect on college completion. Rather, whether a student took calculus would simply be an indirect measure of motivation, which is correlated with success in college.

What's at stake here? Well, if the causal story is right, then requiring a student to take calculus who otherwise wouldn't will help that student complete college by offering better preparation. But if the motivation story is right, then requiring that student to take calculus will not help with college completion. In that story, calculus is just an indicator for intrinsic motivation, which is what is really doing the causal work. Requiring a student to take calculus does not magically make that student more intrinsically motivated. Indeed, requiring that student to take calculus might impose real costs—in terms of self esteem, motivation, or time spent on other activities—without any offsetting benefits.

Indeed, the exact mistake we just described was made in a peer reviewed scientific article. The researchers compared the college performance of people who did and did not take a variety of intensive high school math courses. Based on a positive correlation, they suggested that high school counselors, “use the results of this study to inform students and their parents and guardians of the important role that high school math courses play with regard to subsequent bachelor's degree completion.” That is, they mistake correlation for causation. Based on these correlations, they recommend that students who were not otherwise planning to do so should enroll in intensive math courses to increase their chances of graduating from college.

We'll return to the problem of mistaking correlation for causation in Part III of the book. For now, you should note that, although purported experts do it all the time, in general, it is wrong to infer causality from correlations.

Measuring Correlations

There are several common statistics that can be used to describe and measure the correlation between variables. Here we discuss three of them: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*. But before going through these three different ways of measuring correlations, we are going to learn about means and variances—two informative statistics that help us summarize and understand variables.

Mean and Variance

Let's focus on our Chicago crime and temperature data. Recall that in this data set, each observation is a day in 2018. And for each day we observe two variables, the number of reported crimes and the average temperature as measured in degree Fahrenheit at Midway Airport. We aren't going to reproduce the entire data set here, since it has 365 rows (one for each day of 2018). Figure 2.5 shows what the data looks like for the month of January. For the remainder of this discussion, we will treat the days of January, 2018 as our population of interest.

For any observation, i , call the value of the crime variable $crime_i$ and the value of the temperature variable $temperature_i$. In our data table, i can take any value from 1 through 31, corresponding to the 31 days of January, 2018. So, for instance, the temperature on January 13 was $temperature_{13} = 12.3$ and the number of crimes reported on January 24 was $crime_{24} = 610$.

Now let's calculate a few basic statistics that describe each of these two variables.

day	temperature	crimes
1	-2.7	847
2	-0.9	555
3	14.2	568
4	6.3	600
5	5.4	660
6	7.5	585
7	25.4	535
8	33.9	618
9	30.1	653
10	44.9	709
11	51.7	698
12	21.6	705
13	12.3	617
14	15.7	563
15	16.8	528
16	14.6	612
17	14.7	644
18	25.6	621
19	34.8	707
20	40.4	724
21	42.9	716
22	48.9	722
23	32.3	716
24	29.2	610
25	35.5	640
26	46.0	759
27	45.6	754
28	35.0	668
29	25.2	650
30	24.7	632
31	37.6	708
mean	26.3	655.6
standard deviation	15.1	73.1

Figure 2.5: Average temperature at Chicago Midway Airport and number of crimes reported in Chicago for each day of January, 2018.

First, we can calculate the *mean* of each variable, which we denote by μ (the Greek letter “mu”). The mean is just the average. We find it by summing the values of the observations and dividing by the number of observations. For January of 2018, the means of our two variables are:

$$\mu_{\text{crime}} = \frac{\sum_{i=1}^{31} \text{crime}_i}{31} = \frac{847 + 555 + \dots + 708}{31} = 655.6$$

and

$$\mu_{\text{temperature}} = \frac{\sum_{i=1}^{31} \text{temperature}_i}{31} = \frac{-2.7 + -0.9 + \dots + 37.6}{31} = 26.3$$

The $\sum$ symbols above (the upper-case Greek letter “Sigma”) are shorthand for summation. For example, $\sum_{i=1}^{31} \text{crime}_i$ means that we’re taking the values of crime for days 1 through 31 and add them all up.

A second statistic of interest is the *variance*, which we denote by σ^2 (the lower-case Greek letter “sigma”, squared). The variance is a way of measuring how far from the mean the individual values of the variable tend to be. You might say that the variance measures how variable the variable is.

Here’s how we calculate the variance. Suppose we have a population of size N (in our data, $N = 31$) and some variable X (like crime or temperature). For each observation, calculate its *deviation* from the mean. So, for observation i , that gives us $X_i - \mu_X$. On January 13, 2018 the deviation of the temperature from the mean temperature was $12.3 - 26.3 = -14$. That is, January 13, 2018 was fourteen degrees colder than the average day in January of 2018. By contrast, the deviation of January 23, 2018 was $32.3 - 26.3 = 6$. On January 23, it was six degrees warmer than on the average day in January of 2018.

Note that these deviations can be positive or negative. We are trying to measure how variable the observations are. For that purpose, it doesn’t matter whether any given ob-

ervation is positive or negative, we just want to know how far it is from the mean. So we need to transform the deviations into positive numbers that just measure the distance from the mean, rather than the sign and distance. To do this, we could look at the absolute value of the deviations. But absolute values are sometimes a little tricky to work with. So, instead, we square the deviations, which also gives us a positive number that measures how far each observation is from the mean. To get the variance, we find the average value of these squared deviations:

$$\sigma_X^2 = \frac{\sum_i^N (X_i - \mu_X)^2}{N}.$$

By focusing on the average of the squared deviations, rather than the average of the absolute value of the deviations, the variance is putting more weight on observations that are from the mean. If the richest person in society gets richer, this increases the variance in wealth more than if a moderately rich person gets richer by the same amount. For example, suppose the average wealth is 1. If someone with a wealth of 10 gains 1 more unit of wealth, the variance increases by $\frac{10^2 - 9^2}{N} = \frac{19}{N}$. But if someone with a wealth of 100 gains one more unit of wealth, the variance increases by $\frac{100^2 - 99^2}{N} = \frac{199}{N}$.

The variance is a fine measure of how variable a variable is. But, since we've squared everything, there is a sense in which it is not measured on the same scale as the variable itself. Sometimes we want a measure of variability that is on that same scale. When that is the case, we use the *standard deviation*, which is just the square-root of the variance. We denote the standard deviation by σ (the lower-case Greek letter “sigma”). It is

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum_i^N (X_i - \mu_X)^2}{N}}.$$

The standard deviation roughly corresponds to how far we expect observations to be from the mean, on average. Though, as we've noted, compared to the average absolute value of the deviations, it puts extra weight on observations that are farther from the mean.

For the two variables in our data, the standard deviations are:

$$\sigma_{\text{temperature}} = \sqrt{\frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}} = 15.1$$

and

$$\sigma_{\text{crime}} = \sqrt{\frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}} = 73.1.$$

Now that we understand what a mean, variance, and standard deviation are, we can discuss three important ways in which we measure correlations: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*.

Covariance

Suppose we have two variables, like crime and temperature, and we want to measure the covariance between them. To keep our notation simple, let's call those two variables X and Y . And let's assume we have a population of size N .

Here's how you calculate the covariance. For every observation, calculate the deviations—that is, how far the value of X is from the mean of X , μ_X , and how far the value of Y is from the mean of Y , μ_Y . Now, for each observation, multiply the two deviations together, so you have $(X_i - \mu_X)(Y_i - \mu_Y)$ for each observation i . We could call this the *product of the deviations*. Finally, to find the covariance of X and Y , calculate the average value of this product:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)}{N}.$$

Let's see that the covariance is a measure of the correlation. Consider a particularly strong version of positive correlation: suppose whenever X is bigger than average, Y is also bigger than average and whenever X is smaller than average, Y is also smaller than average. In this case, the product of the deviations will be positive for every observation—

either both deviations will be positive or both deviations will be negative. So the covariance will be positive. Now consider a particularly strong version of negative correlation: suppose whenever X is bigger than average, Y is smaller than average and whenever X is smaller than average, Y is bigger than average. In this case, the product of the deviations will be negative for every observation—one deviation is always negative and the other always positive. So the covariance will be negative. Of course, neither of these extreme cases has to hold. But if a larger-than-average X usually goes with a larger-than-average Y , then the covariance will be positive, reflecting a positive correlation. If a larger-than-average X usually goes with a smaller-than-average Y , then the covariance will be negative, reflecting a negative correlation. And if the values of X and Y are independent of each other, the covariance will be zero, reflecting the fact that the variables are uncorrelated.

Correlation coefficient

While the meaning of the sign of the covariance is clear, its magnitude can be a bit hard to interpret, since the product of the deviations depends on how variable the variables are. We can get a more easily interpretable statistic that still measures the correlation by accounting for the variance of the variables.

The *correlation coefficient* is simply the covariance divided by the product of the standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

When we divide the covariance by the product of the standard deviations, we are normalizing things. The covariance could, in principle, take any value, but the correlation coefficient, always takes a value between -1 and 1 . A value of 0 still indicates no correlation. A value of 1 indicates a positive correlation and perfect linear dependence—that is, if you made a scatter plot of the two variables, you could draw a straight, upward sloping line

through all the points. A value of -1 indicates a negative correlation and perfect linear dependence. A value between 0 and 1 indicates positive correlation but not a perfect linear relationship. And a value between -1 and 0 indicates negative correlation but not a perfect linear relationship.

The correlation coefficient is sometimes denoted by the letter r . And we also sometimes square the correlation coefficient to compute a statistic called r -squared or r^2 . This statistic always lies between 0 and 1.

One potentially attractive feature of the r^2 statistic is that it can be interpreted as a proportion. It's often interpreted as the proportion of the variation in Y "explained" by X , or equivalently, the proportion of X "explained" by Y . As we'll discuss in later chapters, the word explained can be misleading here. It doesn't mean that the variation in X causes the variation in Y or vice versa. It also doesn't account for the possibility that this observed correlation might have arisen by chance rather than reflecting a genuine phenomenon in the world.

Slope of the regression line

One potential concern with the correlation coefficient and the r^2 statistic is that they don't tell you anything substantive about the relationship between X and Y . Suppose our two variables of interest are crime and temperature in Chicago. A correlation coefficient of .8 tells us that there is a strong, positive relationship between the two variables, but it doesn't tell us what that relationship is. It could be that every degree of temperature corresponds with .1 extra crimes, or it could be that every degree of temperature corresponds with 100 extra crimes. Both are possible with a correlation coefficient of .8. But they mean very different things.

For this reason, we don't spend much time thinking about these ways of measuring

correlation. We typically focus on the slope of a line of best fit, as we've already shown you. Moreover, we tend to focus on one particular way of defining which line fits best. Remember, a line of best fit minimizes how far the data points are from the line on average. We typically measure how far the data points are from the line by calculating how far each point is from the line and then squaring that distance (so every value is positive, just like with squaring deviations). We focus on the line of best fit that minimizes the sum of these squared distances. This particular line of best fit is called the *regression line*. All the lines of best fit we drew earlier in this chapter were regression lines.

The slope of the regression line, it turns out, can be calculated from the covariance and variance. The slope of the regression line (also sometimes called the *regression coefficient*) when Y is on the vertical axis and X is on the horizontal axis is:

$$\frac{\text{cov}(X, Y)}{\sigma_X^2}.$$

This number tells us, descriptively, how much Y changes, on average, as X increases by one unit. Had we divided by σ_Y^2 instead of σ_X^2 , then it would tell us how much X changes, on average, as Y increases by one unit. As we've seen, those can be different numbers.

We'll spend a lot more time on regression lines in Chapters 5 and 10.

Populations and samples

Before moving on, there is one last issue that is worth pausing to highlight. We can think about each of the statistics we've talked about—the mean, the variance, the covariance, the correlation coefficient, the slope of the regression line—in two ways. There is a value of each of those statistics that corresponds to the whole population we are interested in. And there is a value of those statistics that corresponds to the sample of data we might happen to have. Either value can be of interest, but they can be importantly different. We

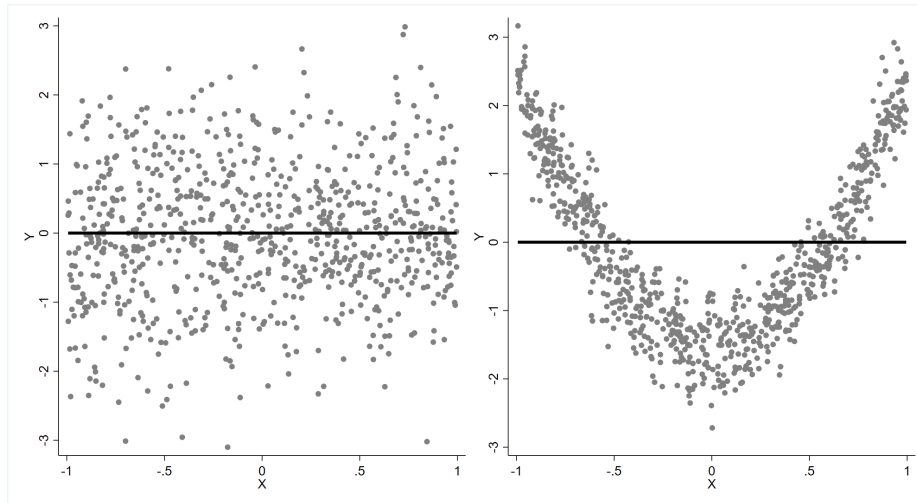


Figure 2.6: Zero correlation can mean many things.

have avoided that issue here by focusing on a case where our data and our population are the same—we have crime and temperature for every day in January of 2018, which we’ve treated as our population and our sample. But this won’t always be the case. We will revisit these issues in Chapter 6.

Straight Talk about Linearity

All of the various ways of measuring correlations that we have discussed focus on assessing linear relationships between variables. We will delve into this topic in more detail later on, especially in Chapter 5 when we return to the topic of age and voter turnout. But for now we will note that linear relationships are often interesting and important, but not all interesting and important relationships are linear. Consider, for example, the two possible relationships between variables X and Y illustrated in Figure 2.6.

As the regression lines make clear, in both these figures, the correlation between X and Y is 0. But these relationships are clearly different, just not in a way that is captured by

the regression line.

In the left panel, there is no correlation between X and Y and there also doesn't seem to be any interesting relationship of any kind. You really can't predict the value of Y from X or vice-versa. In the right panel, there is also no correlation between X and Y —on average, high values of X don't tend to occur with high values of Y nor do low values of X tend to occur with low values of Y . But there is certainly a relationship between these two variables. In fact, X is quite useful in predicting Y in the right panel. An important lesson is that clear thinking about data requires more than just computing correlations. Among other things, it is important to look at your data (e.g., with scatter plots like these), lest you miss interesting nonlinear relationships.

There are lots of statistical approaches for dealing with non-linearity, and we'll discuss some of them in this book. But, as it turns out, linear tools for describing data can still be useful, even when the variables are related in a nonlinear way. For instance, in the right panel of Figure 2.6, there is a strong negative correlation between X and Y when X is less than 0 and a strong positive correlation between X and Y when X is greater than 0. So one thing we could do with linear tools is to draw two lines of best fit, one for when X is less than 0 and one for when it is greater than zero. That would look like Figure 2.7.

Another thing we could do is to transform one of the variables so that the relationship looks more linear. For instance, in our example, although there is no correlation between Y and X , there is a strong linear relationship between Y and X^2 . In Figure 2.8 we plot X^2 on the horizontal axis and Y on the vertical axis. When we transform X into X^2 , negative values of X become positive values of X^2 (e.g., -1 becomes 1), while the positive values stay positive (e.g., 1 stays at 1). So it is as if we are folding the figure in on itself at $X = 0$, and then we're twisting and stretching it a little so that X becomes X^2 (0 stays at 0 , 1 stays at 1 , $.5$ becomes $.5^2 = .25$, and so on).

With this transformation, our regression line shows a strong positive relationship be-

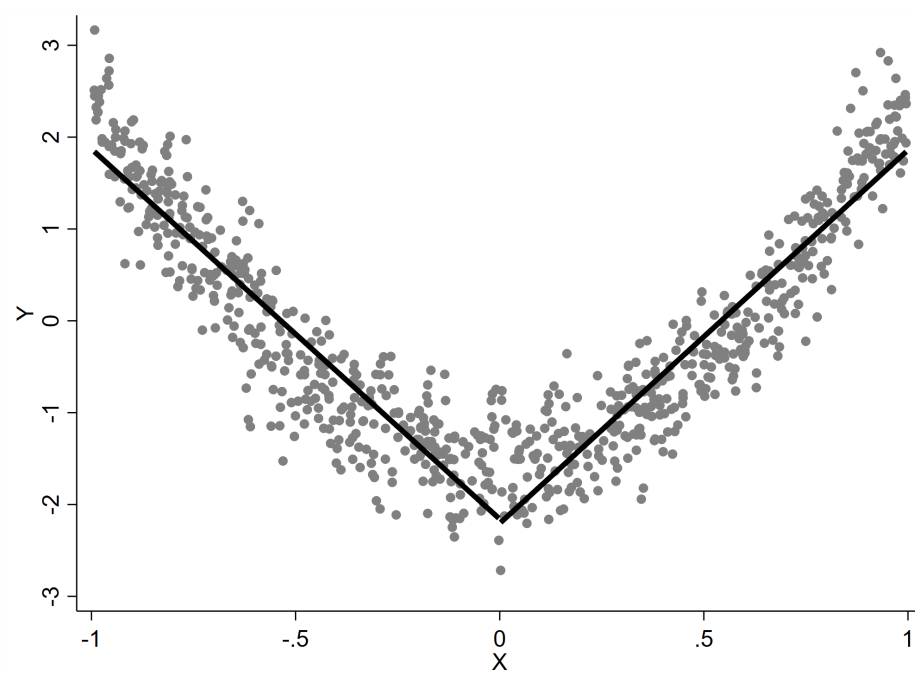


Figure 2.7: Fitting two separate regression lines to a non-linear relationship.

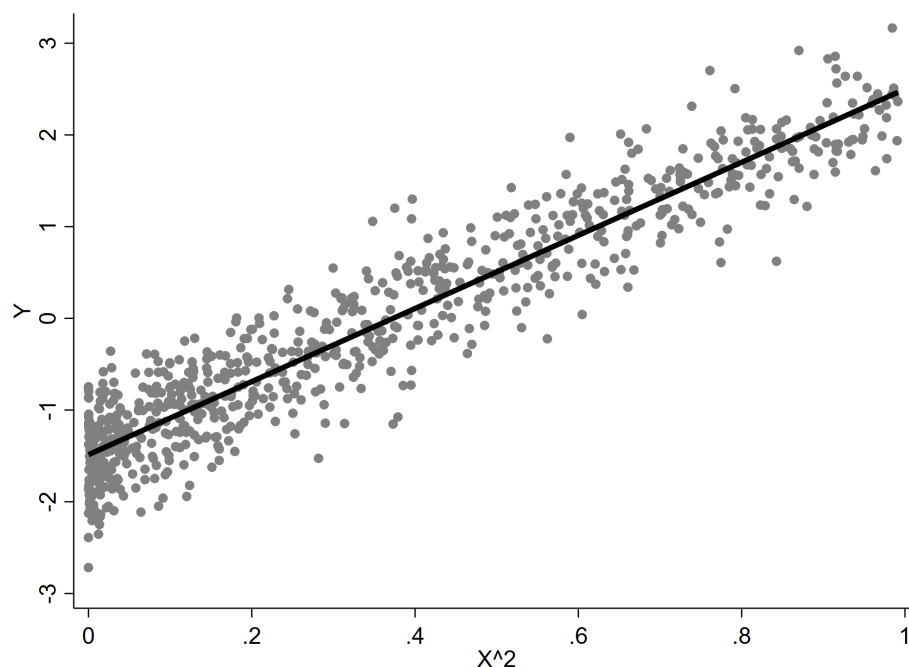


Figure 2.8: Creating a linear relationship by transforming a variable.

tween Y and X^2 and we can do a good job describing the relationship between these variables with our linear tools.

It's also worth pointing out that describing the relationship between two variables with a linear function is always appropriate when we're dealing with binary variables. For example, let's return to the correlation between oil production and autocracy. Figure 2.9 plots the data visually. The scatter plot is completely uninteresting because there are only 4 possible combinations of our two variables. Accordingly, all of the data points lie on one of those 4 dots. However, we can still estimate and plot the slope of the regression line. The slope of this line is simply the proportion of major oil producing countries that are autocracies minus the proportion of non-major oil producing countries that are autocracies. In other words, we learn the same thing from this picture that we learned from the table at the outset of the chapter.

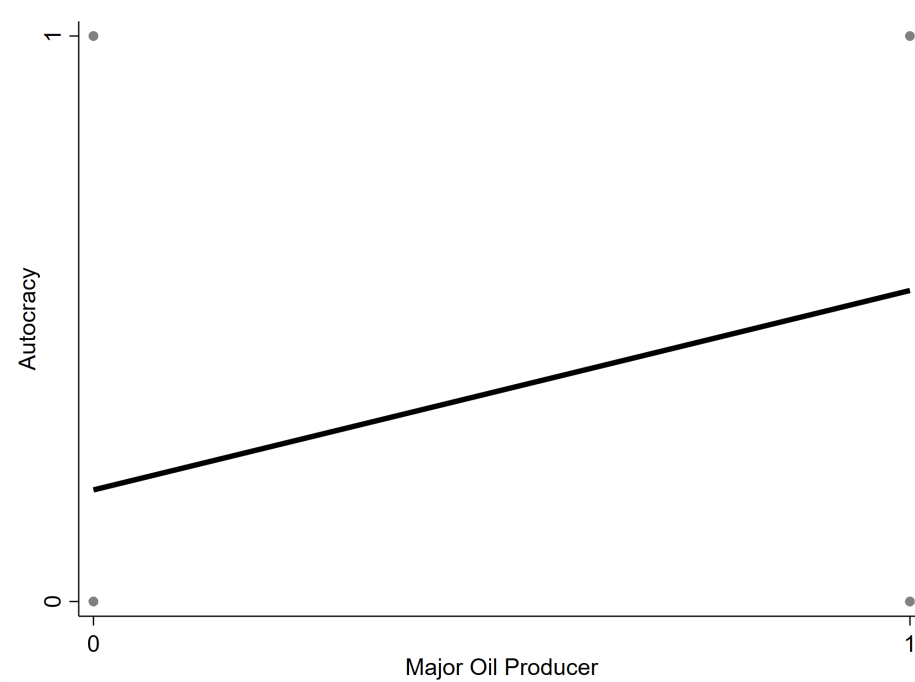


Figure 2.9: A regression line through data with a binary variable gives the difference in means.

One reason that we focus so much on linear relationships is that even nonlinear relationships tend to look approximately linear if you zoom in enough, that is, if you are interested in a sufficiently small range of values of the variable X . We must be particularly cautious about extrapolating when we zoom in like that. As we move further from the range of data where the relationship is approximately linear, our descriptions of the relationship (and, by extension, any predictions we make) will be less and less accurate.

To think more about the dangers of extrapolation, consider an example. Political analysts find that the incumbent party in U.S. presidential elections tends to get about 46 percent of the vote when there is 0 income growth, and an extra 3.5 percentage points of the vote for every percentage point increase in income growth. Of course, they've estimated this relationship using data on income growth levels that have actually occurred. Does this mean that we should predict incumbent vote share will be 81 percent if income growth is 10 percent? Probably not. And the incumbent's vote share definitely would not be 116 percent if income growth were 20 percent. That's impossible! But that doesn't mean a linear description of the data isn't useful for the range of income growths that we actually experience.

Key Terms

- **Correlation:** The correlation between two features of the world is the extent to which they tend to occur together.
- **Positively correlated:** When higher (lower) than average values of one variable tend to occur with higher (lower) than average values of another variable, we say that the two variables are positively correlated.
- **Negatively correlated:** When higher (lower) than average values of one variable tend to occur with lower (higher) than average values of another variable, we say that

the two variables are negatively correlated.

- **Uncorrelated:** When there is no correlation between two variables, meaning that higher (lower) than average values of one variable do not systematically coincide with higher or lower than average values of the other variable, we say that they are uncorrelated.
- **Line of Best Fit:** A line that minimizes how far data points are from the line on average, according to some measure of distance from data to the line.
- **Mean:** The average value of a variable.
- **Deviation from the mean:** The distance between an observation's value for some variable and the mean of that variable.
- **Variance:** A measure of how variable a variable is. It is the average of the square of the deviations from the mean.
- **Standard Deviation:** Another measure of how variable a variable is. The standard deviation is the square root of the variance. It has the advantage of being measured on the same scale as the variable itself and roughly corresponds to how far the typical observation is from the mean (though, like the variance, it over weights observations far from the mean).
- **Covariance:** A measure of the correlation between two variables. It is calculated as the average of the product of the deviations from the mean.
- **Correlation Coefficient:** Another measure of the correlation between two variables. It is calculated as the covariance divided by the product of the variances. The correlation coefficient takes a value between -1 and 1 , with -1 reflecting perfect linear negative dependence, 0 reflecting no correlation, and 1 reflecting perfect linear dependence.

- **r^2 :** The square of the correlation coefficient. It takes values between 0 and 1 and is often interpreted as the proportion of the variation in one variable “explained” by the other variable. But we have to pay careful attention to what we mean by “explained”. Importantly, it doesn’t mean that variation in one variable causes variation in the other.
- **Slope of the Regression Line:** The slope of a line of best fit drawn through the data. It tells us, descriptively, how the value of one variable changes, on average, we the other variable changes. Again, however, we are not entitled to interpret this relationship causally. The slope of the regression line is the covariance of two variables divided by the variance of one of them.

Exercises

1. The table below shows some data on which countries are major oil producers and which countries experienced a civil war between 1946 and 2004. Are being a major oil producer and experiencing civil war positive correlated, negatively correlated, or uncorrelated?

	Civil War	No Civil War
Oil Producer	7	12
Non-Oil Producer	55	94

Table 2.3: Oil production and civil war, 1946–2004

2. The table below provides data about height and income among American men, taken from the National Longitudinal Survey. It is fine to use a calculator for this question, but don’t use a spread sheet or statistical software to compute the answers.

- (a) Calculate the mean of each of these variables.
- (b) Calculate the variance of each of these variables.
- (c) Calculate the standard deviation of each of these variables.
- (d) Calculate the covariance between these two variables.
- (e) Calculate the correlation coefficient for these variables.
- (f) Are the two variables positively correlated, negatively correlated, or uncorrelated? How do you know?

Notes

See <https://www.systemicpeace.org/polity/polity4.htm> for more information about the Polity IV project that classifies countries as democratic vs. autocratic.

The study of advanced math and college completion is:

Trusty, Jerry and Spencer G. Niles. "High-School Math Courses and Completion of the Bachelor's Degree." *Professional School Counseling* 7, no. 2 (2003): 99–107.

If you are interested in examples of the growing use of forecasting and prediction in addressing important policy problems, have a look at

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5): 491–95.

Height in inches	Average Income
60	39,428
61	35,087
62	40,575
63	39,825
64	55,508
65	56,377
66	59,746
67	66,699
68	59,787
69	66,176
70	79,202
71	70,432
72	77,975
73	72,606
74	71,063
75	80,330

Table 2.4: Height and income of American men from the National Longitudinal Survey.

Chapter 3

Causation: What Is It and What Is It Good For?

What you'll learn

- A causal effect is a change in some feature of the world that would result from a change to some other feature of the world.
- Assessing causal relationships is crucial for policy- and decision-making.
- “*What effect did this have on the outcome?*” is a more coherent question than “*What caused the outcome?*”.
- Causal relationships are about comparing *counterfactual* worlds. As a result, they are fundamentally unobservable. But, in certain situations, we can learn about them from data.

As we've just seen, knowledge of correlations is useful for many purposes. Among the most important, but also most vexing, is for learning about causal relationships.

We make claims about causal knowledge all the time. I did poorly on the test because I didn't get enough sleep. Going to college will improve my future job prospects. A political candidate lost an election because of an attack ad. Violent crime is down because a new policing strategy took guns out of the hands of criminals.

Thinking clearly about how to learn whether a causal relationship exists is perhaps the most important conceptual challenge for learning to use information to make better decisions. This is because causal knowledge is the key to understanding how your decisions and actions affect the world around you. If you propose a new tax policy, test-prep strategy, diet plan, or advertising campaign, you're doing so not because you think it is correlated with better outcomes. Rather, you must believe that enacting your proposal will actually cause better outcomes.

Our goal, in this chapter, is to clarify exactly what we mean when we talk about causal relationships. Causality is a deep and perplexing topic to which much attention has been paid by scholars from many different fields. We won't be able to resolve all the thorny philosophical questions here. Instead we've set more modest goals. First, we want to make sure we are all on the same page by defining how we will use causal language for the duration of this book. Then we will explain why the notion of causality we adopt is a particularly valuable one. Finally, we will discuss some other approaches to talking about causality and explain why, from our point of view, they are less helpful than the one we adopt.

What is causation?

When we talk about causation, we're talking about the effect of one thing on another. In plain English:

A *causal effect* is a change in some feature of the world that would result from a change to some other feature of the world.

So, for instance, we would say that the tax rate has a causal effect on government revenue if changing the tax rate would lead to a change in government revenue.

We’ve defined the notion of an effect in plain English, so you might not have noticed that we’ve actually already slipped in a bit of philosophy. What do we mean by *would result*? After all, the world is as it is. Where did this change in some other feature of the world come from?

That’s a good question. In fact, our definition of a causal effect relies on a kind of thought experiment that we need to be explicit about. Let’s start with an example.

The movie star Gwyneth Paltrow runs a company called Goop that sells stickers, called Body Vibes, that are supposed to promote health, wellness, *and* good skin. Here’s what the Goop website says about Body Vibes:¹

Human bodies operate at an ideal energetic frequency, but everyday stresses and anxiety can throw off our internal balance, depleting our energy reserves and weakening our immune systems. Body Vibes stickers come pre-programmed to an ideal frequency, allowing them to target imbalances. While you’re wearing them—close to your heart, on your left shoulder or arm—they’ll fill in the deficiencies in your reserves, creating a calming effect, smoothing out both physical tension and anxiety. The founders, both aestheticians, also say they help clear skin by reducing inflammation and boosting cell turnover.

Suppose you paid the six dollars per sticker that Goop charges for Body Vibes because you really want clear skin. But then your friends started making fun of you for being a

¹<http://goop.com/wearable-stickers-that-promote-healing-really/> (accessed 9/15/17, we aren’t making this up).

sucker. In defending yourself, you'd want to claim that Body Vibes really do have an effect on the clarity of your skin. But what, exactly, would you mean by that claim?

Here's a way of thinking about this. Imagine an alternative world where, at the exact moment you went to stick on your Body Vibes (upper left shoulder, near the heart), unbeknownst to you, one of your friends replaced them with identical looking sticker that cost 10 cents instead of six dollars, but which hadn't been "pre-programmed to an ideal frequency." If your skin clarity would be different in that alternative world, then we would say that Body Vibes have an effect on skin clarity. If your skin clarity would be the same in that alternative world, we're afraid we'd have to conclude that Body Vibes don't have the claimed effect on skin clarity.

We can extend this thought experiment. There's nothing particularly special about the real world. Once we're already thinking about one alternative world, we might as well think about two. For instance, we could think about the effect of 10-cent stickers compared to magical crystals, even if you've never tried either of those approaches to skin care. We just have to compare two make-believe worlds: one where your friends secretly stuck stickers on your upper left shoulder near your heart, and another where they snuck crystals into your pockets.

We call these kinds of comparisons ***counterfactual*** thought experiments because at least one of the worlds we are comparing isn't the real, factual world—it's in our imaginations. We call the comparison of outcomes in such a thought experiment a ***counterfactual comparison***.

We can now make sense of the phrase *would result* in our definition of a causal effect: the change in some feature of the world that would result from a change to some other feature of the world. The *would result* refers to a counterfactual comparison between the outcome in the actual world and the outcome in a counterfactual world that is identical to the actual world up until the point where the feature of the world claimed to have a causal

effect is changed.

This idea of counterfactuals is a bit philosophically subtle. So, to help us make sure we are thinking clearly, we are going to introduce a framework for representing counterfactuals called the *potential outcomes framework*. Using the potential outcomes framework involves a little work mastering some notation. But it isn't too difficult. And once you do it, you will have a much deeper understanding of what causality really is. So let's give it a shot.

Potential Outcomes and Counterfactuals

We are interested in the effect of some *treatment* (say, Body Vibes) on some outcome (say, skin health). Let's call the treatment T . It is a binary variable, taking the value of 0 or 1. If $T = 1$ for some person, that means that person received the Body Vibes treatment. If $T = 0$ for some other person, that means that person didn't receive the Body Vibes treatment. We sometimes say that a unit (here, a person), with $T = 1$ is *treated* and a unit with $T = 0$ is *untreated*, although it's often arbitrary what we call treated and what we call untreated (e.g., we could just as easily talk about the effect of *not* wearing Body Vibes).

Similarly, let's refer to the outcome we are interested in as Y . In our example, Y describes a person's skin health. In a metaphysical sense, for each individual there is some level of skin health that they would have had if they'd used Body Vibes and some level of skin health they would have had if they hadn't used Body Vibes. These are that person's *potential outcomes*. However, at any given moment, we only ever get to observe one of these—each person either is or is not using Body Vibes at any given moment. Nonetheless, it is useful to have notation for both potential outcomes in order to think clearly about counterfactuals:

$$Y_{1i} = \text{outcome for unit } i \text{ if } T = 1$$

Y_{0i} = outcome for unit i if $T = 0$

The effect of wearing Body Vibes on person i 's skin health is just the difference in i 's skin health with and without Body Vibes. In our potential outcomes notation, it is:

$$\text{Effect of Body Vibes on } i\text{'s Skin Health} = Y_{1i} - Y_{0i}.$$

We say that causality is about counterfactual comparisons because we can only observe, at most, one of the two quantities, Y_{1i} or Y_{0i} , for any individual at any particular point in time. This means that we can't directly measure the effect of wearing Body Vibes on an individual's skin health. We suspect this fact is key to the success of Goop's business model.

What is causation good for?

As we've said, knowing about causal effects is super useful when deciding whether to take some action. In particular to weigh costs and benefits, you need to know how your action will affect whatever outcomes you care about.

Knowledge of causation is useful for making predictions about the consequences of an action that changes some feature of the world.

For instance, you can't possibly know if it is a good idea to spend money on a drug to treat heart disease, without knowing about causal effects—whether the drug reduces the risk of heart disease. The same goes for many decisions. When you are deciding whether or not to intervene in the world in some way—with a policy, an exercise plan, a parenting strategy, a new kind of online learning, or what have you—you want to know how that intervention *effects* the outcomes you care about.

Conceptual issues

We like this notion of causality based on counterfactual comparisons. We find it both intellectually clarifying and pragmatically useful for thinking clearly about real world problems. We hope you are fans as well. That said, there are ways in which this notion of causality creates some conceptual issues that might leave you a bit disappointed. But we think that, once you think through the issues clearly, you'll agree that the pain of disappointment is well worth the benefit in terms of conceptual clarity that comes from the counterfactual approach to causality. But let's talk it through.

What is the cause?

The first disappointment is that some of the causal questions that people are accustomed to asking turn out to be a bit hard to talk about with the counterfactual notion of causality. Think of questions like the following: Why did housing prices tank during the latest financial crisis? Why did the Blackhawks win the Stanley Cup? What caused World War I? Questions of causal attribution, like these, are common. But when causation is defined in terms of counterfactual comparisons, they don't make a ton of sense.

Let's think about World War I. A common claim is that World War I was caused by the assassination of Archduke Ferdinand, the heir to the throne of the Austria-Hungary, by Serbian and Bosnian nationalists. The assassins were part of an irredentist movement pushing Serbia to take control over all of the southern Balkans, including Bosnia and Herzegovina, which Austria-Hungary had annexed in 1908. The government of Austria-Hungary responded to the assassination with the July Ultimatum, a list of demands so onerous they were certain to be rejected by the Serbian government. When the ultimatum was rejected, Austria-Hungary declared war causing Russia to mobilize its army to defend Serbia. In response, Germany (an ally of Austria-Hungary) declared war on Russia, France (an ally of

Russia) declared war on Germany, and the whole mess cascaded into World War I. Thus, the claim goes, the assassination of Archduke Ferdinand caused World War I.

Now, there is a sense in which this claim is perfectly simple to think about in our framework. We can ask, in the counterfactual world in which Ferdinand was not assassinated would World War I still have occurred? If in fact World War I would not have occurred in that counterfactual world, then it seems right to say that the assassination had an effect on war breaking out. But that is a far cry from saying that the assassination of the archduke was *the* cause of the war. Surely, there are many factors that, had they been different, would have prevented World War I from being fought? Sure, had Archduke Ferdinand not been assassinated, maybe the war wouldn't have been fought. But also, had Austria-Hungary not annexed Bosnia and Herzegovina, perhaps Ferdinand would have never been assassinated and the war would have never been fought, so the annexation was just as much a cause as the assassination. Similarly, had the Serbian government complied with the July Ultimatum, perhaps the war would have been avoided, so the noncompliance with the ultimatum was also a cause. And to further illustrate how many such causes there are, had some fish-like creature in the Palaeozoic Era swam left instead of right, perhaps the human race as we know it would not exist, and again, World War I would have never been fought. Or, to take an example with some historical gravitas, the 17th century French mathematician, Blaise Pascal, reflecting on Antony's attraction to a long proboscis, quipped, "Cleopatra's nose, had it been shorter, the whole face of the world would have been changed." This led James Fearon, in an essay on counterfactual reasoning, to ask, "Does this imply that the gene controlling the length of Cleopatra's nose was a cause of World War I?" As you can see, then, the problem isn't that it is false that the assassination of Archduke Ferdinand caused World War I. Rather, since so many factors appear to have caused World War I, talk of the war's cause seems almost beside the point.

Once we start thinking about counterfactuals, it becomes pretty clear that things have

lots of causes. That makes it hard to answer “what is *the* cause”-questions. Instead, it pushes us to answer “was this a cause”-questions or “did this have an effect”-questions. This is perhaps disappointing.

One thought you might have, in response, is that surely some causes of a phenomenon are more important or more proximate than others. If that is true, perhaps we can still talk about the *important* or the *proximate* causes of World War I. How might we do this?

An approach that some philosophers advocate goes something like this. Imagine all the counterfactual worlds in which World War I did not occur. Some of these counterfactual worlds are very different from the actual world—e.g., World War I probably doesn’t occur in many counterfactual worlds in which there is no gravity. Others are quite similar to the actual world—perhaps World War I doesn’t occur in a world identical to ours except that Archduke Ferdinand overslept on June 28. We learn about the proximate causes of World War I by comparing the actual world to the counterfactual world in which World War I did not occur that is most similar to the actual world. This kind of analysis may allow us to give reasonable-sounding answers to “what is *the* cause”-questions without abandoning our definition of causation based on counterfactual comparisons. For instance, it seems reasonable to think that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra’s nose, the laws of gravity, or the whims of Paleozoic fish.

There is certainly something to this approach. But, that said, it is often hard to adjudicate between the importance or proximity of one cause versus another in a principled way. If you know a bit of history, you surely can come up with other causes that seem equally proximate. For instance, many scholars have argued that early-20th century military doctrines favoring offensive over defensive strategies played a role in causing World War I. Is the world in which a slightly different military doctrine was adopted more proximate to our world than the one in which Archduke Ferdinand was not assassinated? For that matter, is the world in which one Paleozoic fish took a different turn really so far from our world?

It's hard to say.

To see the problem in a somewhat less lofty and distant setting, consider a basketball game in which the Chicago Bulls are trailing the Los Angeles Lakers by 1 point. The Bulls have just enough time left to take one final shot. They make it, winning the game by one point. Most likely, commentators will fixate on that last shot, and they may even say that the last shot was *the* reason the Bulls won. But think about this counterfactually for a moment. Dozens of shots throughout the game were pivotal. Plausibly, every shot the Bulls made was pivotal—in a counterfactual world in which they missed that shot and everything else played out as it did in the actual world, they would have lost instead of won. Similarly, every shot the Lakers missed was pivotal—in a counterfactual world in which they made it and everything else played out as it actually did, they would have won instead. So what's so special about that last shot? One possibility is that everyone knew that the final shot would be pivotal when it was taken. But very few causes meet this criterion, certainly not the assassination of Archduke Ferdinand. So, in our view, there is no obvious reason to think that the last shot was a more important cause of the Bulls' victory than the other shots. Instead, we think this example illustrates a basic, if frustrating, fact of life: individual events can have many equally consequential causes.

Another surprising fact about the counterfactual approach is that, at least in principle, it's possible for some event to have no causes at all. Suppose that the authors of this book concoct the perfect crime. We both shoot and kill our sworn enemy at the same time, knowing that either bullet would be fatal on its own. When questioned, Anthony says, "Clearly, I can't be charged with a crime. My actions had no effect whatsoever. Had I not fired my gun, the victim would still have died." And similarly, Ethan retorts, "I could not have possibly caused the victim's death either. Had I not shot my gun, he would have still died." While the American justice system might not be impressed by our defense, the counterfactual logic is solid. Some events may be the result of a confluence of factors

whereby no single factor could have changed the outcome. This theoretical possibility is yet another reason that it might not make much sense to ask questions like “what caused World War I?” It could well be that, for all the factors we like to talk about, taking away any one of them would in fact not have sufficed to prevent the war.

The fundamental problem of causal inference

A second potentially disappointing aspect of causal effects as we’ve defined them is that you can never, ever directly observe them. This fact is called the *fundamental problem of causal inference*. Let’s see exactly why and what it implies for our ability to learn about causality.

The effect of going to college on your income is the difference in your income in a world in which you go to college vs. a world in which you are the same up until the college decision, but in which you don’t go to college. At least one of those worlds is counterfactual. You can’t both go to college and not go to college. That is, you have two potential outcomes— Y_{college} and $Y_{\text{no college}}$. But you have only one *actual* outcome, either you went to college or you didn’t. Given this, we can never observe the effect of going to college on your income since we only observe your income in the actual world, not the counterfactual world.

The fundamental problem of causal inference, then, is that, at any given time, we only observe any given unit of analysis (e.g., a person, basketball team, or country) in one state of affairs. So we can’t observe the effect on that unit of being in that state of affairs vs. some other state of affairs, because all the other states of affairs are counterfactual. We can’t know $Y_{\text{college}} - Y_{\text{no college}}$ for you, because we only observe one of the two values.

To return to our World War I example, when we started talking about the counterfactual world in which the Archduke was not assassinated, you might have been thinking, “How could we possibly know what would have happened had the Archduke not been assassinated?”

Maybe no war would have occurred. Or maybe Austria-Hungary and Germany would have found some other pretense for a war they wanted to fight anyway.” That’s the fundamental problem of causal inference. We can’t know.

So how do we make progress on answering causal questions if effects are fundamentally unobservable? Fortunately, there are lots of situations where we don’t necessarily need to know the effect for every individual unit of analysis. Instead, we want to know the average effect across lots of individuals.

Suppose, for instance, that the Food and Drug Administration (FDA) is deciding whether to approve a new drug. To learn about the health effects of the drug, scientists conduct a randomized trial, assigning some people to take the drug (the treated group) and other people to take a placebo (the untreated group). Because of the fundamental problem of causal inference, the scientists can’t observe the effect of taking the drug on any individual. Each person is either taking the drug or not. But by comparing the average health outcomes for people in the untreated group to the average health outcomes for people in the treated group, they can estimate the average effect of the drug. (We’ll talk a lot more about how this works in Parts II and III.) Doing so allows the scientists to answer what turns out to be the key causal question for the FDA’s decision: if we approve the new drug, how will health change in the population on average?

There are many settings in which knowledge about average effects is all you need. But there are some settings where this is not the case and the fundamental problem of causal inference constitutes a real challenge. For instance, some of you may be or become lawyers and judges. You’ll be concerned with applying the “but-for” test to assess liability. That will require answering questions like, “Would a harm to Anthony not have happened but for Ethan’s actions?” The fundamental problem of causal inference says we can never know for sure, since the world in which Ethan did not take his action is counterfactual, so we don’t know what happens to Anthony in that world. Instead, what we’ve just said, and will

cover in much more detail in the rest of the book, is that there are methods for answering a slightly different question like “on average, when people take actions of the sort Ethan took, does it tend to cause harm to other people?” A convincing answer to that latter question may or may not be compelling in a court that wants to answer the former.

It is our hope that this book will give you a useful framework for starting to think clearly about causal questions. Unfortunately, part of that clear thinking involves admitting that sometimes we cannot answer certain questions with confidence, even when those questions are very important.

Causality and counterexamples

One common skeptical reaction to evidence showing the existence of an average effect is to point to counter-examples. Perhaps you’ve had an experience like the following at a family gathering. You read a study showing that, on average, flu shots reduce the risk of contracting flu. You mention this over Thanksgiving dinner, encouraging your loved ones to get the vaccine. But your vaccine-skeptic relative says, “I don’t know, I got the flu shot last year and I still got the flu.” Many people nod and agree, perhaps pointing out that their friend so-and-so also got the flu shot and still got sick.

The intuition behind this kind of objection-by-way-of-counterexample is something like: “If flu shots really prevent the flu, then no one who got a flu shot would get the flu. Thus, my one counterexample means the vaccine doesn’t work. Also, I don’t like shots, so I’ll just stick with using Body Vibes from Goop to prevent flu.”

This is not a good argument. The evidence says that the flu shot causes flu risk to go down averaging across lots of people, each with their unique biology, level of flu exposure, environment, and so on. It doesn’t say that it eliminated flu risk for each and every individual. Indeed, because of the fundamental problem of causal inference, it can’t say

that. But to get flu risk to go down on average the flu shot must have prevented the flu (i.e., had a causal effect) for at least some people. We just don't know exactly which ones experienced the effect.

Let's think about this in our potential outcomes notation. Think of the potential outcomes as whether or not you get the flu. We'll say $Y = 1$ if you got the flu and $Y = 0$ if you didn't. And think of the treatment as whether you got the flu shot, with $T = 1$ meaning you got the shot and $T = 0$ meaning you didn't.

Maybe there are three different kinds of people—call them the *never sick*, the *always sick*, and the *vaccine responders*. The never sick and the always sick have potential outcomes that don't respond to treatment. The never sick don't get the flu and the always sick get the flu, regardless of whether they get the flu shot. In our notation:

$$Y_{1,\text{never sick}} = 0 \quad Y_{0,\text{never sick}} = 0$$

and

$$Y_{1,\text{always sick}} = 1 \quad Y_{0,\text{always sick}} = 1.$$

But the vaccine responders are different, they get the flu if they don't get the shot and don't get the flu if they do get the shot:

$$Y_{1,\text{vaccine responder}} = 0 \quad Y_{0,\text{vaccine responder}} = 1.$$

In a population made up of these three groups of people, the flu shot reduces flu on average. You don't know which group you are in. But there is a chance you are a vaccine responder. So, while there is a chance that the flu shot won't do anything for you, getting a flu shot still reduces your risk of getting sick.

Importantly, pointing to one counterexample is neither here nor there with respect to

such evidence. Perhaps your unlucky relative was one of the people whose confluence of circumstances were such that the flu shot didn't have an effect (i.e., they were an always sick). That doesn't mean it didn't have an effect for lots of other people. And it doesn't even mean that the flu shot won't prevent the flu for that same relative next year or that it won't help you. Absent any further information about which group they are in, any individual's best guess is that the flu shot will reduce their chances of contracting the flu since it does so on average.

Of course, the possibility that effects are different for different people presents another set of important conceptual challenges. We might be able to detect such heterogeneous effects, especially if they correspond with fairly straightforward categories (e.g., men versus women, old versus young, healthy versus sick). To identify such heterogeneous effects, we could run a separate experiment for each group, which would tell us the average effect for each group, rather than for the population as a whole. But what if effects differ across people for complicated or obscure reasons that might never occur to us? Then, when we go to look at the effect of some intervention, it is very important to keep in mind that we are estimating an average effect. Some people may have effects much larger than the average. Others may have effects much smaller than the average. Indeed, some people may have no effect at all or an effect in the opposite direction from the average. If we don't know the source of this heterogeneity, all we will be able to say is something about the average which, as we've already discussed, may still be valuable.

Causality and the law

One place where these philosophical questions become of serious practical import is in the law. Administering justice requires assigning blame and assessing liability. If we want to know whether, say, Ethan should be held liable for some harm suffered by Anthony, surely we need to know whether Ethan's actions caused that harm. But, as we've just discussed,

talking about causes in this way is, at best, conceptually fraught. Many things, from the behavior of a Paleozoic fish to Ethan's alleged negligence, may have had a causal effect on the harm Anthony suffered. Is the fish liable too?

The law is aware of the philosophical conundrum. But it must ultimately come up with some pragmatic resolution that allows judges and lawyers to get on with the business of administering justice. Here's, roughly, where it comes down.

In the Common Law, causality is thought of in terms of two conditions that are closely related to things we've talked about. These are referred to as *cause-in-fact* and *proximate causality*.

Cause-in-fact is essentially counter-factual causality. Whether Ethan's actions are a cause-in-fact of Anthony's suffering is determined by a "but for" test. Ethan's actions are a cause-in-fact of Anthony's suffering if Anthony wouldn't have suffered *but for* Ethan's actions.

Of course, as you already know, a counter-factual standard like the "but for" test isn't very stringent. World War I wouldn't have happened, but for a Paleozoic fish turning the wrong direction. Does that mean we should hold the poor fish liable for World War I?

The law's answer is no. The fish is off the hook, so to speak. This is where proximity comes in. For there to be liability, the law requires that some cause-in-fact be close enough in the causal chain. This thought is also familiar, for instance from our argument that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra's nose.

So an assessment of legal causality might go something like this. Suppose Ethan chairs a search committee that hires a new professor whose brother comes to visit, gets drunk, and crashes into Anthony's car. Is Ethan liable for Anthony's suffering? It is plausible that, but for decisions Ethan made a search committee chair, Anthony's car would not have been

hit. So Ethan's actions are probably a cause-in-fact of Anthony's suffering. But there are many steps in the causal chain between Ethan's actions and the car crash, all of which are out of Ethan's hands. So the law would not find Ethan liable for the damage to Anthony's car.

Of course, as we've discussed, knowing exactly how to apply either of these criteria is tricky. To apply the but-for test, we have to know what the right counter-factual world is. And defining how close is close-enough for a proximity test is a fraught problem, full of judgment calls. All of which is to say that these questions about causality are vexing and of great practical importance.

Causality works in surprising ways

Once you understand the relationship between counterfactual comparisons and causality, you can see that causality can work in some ways that might seem surprising. Consider a few examples.

Causality can run backwards in time

One common intuition is that causality must run forwards in time. That is, an event that happens now can have an effect on events that happen in the future. But surely, the thought goes, events that happen in the future can't affect events in the past. After all, the events in the future haven't happened yet.

Let's check this intuition by thinking about Christmas. Here's a correlation in the world: the number of Christmas cards that get mailed in a given week is strongly correlated with it being one of the three weeks prior to Christmas. That is, many more Christmas cards are mailed in the three weeks before Christmas than in any other week of the year.

Now, although correlation need not imply causation, we suspect that there is a causal

relationship here. But that causal relationship does not run forward in time. Sending Christmas cards does not affect whether Christmas occurs. In a counterfactual world in which greeting cards cease to exist, Christmas will still occur on December 25. Instead, the causal relationship runs backward in time. Christmas exerts an effect on the sending of Christmas cards. In the counterfactual world in which Christmas will not exist on December 25, fewer Christmas cards will be sent in the first three weeks of December. Thus, on our counterfactual definition, Christmas exerts a causal effect on Christmas cards. Causality can run backwards in time.

Causality does not require physical connection

Another standard intuition is that causation necessarily has to do with physical connection. One billiard ball affects another by bumping into it. Maybe such physical connection always underlies causal relationships.

While, of course, there are many examples of causal effects that occur through physical connection, such physical connection is not required. Think of a person who is deterred from robbing a store by worry over imprisonment. Such a person's behavior is affected by the existence of the police, the courts, the penal code, and the prison system. Each of these features of the world has an effect on whether this person commits a crime, even though there is no physical connection between them. Indeed, think of our previous discussion of the effect of Christmas on the sending of Christmas cards. Christmas isn't a physical thing in the world at all. It is hard to see even what it would mean for the causal relationship between Christmas and the sending of Christmas cards to occur through physical connection.

A defender of "physicalism" might say that with enough creativity, we can describe the effect of the police on a criminal's behavior in purely physical terms. Perhaps the police arrested other criminals in the past, leading reporters to write about this activity in newspapers, leading the criminal in question to read about these arrests in the newspaper,

detering him from committing the crime. If you are willing to really tie yourself into knots, you could do a similar exercise for Christmas and Christmas cards. But this all seems rather strained. The important point is that counterfactually defined causal relationships surely do not depend on anything like the simple, common sense kind of physical connections suggested by the billiard ball example.

Causation need not imply correlation

We've agreed that correlation need not imply causation. But, perhaps more surprisingly, causation also need not imply correlation and certainly not correlation in the "expected" direction. There are many situations in which some feature of the world has (say) a *negative* effect on some other feature of the world, but those two features of the world are *positively* correlated (or vice versa).

You'd probably find a strong, positive correlation between the number of firefighters who have recently visited a house and the amount of fire damage to that house. But if we had to guess, we'd suspect that firefighters, on average, reduce fire damage. In other words, if fewer firefighters had visited, we suspect there would be even more fire damage.

So why is the correlation positive? Firefighters tend to visit houses that are on fire. So, although firefighters reduce fire damage to some degree, the houses that have been visited by firefighters tend to have more fire damage. Hence, not only should one not conclude from a correlation that there must be a causal relationship. One should not assume that, just because a causal relationship exists, the correlations found in the world will correspond to those causal relationships in some straightforward way.

Key Terms

- **Causal Effect:** Informally, the change in some feature of the world that would result from a change to some other feature of the world. Formally, the difference in the potential outcomes for some unit under two different treatment statuses.
- **Body Vibes:** Stickers that a company called Goop claims improve your skin.
- **Counterfactual Comparison:** A comparison that compares things in two different worlds or states of affairs, at least one of which does not actually exist.
- **Treatment:** Terminology we use to describe any intervention in the world. We usually use this terminology when we are thinking about the causal effect of the treatment, so we want to know what happens with and without the treatment. Importantly, although it sounds like medical terminology, treatment as we use it can refer to *anything* that happens in the world that might have an effect on something else.
- **Potential Outcomes Framework:** A mathematical framework for representing counterfactuals. The potential outcome for some unit under some treatment is the outcome that unit would experience under that (potentially counterfactual) treatment.
- **Fundamental Problem of Causal Inference:** This refers to the fact that, since we only observe any given unit in one treatment status at any one time, we can never directly observe the causal effect of a treatment.

Exercises

1. Sarah says that she is hungry. John hands her a piece of pizza. Sarah eats the pizza and then declares that she is no longer hungry.

- (a) The fundamental problem of causal inference seems to say that you can't know that Sarah eating the pizza had a causal effect on her no longer being hungry. Is that right? Explain.
 - (b) Do you think you nonetheless have good reasons to believe that eating the pizza had an effect on Sarah no longer being hungry? Explain why or why not.
 - (c) Do you have good reasons for believing that John handing Sarah the pizza had a causal effect on her no longer being hungry? In your assessment, are the reasons to believe John's actions had a causal effect better or worse than the reasons to believe Sarah eating the piece of pizza had a causal effect?
2. A government is considering making alcohol consumption illegal as part of a public health campaign. Let's think of making alcohol illegal as the treatment, T . Write $T = 1$ if the government makes alcohol illegal and $T = 0$ if the government leaves alcohol legal.

We will think of a binary outcome for each person, either they drink alcohol or they do not. If person i drinks at treatment status T , we write her potential outcome as $Y_{Ti} = 1$ and if she doesn't drink, we write it as $Y_{Ti} = 0$.

Suppose the society is made up of three groups: the always drinkers, the legal drinkers, and the never drinkers. The always drinkers will drink whether or not alcohol is legal. The legal drinkers will drink if and only if alcohol is legal. The never drinkers won't drink whether or not alcohol is legal.

- (a) Write down, in potential outcomes notation and as a number, each of the two potential outcomes for each of the three groups.
- (b) Write down, in both potential outcomes notation and as a number, the causal effect of making alcohol illegal on drinking for each of the three groups.
- (c) Is there an effect, on average, of banning alcohol in this society?

- (d) Suppose you are out to lunch with some friends and one of them says, “My uncle lives in a place where they banned alcohol and all of his friends kept drinking, so I don’t think it does anything.” Explain, in terms of our example, why this isn’t a convincing argument.

Notes

The quote from Blaise Pascal on Cleopatra’s nose is from his 17th century collection entitled *Pensées*.

The essay about counterfactual reasoning discussing the gene controlling the length of Cleopatra’s nose is

James D. Fearon 2011. “Counterfactuals and Hypothesis Testing in Political Science.” *World Politics* 43(2):169-195.

If you’d like to read more about the counterfactual definition of causality, potential outcomes, and surrounding discussions and debates have a look at:

David Lewis. 1973. Causation. *Journal of Philosophy* 70: 556–67.

Paul W. Holland. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396):945–960.

Stephen Mumford and Rani Lill Anjum. 2014. *Causality: A Very Short Introduction*. Oxford University Press.

There is also a nice entry by Peter Menzies and Helen Beebe in the Stanford Encyclopedia of Philosophy: <https://plato.stanford.edu/entries/causation-counterfactual/>.

Part II

Does a Relationship Exist?

Chapter 4

Correlation Requires Variation

What you'll learn:

- You can't learn about a correlation without variation in both variables of interest.
- In many realms of life—from education to medicine to rocket science—people fall into the trap of trying to make claims about correlations without such variation.
- Many institutional procedures push us to make this mistake without noticing it.
- We'll show you a simple way to check for this mistake in your own thinking or that of others.

In Chapter 2 we discussed the idea that the correlation between two features of the world is the extent to which they tend to occur together. We opened our discussion of correlation by thinking about whether oil production and autocracy are correlated. To figure this out we looked at the country-level data represened in Table 4.1.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

Table 4.1: Oil Production and Type of Government.

Determining whether there is a correlation between oil production and autocracy involved comparing the percentage of major oil producers that are autocracies to the percentage of countries that aren't major oil producers that are autocracies. To make this comparison, we needed four pieces of information: the number of major oil producers that are autocracies, the number of major oil producers that are democracies, the number of countries that aren't major oil producers that are autocracies, and the number of countries that are not major oil producers that are democracies. Had we been lacking any of these pieces of information, we would not have been able to figure out whether oil production and autocracy are correlated.

To see why, suppose we didn't know that the number of major oil producers that are democracies. (Of course, we'd also have to not know the total number of countries, so we couldn't just back-out the 9 by subtracting the number of countries in the other three categories from the total number of countries.) We still know that about 20 percent (29/147) of countries that aren't major oil producers are autocracies. But now we can't figure out

what proportion of the major oil producers are autocracies. It could be anything. If the number of democracies that are major oil producers turned out to be (say) 11, then 50 percent ($11/22$) of major oil producers would be autocracies and there would be a strong positive correlation. If the number of democracies that are major oil producers turned out to be (say) 99, then only 10 percent ($11/110$) of major oil producers would be autocracies, so there would be a strong negative correlation. If the number of democracies that are major oil producers turned out to be 44, then 20 percent ($11/55$) of major oil producers would be autocracies—the same as for countries that are not major oil producers—and there would be no correlation at all. So, just as we saw in our discussion of governors and fraud in Chapter 2, we need to observe all four pieces of information to figure out the correlation.

This is what we mean when we say that correlation requires variation. If you want to figure out whether two variables are correlated, you have to observe variation in both of them. You must observe the number of countries that are and are not major oil producers. And you must observe the number of autocracies and democracies in each group. Just observing variation in one or the other variable is not enough. In Chapter 2, when we asked which of five factual statements described a correlation, the problem with the three statements that did not was a lack of variation in one of the variables.

It may seem obvious, based on our simple binary example, that correlation requires variation. In our experience, it is anything but. Indeed, failing to look for variation in one or another variable while trying to establish a correlation is an exceptionally common mistake.

In this chapter, we explore this mistake and try to unpack why it is so common. Broadly, we think there are two, closely related, reasons that people so frequently try to establish a correlation without variation.

The first reason is called *selecting on the dependent variable*. If you want to forecast or explain some phenomenon, it is a natural impulse to start by examining previous instances

where that phenomenon occurred. But if you look only at instances where the phenomenon occurred, you are trying to assess a correlation without variation, since you have no variation in whether or not the phenomenon occurred. This is like looking for correlates of autocracy without examining any democracies. It won't work.

The phrase *dependent variable* refers to the variable representing the phenomenon you are trying to forecast or explain. This mistake is referred to as selecting on the dependent variable because you are selecting which cases to look at based on the value of the dependent variable (e.g., only looking at autocracies), rather than looking at variation in the dependent variable (e.g., comparing autocracies and democracies).

The second reason is that the world is often organized in ways that push us to make this mistake. For instance, in many situations, specialists are only asked to weigh in when things are going wrong. Those specialists end up not comparing situations where things are going wrong to situations where things are going right simply because, when things are going right, no one thinks to ask for their input.

Certain kinds of medical care present classic examples of this problem. By and large, people only go to the doctor when they are ill. So doctors may have scant opportunity to compare the ill to the well, which can sometimes lead them to mistaken conclusions about the correlation between symptoms and diagnoses. We'll see an example of this later in the chapter.

Institutional procedures are also designed to make this mistake. Organizations frequently have mandatory self-studies or post-mortems to try to figure out what went wrong after a disaster. But those involved seldom think to compare the disastrous case to cases where things went alright. This sets the organization up to make the mistake of looking for correlation without variation, without anyone ever noticing.

This chapter, more than most other chapters in the book, is built around examples. We

do this for a reason. In our experience, once we explain that correlation requires variation, people tend to nod their head in agreement, appearing to understand. Indeed, because the point seems obvious when put in plain English, many people are skeptical that this could be such a big problem. And, yet, they themselves go right back to making the same mistake. We hope that by showing you lots of examples of very smart people making this mistake in high stakes environments, we will convince you that this is a real problem and that avoiding this error requires clear thinking and genuine effort and concentration.

Selecting on the Dependent Variable

As we've said, if you want to forecast or explain some phenomenon, it is a natural impulse to start by examining previous instances of that phenomenon occurring. Consider a few examples. Following the financial crisis of 2008, both scholars and journalists who wanted to understand how to predict future financial crises invested enormous time and energy examining the historic record to look for patterns in previous crises. Malcolm Gladwell, in his book *Outliers*, tries to understand the correlates of personal success by recounting the lives of highly accomplished people, looking for similarities. Congress, considering a change to American counterinsurgency strategy in Afghanistan, heard testimony on the correlates of suicide terrorism from an academic expert who had done an exhaustive study of all suicide terrorist campaigns since 1980 looking for shared characteristics.

Natural as it seems to look for commonalities in past instances of events you want to forecast, it won't work. None of the studies just described provide evidence about the correlates of the phenomena with which they are concerned. And this is because each is looking for correlation without variation.

The claim that we can't learn about the correlates of suicide terrorism or financial crises by looking for commonalities among historic cases of similar events may seem counterintu-

itive. But, since we know that correlation requires variation, the mistake is actually quite simple to grasp. Put in the terms of our earlier example, each of these examples is analogous to looking for correlates of oil production without any data on non-oil-producing countries!

To see the key conceptual flaw in all of these arguments in another way, let's start by considering the central claim in Malcolm Gladwell's *Outliers*, the so-called *10,000-hour rule*.

The 10,000 hour rule

Gladwell's idea is that it takes about 10,000 hours of serious practice to master any difficult skill. As such, he argues, a key correlate of being a great achiever is having the will and the opportunity to put in those 10,000 hours. Talent might matter too, but first and foremost, if you are looking for a great achiever, look for someone who was in the right place at the right time to get that 10,000 hours of practice.

Now, of course, Gladwell isn't just interested in forecasting great success. He thinks the 10,000 hour rule might be causal. If true, this would have far reaching consequences. Given the right conditions, perhaps any of us could achieve almost anything.

But talk of causality is premature. Before we can think about causality, we need to figure out whether Gladwell's evidence is even compelling for the claim of a correlation between 10,000 hours of practice and great success. So let's start there.

Gladwell asks, "[i]s the ten-thousand-hour rule a general rule of success?...[i]f we scratch below the surface of every great achiever, do we always find...some sort of special opportunity for practice?" (p. 47) His book goes on to show just that. In case after case, from Bill Gates to the Beatles, great achievers put in their 10,000 hours. Overwhelming evidence, Gladwell concludes, that opportunities for practice predict success.

Let's try to think a little more clearly about Gladwell's evidence. What has Gladwell shown us? Of course, he hasn't actually looked at every great achiever. But he's shown us

evidence that lots of great achievers practice at least 10,000 hours. The big problem is that he's told us nothing about all the people who aren't great achievers. A table of evidence for *Outliers* would look something like Table 4.2.

	Great Achiever	Not Great Achiever	Total
10,000 hours of practice	Most	?	
Not 10,000 hour of practice	Very few	?	
Total			

Table 4.2: Great achievers practice more than 10,000.

Even granting that Gladwell is correct that most great achievers put in 10,000 hours of practice, this tells us nothing about whether the 10,000 hour rule is true. Correlation requires variation. But Gladwell's data lack variation on one of his two variables, achievement. If you want to know whether putting in 10,000 hours of practice correlates with success, you need variation in both practice and achievement. It is not enough to observe that most great achievers put in 10,000 hours of practice. We need to know about the non-achievers' practice habits as well.

Hopefully the table is convincing to you. But to build your intuition further, let's see the problem with claims like Gladwell's in another setting. We are going to repeat his exact argument, but in a fictional example that hopefully will make the problem even clearer.

Suppose a town of ten thousand people experiences a surprising spate of illness.

In the course of a month, five hundred people are taken ill with the same symptoms. Local health officials want to determine the cause of the illness. They take case histories of the five hundred sick people, looking for commonalities.

In the course of this investigation, they find that all 500 people consumed the same beverage, from the same source, the day before they were hospitalized.

The facts above about the beverage and the illness correspond exactly to the facts about practice and success from *Outliers*. Everyone who gets sick (succeeds) drank the same beverage (put in 10,000 hours). Surely, then, drinking that beverage (practicing 10,000 hours) is an important predictor of illness (great success). If we want to know who else is likely to get sick, we should survey town and find out who else drank the same beverage. Right?

Table 4.3 shows data corresponding to our fictionalized story.

	Sick	Not Sick	Total
Drank Beverage	500		
Didn't Drink Beverage	0		
Total	500		

Table 4.3: What sick people drank (made up data).

Suppose we tell you that the beverage in question is tap water. The claim that the “pattern” of illness suggests a correlation between the beverage and the disease is now obviously wrong. Why? Because essentially everyone consumes tap water every day. Sure, all 500 people who got sick consumed tap water, but so too did the 9,500 who didn't get sick. As the following table makes clear, there is in fact no correlation between the beverage and getting sick: 100% of sick people and 100% of healthy people drank the beverage.

The 10,000 hour rule is similarly unsubstantiated by data of the sort presented by

	Sick	Not Sick	Total
Drank Beverage	500	9500	10,000
Didn't Drink Beverage	0	0	0
Total	500	9500	10,000

Gladwell. Yes, lots of successful people practice very hard. So too do lots of unsuccessful people. Think of all the bands that practiced countless hours, played countless gigs, and did not become the Beatles.

Corrupting the youth

Any American kid who liked rock music in 1980s (ask you parents) remembers the Parents Music Resource Center (PMRC). The PMRC was a group of important Washington women who opposed what they perceived to be the increasingly inappropriate content of rock music. Most famous among the founders of the PMRC was Tipper Gore, wife of then Senator and later Vice President Al Gore, who started the group after being shocked by the lyrics of a Prince song. There were two great villains of Ethan's youth: the Soviet Olympic team and the PMRC.

The PMRC claimed that explicit lyrics were corrupting the youth, causing suicide, sexual violence, and even murder. They denounced "porn rock"—a category that included Springsteen because the song "I'm on Fire" offended their sense of sexual propriety—and demanded warning labels be placed on albums. In 1985 the Senate Commerce, Science, and Transportation Committee held hearings. Musicians from across the musical spectrum, from the country singer John Denver to Twisted Sister's Dee Snider (again, ask your parents), testified against the PMRC's position. But Tipper Gore and her allies won the day.

Let's listen to just a bit of the argument. Here is the testimony of Jeff Ling, a PMRC consultant:

Many albums today include songs that encourage suicide, violent revenge, sexual violence, and violence just for violence's sake... This is Steve Boucher. Steve died while listening to AC/DC's "Shoot to Thrill." Steve fired his father's gun into his mouth... A few days ago I was speaking in San Antonio. The day before I arrived, they buried a young high school student. This young man had taken his tape deck to the football field. He hung himself while listening to AC/DC's "Shoot to Thrill." Suicide has become epidemic in our country among teenagers. Some 6,000 will take their lives this year. Many of these young people find encouragement from some rock stars who present death as a positive, almost attractive alternative... Of course, AC/DC is no stranger to violent material... One of their fans I know you are aware of is the accused Night Stalker.

Ling's argument, which is typical of crusaders against corruption of the youth, amounts to this. Some young people behave badly. The youth who behave badly all listen to this terrible rock music. The music must be the cause of the bad behavior. (Of course, talk of causality is again premature. We'll focus on whether such evidence even suggests a correlation.)

Thirty years earlier, in 1954, the Senate heard astoundingly similar testimony about that generation's scourge of the youth, comic books. Here is the neurologist and psychiatrist Frederic Wertham testifying before a Senate subcommittee:

There is a school in a town in New York State where there has been a great deal of stealing. Some time ago some boys attacked another boy and they twisted

his arm so viciously that it broke in two places, and, just like in a comic book, the bone came through the skin.

In the same school about 10 days later 7 boys pounced on another boy and pushed his head against the concrete so that the boy was unconscious and had to be taken to the hospital. He had a concussion of the brain.

In this same high school in 1 year 26 girls became pregnant. The score this year, I think, is eight. Maybe it is nine by now.

Now, Mr. Chairman, this is what I call ethical and moral confusion. I don't think that any of these boys or girls individually vary very much. It cannot be explained individually, alone.

Here is a general moral confusion and I think that these girls were seduced mentally long before they were seduced physically, and, of course, all those people there are very, very great—not all of them, but most of them, are very great comic book readers, have been and are.

We have all heard, and perhaps even made, similar arguments about the insidious effects of television or video games or social media. For instance, following the horrific shootings at Columbine High School, the Department of Education and the Secret Service set up a joint task force to determine what factors would allow school officials to anticipate and prevent school violence. The task force studied all 37 incidents of school violence from 1974 through 2000. While concluding that there is no single profile of a school shooter, they also reported the following (among many other things):

1. "Many attackers felt bullied, persecuted, or injured by others prior to the attack."
2. "Most attackers were known to have had difficulty coping with significant losses or personal failures."

3. “Most attackers engaged in some behavior, prior to the incident, that caused others concern or indicated a need for help.”
4. “Over half of the attacks demonstrated some interest in violence, through movies, video games, books, and other media”

Such analyses are seemingly inevitable after acts of youth violence. But, for reasons we’ve already seen, these arguments, like the Senate testimonies above, are misleading. Even if it were true that every young person who behaves badly also listens to rock, reads comic books, or plays video games, this would not constitute evidence for a correlation between such behavior and these supposed corrupters of the youth. Correlation requires variation. Evidence for the proposition that kids who engage in those activities are *more* likely to be violent than kids who do not engage in those activities must involve a comparison of these two types of kids. Just reporting that violent kids feel bullied or have difficulty coping with personal failures leaves open the possibility that the same is true of kids who are not violent. After all, who among us never felt bullied or never had difficulty coping with personal failures?

Indeed, in these cases, a stark version of the problem is pretty close to right. In the 1980s, a huge proportion of American teenagers listened to music the PMRC found objectionable. And, today, virtually every teenager is on social media. Given this, it will certainly be the case that every teenager who committed an act of violence in the 1980s listened to rock music and every one today is on social media. So someone who doesn’t like, say, rock music could argue, “I’ve looked at every violent teenager and every single one of them listened to AC/DC or its equivalent!” But a music lover could similarly point out, “I’ve looked at every teenager admitted to Harvard and every single one of them listened to AC/DC or its equivalent!” Both statements are true. Neither is evidence for much of anything.

If we want to know if there is a relationship between some putative scourge of the youth

and violence, we must compare violent kids to non-violent kids and see whether violent kids are more likely to engage in that scourge than non-violent kids. (Again, even then we can't say the relationship is causal.) It is upsetting that the so-called experts who testify before Congress and write important government reports have not internalized this simple principle of clear thinking. The fact that they haven't means that, for all the expert opinion offered on the topic, we actually have no idea if there is a link between violence and listening to rock music, reading comic books, playing video games, or using social media.

High school dropouts

Let's stick, for the moment, with troubled youth. Early 21st century America has a high school dropout problem. At a time when the economic returns to education are at an all time high, almost a third of students in the public schools fail to complete high school.

To its credit, in 2006, the Bill and Melinda Gates Foundation decided to put some resources into addressing this issue. As one step in trying to find a solution, they commissioned a study on the correlates of dropping out of high school. The report's main thrust is that high school dropout is not primarily associated with the things you might have guessed—problems at home, lack of academic preparation, or listening to AC/DC. Rather, the big problem seems to be that kids aren't engaged by the educational environment and find school boring.

As the report states, “nearly half (47 percent) [of dropouts] said a major reason for dropping out was that classes were not interesting.” And “nearly 7 in 10 respondents (69 percent) said they were not motivated or inspired to work hard.”

Unfortunately, because correlation requires variation, the evidence in this Gates Foundation study, just like the evidence presented by the PMRC and the anti-comic book lobby before it, is pretty uninformative.

The fact that half of high school dropouts report finding school uninteresting does not constitute evidence that finding school uninteresting correlates with dropout. Because correlation requires variation, such evidence would have to involve comparing dropouts to non-dropouts and seeing whether dropouts were more likely to find school uninteresting. The Gates Foundation study, by looking only at high school dropouts, is unable to shed light on the question.

This point isn't just pedantic. Think about it for a second. We both went to high school. Neither of us dropped out. However, both of us recall finding some classes uninteresting. Didn't you?

Now, our personal experiences also don't constitute compelling evidence. So let's see if we can do a little better in figuring out whether finding classes boring is really a key predictor of dropout. Researchers at Indiana University did a nationally representative survey of high school students in 2009. Most of these students are not going to drop out, yet they report that "[t]wo out of three respondents (66%) in 2009 are bored at least every day in class." That's even more than the 50% of dropouts who find school boring in the Gates Foundation study.

There are many reasons the Gates Foundation survey and the Indiana University survey can't be compared. They sample different groups of students, ask different questions, and are from different years. So we don't want to leap to conclusions. But at the very least, the Indiana University survey should make you worry that finding school boring is in fact a very common experience for high school students. And, so, the fact that half of high school dropouts report having found school boring might not actually be compelling evidence of a correlation.

The future of American education is serious stuff. It is admirable that the Gates Foundation is trying to improve education. But their research ignores a key principle of clear thinking; they are trying to learn about the correlates of educational failures without any

variation in failure versus success. This approach cannot work.

Suicide attacks

In 2009, noted terrorism expert, Robert Pape, testified to the House of Representatives Armed Services Subcommittee on Terrorism. The topic was General Stanley McChrystal's proposal for a 40,000 troop surge to fight the Taliban insurgency in Afghanistan. Here is what Pape had to say:

The picture is clear, the more Western troops have gone to Afghanistan, the more local residents have viewed themselves as under foreign occupation—and are using suicide and other terrorism to resist it. . . . As my study of suicide terrorism around the world since 1980 shows, what motivates suicide terrorists is not the existence of a terrorist sanctuary, but the presence of foreign forces on land they prize. So, it is little surprise that US troops are producing anti-American suicide attackers.

Pape goes on to recommend a major rethinking of American military strategy in Afghanistan. His argument is based on the claim that suicide attacks are primarily motivated by foreign occupation. His evidence is the data he collected and analyzed in an article and two books on every suicide terrorist campaign in the world since 1980.

The argument sounds intuitive. In Afghanistan, the United States was being attacked by suicide bombers who wanted us to leave. Tamil Tiger suicide bombers attacked a government in Sri Lanka they believed was occupying their homeland. Palestinian suicide bombers attack Israelis, arguing that they are foreign occupiers. It sure seems like occupation is a major correlate of suicide attacks.

Now, the claim that virtually every suicide attack is targeted against a foreign occupier is, we think, debatable. For instance, do we really want to say that the 9/11 bombers

attacked a foreign occupier because Osama bin Laden asserted that the United States is occupying Saudi Arabia? But, for the sake of argument, let's assume that the basic factual claim is correct. Does this mean that there is a correlation between foreign occupation and suicide attacks?

The answer is, of course, no. Correlation requires variation. To understand the correlates of suicide attacks, you can't just study every single instance of a suicide attack and look for commonalities. You must compare conflicts with suicide attacks to those without.

An easy thing to do in this case is to simply look at every single country and ask: Are foreign-occupied countries more likely to experience suicide attacks than countries that are not foreign occupied? It turns out that a recent study did precisely that comparison and found that the answer was no. In particular, if we compare occupied to non-occupied countries, the difference in likelihood of experiencing suicide violence is less than 1 percentage point!

What is going on? All those examples of suicide bombers that we listed involved attacking foreign occupiers. How could it be that there is almost no correlation between foreign occupation and suicide attacks?

The way to get some intuition is to think about how many foreign occupations there are that don't lead to suicide terrorism. The British occupation of Ireland, despite sparking a decades-long campaign of violent resistance, never gave rise to suicide terrorism. Basque separatists in Spain fought a decades-long campaign and never resorted to suicide attacks. At various points during the Cold War (and beyond), the United States stationed troops in Germany, Japan, South Korea, Grenada, Panama, and Haiti (all as much occupations as the putative occupation of Saudi Arabia) but suffered not even one suicide attack in any of these locations. If occupation predicts suicide violence, what was going on in all these places?

This example has another nice feature. It not only illustrates the mistake of looking for correlation without variation. It shows you how misled you can be by trying to reach conclusions by only looking at cases where the phenomenon of interest (here, suicide attack) occurs. To see this, it helps to go back in history a little.

Suppose you'd started collecting data on suicide violence in the early 1980s. By 1986 you'd have recorded 33 attacks and over 1,000 deaths. Essentially every single one of those attacks was carried about by the armed Shi'a militia Hezbollah against American, Israeli, and French targets in Lebanon, including the attack on the U.S. Marines Barracks in Beirut which killed 320 people.

If you'd looked for commonalities amongst every suicide attack ever committed in 1986 you might have noticed that they were all carried out by Muslims in the Middle East. Using the same logic that led to the conclusion that occupation is a major predictor of suicide attacks, you might have concluded that Islam was the key correlate.

Of course, if you had done a proper comparison, you wouldn't have reached this conclusion. There are a whole lot of Muslim countries in the world. In 1986, almost none of them had experienced suicide violence.

Moreover, this conclusion in 1986 would have led you terribly astray. In 1987, the world saw the first suicide attack by the Liberation Tigers of Tamil Eelam (Tamil Tigers), a group of secular separatists in Sri Lanka with no ties to Islam. The attack marked the beginning of what would become the largest campaign of suicide violence the world had ever seen. When you try to establish correlation without variation, you can get things colossally wrong.

The World is Conspiring to Make Us Select on the Dependent Variable

As we've seen, it is incredibly easy to fall into the trap of selecting on the dependent variable simply by failing to think clearly. But matters are even worse than that. The world sometimes seems to be organized in a way that almost forces us to look for correlation without variation. In this section we look at three ways in which that is true: the organization of certain professions, the practice of post-mortem analyses following disasters, and the way we seek life advice.

Back pain

Anyone who has suffered from significant back pain knows that it is rough. Ethan spent a particularly memorable semester of graduate school lying on the floor at the back of his classes because he couldn't sit. A tip: graduate school and prescription muscle relaxants are not a great combination.

When you have bad back pain, they send you to a doctor, who sends you to get an MRI. Usually, the MRI shows some discs in the afflicted back bulging or herniated. These bulging discs are taken to be the cause, in some not fully understood way, of the back pain. (Maybe hitting a nerve, maybe something else.)

In Ethan's experience, the recommendation following this diagnosis depends a lot on the kind of doctor you've gone to. Some want to operate. Some send you to a pain clinic where other doctors recommend sticking you with giant needles with medication that will dull the pain and maybe reduce the swelling. Some say to try physical therapy and take lots of painkillers. Ethan recommends taking up martial arts and losing a bunch of weight.

Here's the kicker. As best we can tell, there is precious little evidence that having a

bulging disc is correlated with back pain. Here are the facts. People with back pain are quite likely to exhibit disc herniation. Indeed, in a 2011 British study published in the journal *Pain*, about two-thirds of back pain sufferers who were referred for an MRI had nerve compression as a result of a disc bulge or herniation. This seems like evidence that those bulging discs really are a problem.

But remember, correlation requires variation. You should be asking yourself: what about people without back pain? How do their discs look? Good question. The answer is, they look exactly the same as the people who do have back pain! A 1994 study published in the *New England Journal of Medicine* found that about two-thirds of people who do not suffer from back pain also have a disc bulge or herniation. Once you compare, the apparent association between bulging discs and back pain disappears.

It is easy to see how doctors could end up associating bulging discs with back pain. Even if they are thinking clearly, by dint of profession, a doctor is almost doomed not to look at variation. Sick people go to the doctor. Healthy people don't. Your typical back doctor just doesn't get much of an opportunity to look at the MRIs of people with well-functioning backs.

Post-mortems

Another way the world conspires to make us look for correlation without variation is through institutional rules or procedures. A particularly common example is the way organizations respond to both great failures and great successes.

Following a crisis or disaster, organizations want to know what went wrong so they can avoid making similar mistakes in the future. Likewise following great successes they want to know what went right to establish best practices. Looking closely at an instance of great failure or great success is not, in and of itself, a mistake. Indeed, it is a very sensible starting

point. But, if you think clearly, you should already be able to see that, on their own, such post-mortem procedures are not sufficient to establish correlations between what went wrong (or right) and existing practices.

The question you are trying to answer when assessing lessons learned from a crisis is, “which decisions should have been made differently to avoid the crisis, given what we knew at the time?” However, when assessing lessons learned, we often slip into answering a slightly different question: “which decisions should have been made differently to avoid the crisis, given what we know now?”

The latter isn’t a terribly useful question to answer, for the reasons we’ve already talked about in this chapter. Suppose you find some decision that, it turns out, seems to have led directly to the disaster. After the fact, it is easy to say, “had we not taken that action, the disaster wouldn’t have happened.” But does that mean that you shouldn’t take similar such actions in the future? To know the answer to that, you’d want to know whether disasters are more likely to occur in the presence of such actions than in their absence. That is, you want to know whether there is a correlation between taking such actions and disasters occurring. To establish a correlation, you need variation. But a post-mortem, almost by definition, has no variation. You are only looking at an instance of the disaster occurring.

To see what we mean a little more intuitively, let’s start with an fictional example. Then we’ll turn to some real cases.

Imagine you are a high school band director preparing for a regional competition in a week. You have to decide whether to push the kids hard with a grueling schedule of rehearsals or give them time off so they go into the competition relaxed. You weigh the pros and cons, deciding preparation is more important than mental state. So you schedule a week of extra rehearsals. Unfortunately, the band doesn’t play terribly well on the day of the competition, and you are eliminated in the first round.

In your post-mortem analysis you ask the question, “what should I have done to avoid the loss?” It occurs to you that you’ve seen a lot of bands lose competitions in this same way (i.e., having rehearsed themselves to death the week before), so you decide to collect some data. You look at the history of all the competitions in which your band was eliminated in the early rounds. Just like in this year’s competition, you find that in almost every one of these competitions, you scheduled a heavy rehearsal schedule in the week leading up to the competition.

Let’s say you did a week of intensive rehearsing prior to 80 out of 88 losses. The post-mortem conclusion seems clear. In over 90% of the cases where your band was eliminated early, it was after a week of exhausting rehearsal. Now you feel even more sure: intensive rehearsal is the wrong strategy.

But this conclusions doesn’t necessarily follow from the data you’ve collected. In fact, there’s no way to know whether those rehearsals are associated with performing well or poorly from this data alone. Table 4.4 summarizes what you know so far from your post-mortem analysis.

	Do well	Do poorly	Total
Extra rehearsals	?	80	?
Take it easy	?	8	?
Total	?	88	?

Table 4.4: Rehearsal strategies in the week before competitions where your band performed poorly (made up data).

You have answered the wrong question. You don’t want to know if band did extra rehearsals prior to most of the competitions where they performed poorly. You want to

know if extra rehearsals are positively or negatively correlated with performing well. The answer to that question will help you know whether those extra rehearsals are a good idea for the next competition.

To answer that question, you have to look at the correlation between extra rehearsals and performing well in competition. But you can't know the correlation from your post-mortem analysis. Correlation requires variation. Your post-mortem, by focusing only on poor performances, guarantees that you lack the variation needed to establish a correlation.

To do a better job, you could look at the history of all the band competitions you've participated in, whether you performed well or poorly. Now you have variation in both variables and can fill in all the data, as shown in Table 4.5.

	Do well	Do poorly	Total
Extra rehearsals	300	80	380
Take it easy	12	8	20
Total	312	88	400

Table 4.5: Rehearsal strategy in the week before competitions where your band performed well or poorly (made up data).

From this table it is clear that there is in fact a strong positive correlation between scheduling extra rehearsals and performing well. The probability of your band performing well when you rehearsed hard is about 79% ($\frac{300}{380} \approx .79$). By contrast, the probability of your band performing well when you took it easy the week prior to a competition is only 60% ($\frac{12}{20} = .60$). The only reason that the post-mortem turned up the finding that almost every poor performance involved intensive rehearsals is that those extra rehearsals are so effective that sensible band directors almost always schedule them.

By finding the variation needed to establish the correlation that is actually relevant to the question at hand, you reach a very different conclusion than you did in your original post-mortem. Following the loss, it seemed like intensive rehearsals were a bad idea. But before the fact, given the information available, rehearsing hard was exactly the right call. Faced with the same situation again, you should probably make the same decision.

This problem is endemic to the process of post-mortems following disasters. We tend to look at the factors that seem like they contributed to the disaster, ask if they were also present in past disasters, and when they were, conclude that we should eliminate those factors in the future. But, in so doing, we are making the same mistake as the band director. Without variation in whether or not a disaster occurred, we can't actually learn whether the presence of those factors is correlated with the occurrence of a disaster. So we don't know if there are lessons to be learned.

We are going to show you what we mean with two examples of post-mortems that followed major disasters—the Challenger shuttle explosion and the financial crisis of 2008. In each case, we will see that, while after the fact it sure looks like some serious and obvious mistakes were made, it is less clear that the decision makers could have known that they were making mistakes before the fact. Moreover, once we've grasped this, we will be able to think more clearly about how to design post-mortems that might be more informative about lessons learned.

The Challenger disaster

On January 28, 1986, the Space Shuttle Challenger disintegrated off the coast of Cape Canaveral less than two minutes after launch. Seven crew members were lost. The night before the Challenger exploded, a small group of engineers from the NASA contractor responsible for the shuttle's solid rocket boosters predicted that the cold weather would lead to a catastrophic failure that might well compromise the shuttle. The concern was that the

critical O-ring seals responsible for containing gases produced by burning the rocket fuel were not certified to operate at the low temperatures that preceded this particular launch. If the O-ring seals failed, the engineers argued, the hot pressurized gas could burn through the rocket's casing, causing catastrophic failure.

These predictions were shunted aside by managers at NASA and the engineers' own firm, but proved tragically correct. Many post-mortem analyses focused on NASA's failure to take these concerns seriously. The conclusion most observers reached was that the disaster was caused by organizational and cultural failures at NASA that facilitated group-think and led managers to systematically ignore important objections from experts. For instance, the Report of the Presidential Commission on the Space Shuttle Challenger Accident (the Rogers Commission) concluded, "failures in communication...resulted in a decision to launch 51-L based on incomplete and sometimes misleading information, a conflict between engineering data and management judgements, and a NASA management structure that permitted internal flight safety problems to bypass key Shuttle managers."

The Challenger case is interesting. No one questions the physics behind the conclusion that the O-ring failed because of cold temperatures. Indeed, the Rogers commission included the Nobel prize winning physicist Richard Feynman precisely so they could say with authority whether the engineers were right on the science. They were. And so, in this sense, launching the shuttle was clearly a mistake.

Because the science is so clear, it seems natural for a post-mortem to ask what it was about the process that led decision makers to ignore engineers making good scientific arguments. But it is here that our knowledge of the pitfalls of post-mortems should make us stop and think. We know that, after the fact, the decision to launch was tragically flawed. But we want to evaluate whether it was a bad decision at the time it was taken. To do so, we need to know about the correlation between the presence of scientifically valid engineering concerns and the success of shuttle launches. And to know about that correlation, we need

variation, we must compare disastrous launches to successful launches.

We aren't engineers, so we aren't going to try to weigh in on whether or not the decision to launch Challenger was reasonable at the time it was made. But we can see how, to get at this, a post-mortem commission would need to ask questions they aren't accustomed to asking. Post-mortem commissions ask what led to the disaster, whether people had raised the relevant objections, and if so why those objections weren't listened to. In addition, we would like them to ask whether engineers also raised scientifically valid concerns prior to lots of successful launches. This doesn't seem implausible. Space shuttle launches are incredibly complex and dangerous undertakings. Perhaps there is almost always a scientifically valid reason for serious concern. If so, then there actually wouldn't be much (if any) correlation between the presence of such concerns and launch success. If this is the case, unless you are prepared to simply shut down the space program, it isn't fair to say that launching following a scientifically plausible objection by an engineer is always a mistake. This is the sort of thing one would want to know from a post-mortem commission before reaching conclusions about changing the organizational culture, management practices, or what have you.

Financial crisis of 2008

The financial crisis that shook the world economy in 2007 and 2008 began with a crash in the US subprime housing market. This crash had ripple effects across the banking sector that eventually spread throughout the world economy. Understandably, in the wake of this crisis—at the time, the worst since the Great Depression—policymakers and the public alike were interested in early warning indicators that might help them forecast and forestall future crises.

Perhaps the most important post-mortem analysis attempting to provide such early warning indicators was the book *This Time is Different* by the economists Carmen M. Reinhart and Kenneth S. Rogoff. Reinhart and Rogoff collected and analyzed data on every

major financial crisis of the last 800 years. By doing so, they argued, they can identify a few key indicators that almost always precede such a crisis. These include uncommonly large current account deficits (that is, goods and services exported minus imported net of income from abroad), asset price bubbles, and excessive borrowing. For instance, in 2006 the United States had a current account deficit close to 7% of GDP, a bubble in the housing market, and ballooning federal debt. Thus, Reinhart and Rogoff conclude, “we’ve been here before.” The implication is that the U.S. financial crisis could have been predicted by the presence of those same factors which seem to characterize financial crises across time and around the globe. Similar patterns were true before the financial crises in Latin America in the early 2000s, East Asia in the 1990s, Nordic countries in the 1980s, and so on into history.

The problem with this argument is the same as in our earlier examples. Early warning indicators should be correlates of financial crises. And, because correlation requires variation, to know if current account deficits, soaring asset prices, or heavy borrowing correlate with financial crises, we need variation in crises. That is, we need to know not only that these factors tend to be present when crises occur. We need to know how frequently they are present when crises do not occur. Without such variation, we cannot establish a correlation.

Reinhart and Rogoff’s plan of studying every major financial crisis for 800 years cannot answer the question. And there are reasons to be worried about their conclusions. As the MIT political scientist David Andrew Singer points out, one need only look at recent history to cast some doubt on the story. For instance, in the late 1990s the United States had all the early warning signs for a financial crisis. There were large current account deficits as a results of massive foreign investment in dot.com’s. Moreover, when the dot.com bubble burst, “it wiped out approximately \$5 trillion in market capitalization.” Yet no financial crisis occurred. This, of course, is just one anecdote. But it should make you wonder whether the factors Reinhart and Rogoff point to are really good predictors of financial

crises, or are they common features of the world that happen to be around when financial crises occur and also when they don't occur.

Life advice

We've been arguing that our world is organized in ways that push us to try to figure out the correlates of success or failure without looking at variation, even though it won't work. It is important to see that this problem isn't confined to big institutional settings. We are all victims of it every day in many small ways.

One simple example is the ways in which we seek life advice, which almost always involves asking successful people how it is that they succeeded. In our business, for instance, graduate students are encouraged to ask senior professors what they did to succeed on the job market. We imagine the same is true in other professions. There is certainly no shortage of self-help books describing the habits of successful people.

But such wisdom suffers from exactly the problems we've been pointing to. Successful people, reflecting on their lives, are inclined to identify a few decisions they made or a few personal characteristics that seem important and offer them as advice to the next generation. But those successful people typically have no idea whether many other, less successful people, made similar decisions or had similar characteristics. That is, their introspection about the correlates of success lacks variation. As such, successful people don't really know whether the lessons they point to in telling their personal stories are correlates of success or not. And so, we leave you with this happy bit of wisdom of our own. Beware life advice. It is probably mostly nonsense.

Key Terms

- **Selecting on the Dependent Variable:** Examining only instances where the phenomenon of interest occurred, rather than comparing cases where it occurred to cases where it did not occur.

Exercises

1. In Chapter 2 we discussed the differences between statements about correlations and other factual statements that do not convey information about a correlation. Now that you have a deeper understanding that correlation requires variation, consider the following statements. Which ones describe a correlation and which ones do not?
 - (a) Most top performing schools have small student bodies.
 - (b) Married people are typically happier than unmarried people.
 - (c) Among professionals, taller basketball players tend to have lower free-throw percentages than shorter players.
 - (d) The locations in the U.S. with the highest cancer rates are typically small towns.
 - (e) Older houses are more likely to have lead paint than newer ones.
 - (f) Most colds caught in Cook County are caught on cold days. (This one also doubles as a tongue twister.)
2. At least 20 billionaires dropped out of college before earning their fortunes including Bill Gates and Mark Zuckerberg.
 - (a) Does this mean that dropping out of college is correlated with becoming a billionaire? Why or why not?

- (b) Draw the two-by-two table that would allow you to assess whether dropping out of college is correlated with becoming a billionaire. Let's assume that exactly 20 people have dropped out of college and become billionaires, so you know what to put in one of the four cells. Make your best guess for the other cells. At the time of this writing, there are about 7.8 billion people in the world, and about 2,000 billionaires. Do you think there is a positive or negative correlation between dropping out of college and becoming a billionaire?
- (c) What proportion of the non-billionaires would need to be college dropouts in order for the correlation to be negative? What proportion of the non-billionaires would need to be college dropouts in order for the correlation to be positive?
- (d) If you're currently a college student deciding whether you want to drop out in the hopes of becoming a billionaire, you may want to restrict attention to people who actually started college. Do you think the correlation between dropping out of college and becoming a billionaire is more or less likely to be positive if we restrict attention to just people who start college?
- (e) About 7 percent of the world's population has a college degree. And about a third of those who start college complete it. If we assume that everyone who becomes a billionaire started college, you should now have all the information you need to assess this correlation between becoming a billionaire and dropping out of college among those who start college. Is it positive, negative, or zero?

Notes

For more information on Frederic Wertham and his flawed argument about comic books, we recommend

David Hajdu. 2009. *The Ten-Cent Plague: The Great Comic-Book Scare and How it*

Changed America. Picador.

And although Wertham's conclusions don't follow from his empirical findings because correlation requires variation, it also turns out that Wertham may have manipulated and fabricated his data. See

Carol L. Tilley. 2012. Seducing the Innocent: Fredric Wertham and the Falsifications That Helped Condemn Comics. *Information & Culture: A Journal of History* 47(4):383–413.

The survey on boredom in school is called the High School Survey of Student Engagement, and it's administered by the Center for Evaluation and Education Policy at Indiana University. The quote was taken from their 2010 study. For more information, see <http://newsinfo.iu.edu/news-archive/14593.html>.

For more on suicide attacks, including why we can't learn about the causes or correlates of suicide violence from studying only cases where it occurs, see

Robert A. Pape 2003. "The Strategic Logic of Suicide Terrorism." *American Political Science Review* 97(3):343–361.

Scott Ashworth, Joshua D. Clinton, Adam Meirowitz, and Kristopher W. Ramsay. 2008. "Design, Inference, and the Strategic Logic of Suicide Terrorism." *American Political Science Review* 102(2):269–273.

Robert A. Pape 2008. "Methods and Findings in the Study of Suicide Terrorism." *American Political Science Review* 102(2):275–277.

Scott Ashworth, Joshua D. Clinton, Adam Meirowitz, and Kristopher W. Ramsay. 2008. "Design, Inference, and the Strategic Logic of Suicide Terrorism: A Rejoinder." Unpublished note: <http://home.uchicago.edu/~sashwort/rejoinder3.tif>.

To read about the high rates of disc bulges and herniation among people with and without back pain (and therefore the lack of correlation between these characteristics and back pain) see

Michael J. DePalma, Jessica M. Ketchum, and Thomas Saullo. 2011. “What Is the Source of Chronic Low Back Pain and Does Age Play a Role?” *Pain Medicine* 12(2):224–233.

Maureen C. Jensen, Michael N. Brant-Zawadzki, Nancy Obuchowski, Michael T. Modic, Dennis Malkasian, and Jeffrey S. Ross. 1994. “Magnetic Resonance Imaging of the Lumbar Spine in People without Back Pain.” *New England Journal of Medicine* 331:69–73.

For more on why we can’t learn about the causes and correlates of financial crises by studying only financial crises, see

David A. Singer. 2010. “Is This Time Different?” *The Political Economist*, Fall pp. 4–5.

Chapter 5

Regression for Describing and Forecasting

What you'll learn

- Regression involves finding the *line of best fit* through some data. It is perhaps the most important tool for describing the relationship between two or more variables.
- Under certain conditions, regression can be useful for forecasting or making predictions.
- Things can go wrong with regression, especially if we have a small amount of data. Among the most important problems that can arise is *overfitting*.
- A little intellectual history about where regression came from.

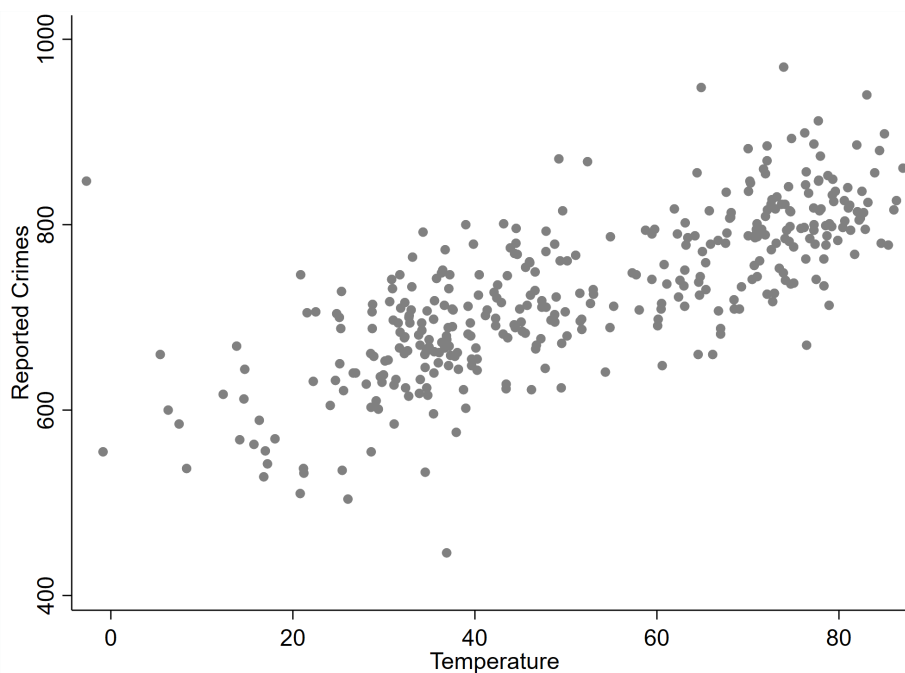


Figure 5.1: Crime and temperature in Chicago across days in 2018.

In Chapter 2, we defined correlation and discussed its three uses: description, forecasting, and causal inference. We also talked about a variety of ways to quantify correlations, including the slope of the regression line, the covariance, and the correlation coefficient. Regression lines are the most common and useful of these. And so, in this chapter, we are going to take a deeper dive into regression to make sure we are all thinking clearly about this important technique.

Regression Basics: Crime and Temperature Revisited

Let's return to the data on crime and temperature in Chicago that we discussed back in Chapter 2. Figure 5.1 reminds you what a scatter plot of that data looks like.

As you can see just by looking at the data, generally speaking, warmer days have more

crime. But you sometimes want to be more precise about the relationship. If you worked for the Chicago Police Department and your boss asked you to summarize the relationship between temperature and crime, they probably wouldn't be particularly pleased if you just showed up with this graph. They might want a simple summary of the relationship that's easy to understand and communicate to people making policy decisions. This is where linear regression comes in.

The line that best fits the data provides just the kind of accessible summary of the relationship between temperature and crime that we are looking for. In particular, as we discussed in Chapter 2, the slope of that line tells us about the sign and magnitude of the correlation. So let's figure out how exactly we find that line, so that we can think clearly about how to interpret and communicate what it has to tell us.

With the exception of completely vertical lines (which wouldn't provide a useful description or prediction anyway), all hypothetical lines that we could draw on this graph can be described by an equation of the following form:

$$\text{Predicted Crime} = \alpha + \beta \cdot \text{Temperature}.$$

The parameter α is called the *intercept*; it is the predicted number of a crimes on a day where the average temperature is 0 degrees Fahrenheit. The parameter β is the *slope*; it is the amount that predicted crime goes up with each degree Fahrenheit. Any possible line on the graph corresponds to one particular combination of α and β , and if we know α and β , we can perfectly reproduce that line and generate predicted crime for any average temperature.

The method of summarizing or predicting crime using an equation like this one is called *linear regression*. It's *linear* because we can represent this equation graphically with a line, and it's called *regression* for historical reasons that we are getting to.

Of course, we don't want to try to predict crime from temperature using any arbitrary line. The wrong line will yield really bad predictions. We want to predict crime from temperature using the line that best fits the data.

In order to find the values of α and β that give us the line that best fits, we need to start by defining what the term *best fits* means. We do so quantitatively by choosing a statistic to measure how well any given line does at summarizing or fitting the data. Then we (or, at least our computer) finds the values of α and β that result in the best possible value of that statistic. Those values of α and β describe the line that best fits the data according to the statistic we chose.

Clearly, the statistic we choose to evaluate fit is quite important. And we have some options. The goal is to find a statistic that has some sensible properties. Most importantly, we want it to like values of α and β that result in a line that is close to the data on average. As we mentioned briefly in Chapter 2, the most commonly used such statistic (and the one on which we focus) is the *sum of squared error*. Figure 5.2 illustrates how it is constructed.

First, draw the line associated with some α and β on the same graph as the data. In the figure, we drew the line associated with $\alpha = 650$ and $\beta = 2$ through our crime and temperature data. (As we will see later, this turns out not to be the line of best fit.)

Next, for every data point, measure the vertical distance from the data point to the line. That distance is called the *error* for that data point. In the figure, these errors are drawn in for a few data points. They are the vertical lines from a data point to the line. We didn't draw all the errors because the figure would have been too messy. Notice, the error for any given data point can be positive (if the data point lies above the line) or negative (if the data point lies below the line). But we want a measure of how far the data point is from the line. We don't care whether it is above or below. So, to get such a measure, we next square the error for each data point. The squared error is positive regardless of whether the data point lies above or below the line. It is just a measure of how far the data point

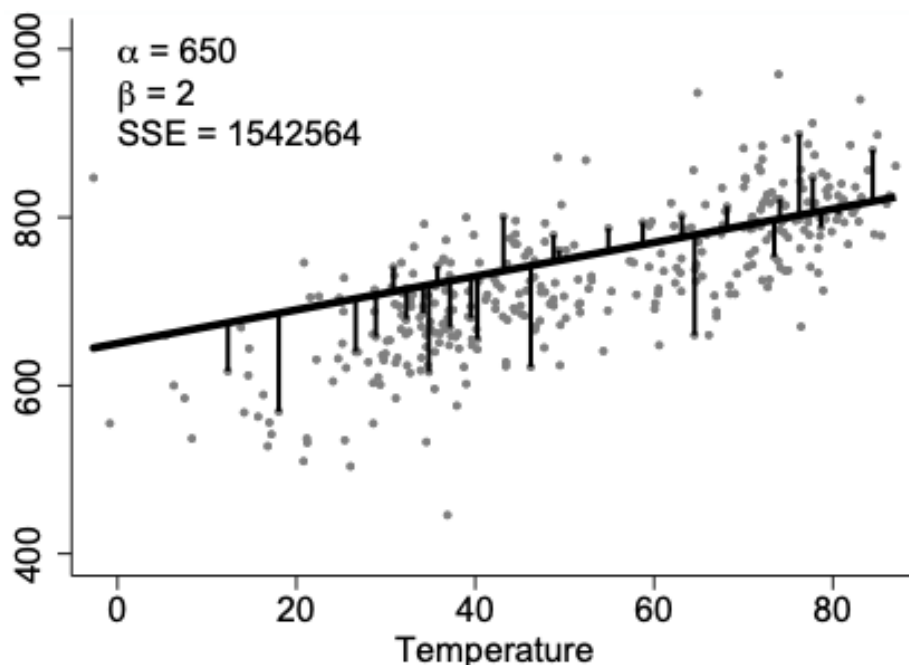


Figure 5.2: Fitting a line through crime and temperature, showing some of the errors.

is from the line.

Next, add up all those squared errors to get the sum of squared errors. The figure reports the sum of squared errors for this particular line. The bigger the sum of squared errors, the further the data is from the line, on average. We can follow this procedure to get the sum of squared errors for any line associated with an α and β .

Finally, the line we are looking for is the one with the smallest sum of squared errors. So find the values of the parameters α and β that minimize the sum of squared errors.

These values of the parameters are called the *ordinary least squares (OLS) regression coefficients*. The line associated with these parameters is the *OLS regression line*. It is our line of best fit.

Figure 5.3 below shows the crime and temperature data with 4 different lines drawn through it, corresponding to different combinations of α and β . For each line, the figure

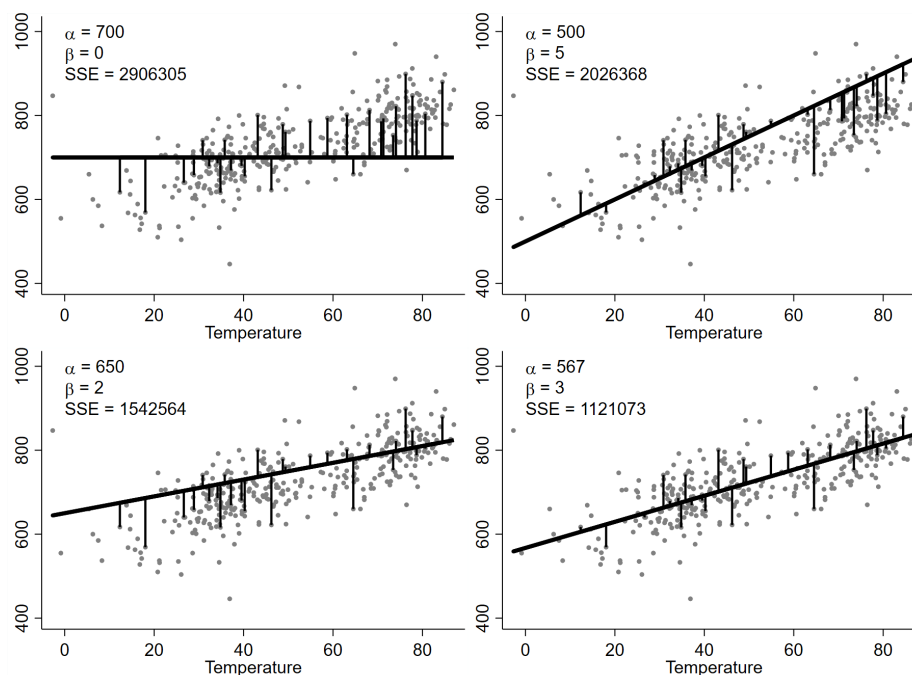


Figure 5.3: Fitting different lines through crime and temperature, showing some of the errors.

reports the α , β , and the sum of squared errors (SSE). A few of the errors are shown visually with vertical black lines. The bottom-right panel shows the OLS regression line—i.e., the line that minimizes the sum of squared errors. Visually, we can see that this line is a better approximation of the data than other other three options.

How do we interpret the OLS regression line? As we see in the figure, rounding to the nearest integer, the intercept (α) is 567 and the slope (β) is 3. In other words, the OLS regression line is telling us that, in 2018, on days when the average temperature at Midway Airport was 0 degrees, there were about 567 crimes on average, and for every additional degree Fahrenheit, the average number of crimes increased by about 3.

We didn't have to choose our regression line by minimizing the sum of squared error. For example, we could have instead minimized the sum of the absolute value of the errors.

Or we could have minimized the sum of errors raised to the fourth power. The possibilities are endless.

We like the sum of squared errors for a couple reasons. First, minimizing the sum of squared errors happens to provide the best linear approximation to another function you might care about: the *conditional mean function*. The conditional mean function is a function that tells you the mean (average) of some variable conditional on the value of some other variables. Here the particular conditional mean function we are interested in is the one that gives the mean number of crimes conditional on temperature. Suppose that for, say, each degree Fahrenheit, you calculated the average number of crimes on days with that temperature and plotted them. That gives you a graph of a conditional mean function—for each degree of temperature, it tells you the average number of crimes. In Figure 5.4, the light gray dots are our raw crime and temperature data and the large black dots are the mean number of crimes conditional on being in a 5 degree Fahrenheit bin (0–5 degrees, 6–10 degrees, and so on). As you can see, the regression line, in addition to being the line of best fit through the raw data, is also a very good approximation of these conditional means. So, if you are interested in averages, the line that minimizes the sum of squared errors is the way to go.

Of course, you might not be interested in averages. Perhaps, instead, you want to estimate the median. In that case, it turns out that you’d want to draw the line that minimizes the sum of the absolute values of the errors. As we said, there are a variety of reasonable choices.

A second reason that people focus on minimizing the sum of squared error is historical. There is an easy way to calculate the values of α and β that minimize the sum of squared errors using a set of mathematical tools from the field of linear algebra. As a result, OLS coefficients can be calculated quite quickly. Back when people did this by hand, or even when computers were much slower, this was important. As computational speeds improve,

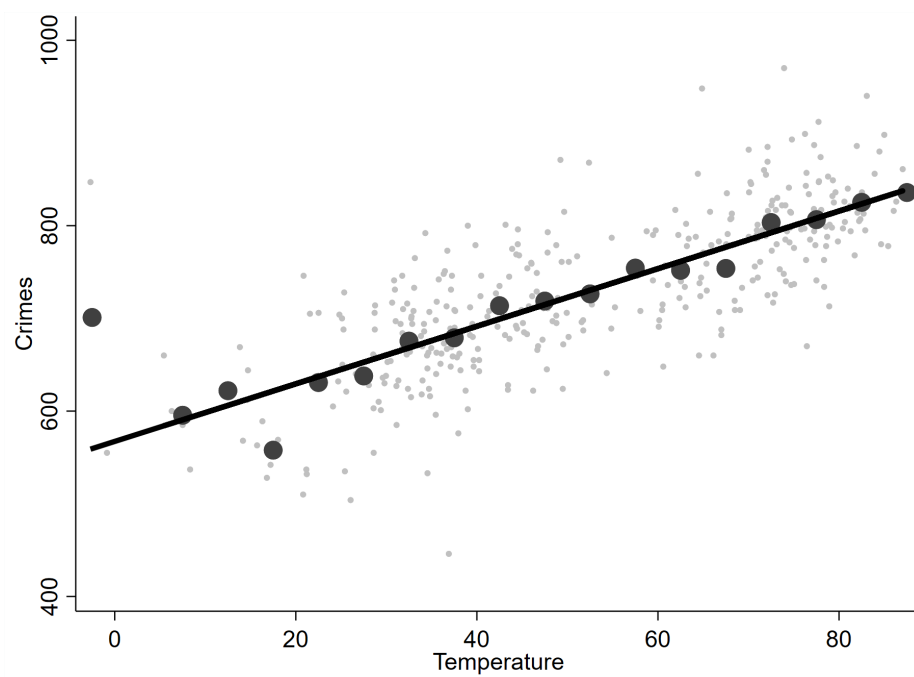


Figure 5.4: The regression line through the data is also the best linear approximation to the conditional means.

however, this consideration is becoming less relevant.

Linear Regression, Non-Linear Data: Age and Voter Turnout

What do we do when we want to use linear regression, but our data is not terribly well described by a line? To start to think about this, let's return to the data we looked at in our discussion of voter turnout in Chapter 2. Remember, there we wanted to describe the relationship between age and voter turnout—perhaps to know whether young people are politically underrepresented or perhaps to decide whom to target with a get-out-the-vote drive.

Figure 5.5 shows the voter turnout rate for each year of age between 18 and 68 in the 2014 midterm elections. Notice, in this data, an observation is not an individual, it is an age cohort. As with temperature and crime, the relationship between age and turnout is potentially quite complex. What if we want to summarize the average relationship between age and turnout in a simple way? Or what if we didn't have the data for 31 year olds (omitted from the figure) and we wanted to come up with our best guess for their turnout level? Or what if we wanted to predict turnout based on age in the 2018 election? Linear regression could be useful for all of these purposes.

Looking at the graph, the relationship between age and turnout appears approximately linear, at least for this range of the data. In other words, we could probably draw a line on this graph that comes pretty close to each of the data points. And if we did draw such a line, this would be fairly useful for both description and prediction.

Let's try out OLS regression with our voter turnout data. We could again describe any line with the following regression:

$$\text{Predicted Turnout} = \alpha + \beta \cdot \text{age}.$$

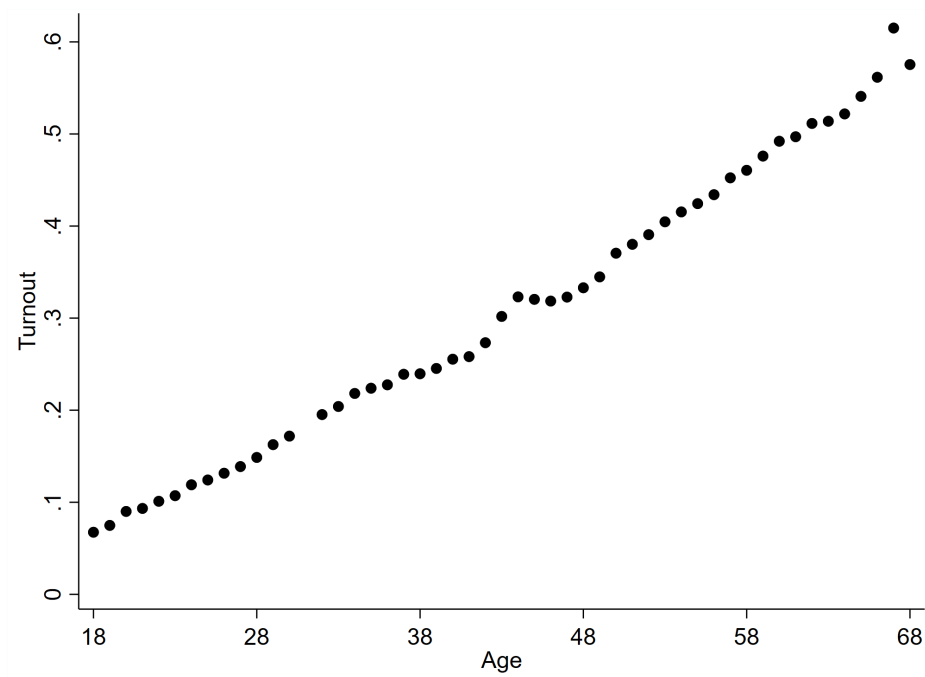


Figure 5.5: Voter turnout rate by age for 2014 midterms.

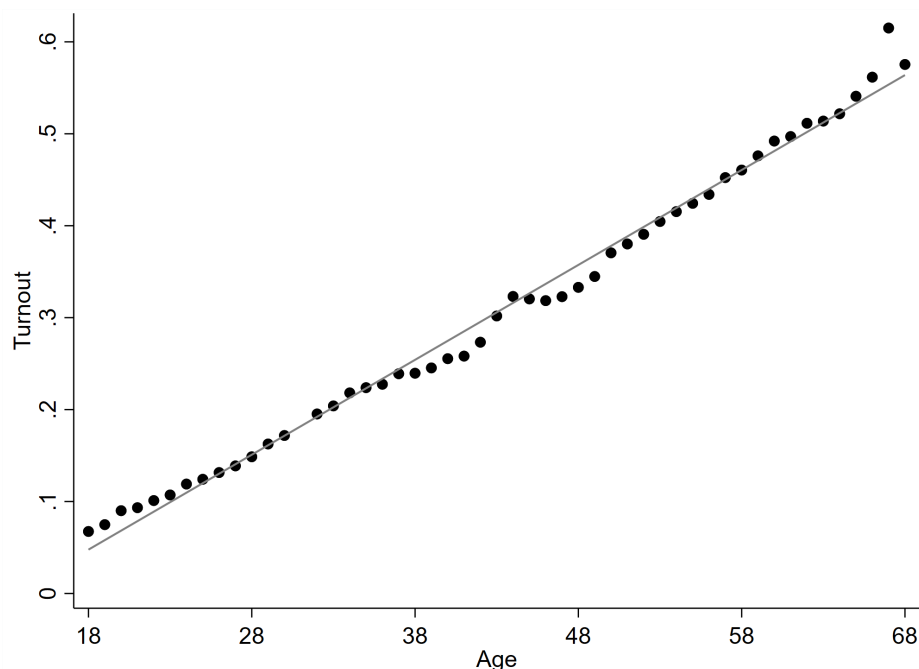


Figure 5.6: A regression line through voter turnout rate by age for 2014 midterms.

There’s a lot of lingo to describe finding the α and β that minimize the sum of squared errors. Sometimes we say that we’re “regressing turnout on age.” When we’re in a long-winded mood, we’ll say that we’re “running a linear ordinary least squares regression where turnout is the *dependent* (or *outcome*) variable and age is the *independent* (or *explanatory*) variable.” In any event, our statistical software program tells us that, for this data, $\alpha = -0.1381$ and $\beta = 0.0103$. With these two numbers, we can draw the line that best fits the data, and we can generate predicted turnout for any given age. Figure 5.6 shows how the OLS regression line looks.

It is important to pause and think clearly about the substantive meaning of the regression line.

The number α is the intercept. It tells us that the predicted turnout rate of people aged zero is -0.1381 , or about -14 percent. That’s a pretty weird prediction. Turnout rates

can't be negative. And infants can't vote. We know that the turnout rate for zero year olds is zero.

Does this mean that the regression is meaningless or wrong? No. It reflects the fact that the regression isn't super useful for forecasting the voting behavior of babies. That isn't surprising. Our regression line was chosen to do a good job of approximating our data. We shouldn't expect it to do a terribly good job approximating the behavior of people with ages well outside the range of our data. And we don't have any data on people younger than 18.

The number β is the slope. It tells us that, on average, within the range of our data, each additional year of age corresponds with an increase in turnout of just over 1 percentage point. In other words, on average, between the ages of 19 and 68, people are about 1 percentage point more likely to vote than people who are just one year younger than themselves. That's interesting. And it accumulates across years, implying that 68 year olds are approximately 50 percentage points more likely to vote than are 18 year olds, which is exactly what we see in the data.

The regression line is doing its job pretty well. It gives us a fairly simple and quick summary of the relationship between age and turnout for people between the ages of 18 and 68 in the 2014 election. In this particular election, 18 year olds voted at an approximate rate of 4.8 percent $((-.1381 + .0103 \cdot 18) \cdot 100 \approx 4.8)$, and then turnout increases by just over 1 percentage point for every additional year in age. Although this summary doesn't get turnout exactly right for each age group, it gets pretty darn close. And, in our view, what is lost in accuracy (compared to, say, just listing turnout rates by age) is more than made up for in parsimony and ease of communication.

We can also use α and β to predict turnout levels for voters whose ages are not in our data. For reasons we've already discussed, we don't want to extrapolate too far. We aren't going to do a good job if we extrapolate back to infants, or even 17-year-olds, since they

aren't eligible to vote. We probably also don't want to extrapolate to people too much older than 68. Our predictions are likely to be pretty good for 69 or 70 year olds, for whom we predict turnout rates of approximately 57.3 percent $((-.1381 + .0103 \cdot 69) \cdot 100)$ and 58.3 percent $((-.1381 + .0103 \cdot 70) \cdot 100)$, respectively. But the further we get away from the range of our actual data, the more we should worry about the reliability of our predictions.

The spot where we might be most confident in our predictions is for 31 year olds. For whatever reason, our graph shows no data on 31 year olds. (Here that's because we purposefully omitted it for illustrative purposes. But if you start working with data you'll find that this sort of thing happens all the time. Maybe the county clerk spilled coffee on the voter returns for 31 year olds.) But we have lots of data on people with ages on both sides of 31. So we can probably generate pretty good predictions for turnout by 31 years olds. Let's see.

Our regression equation predicts a turnout rate for 31 year olds of just over 18 percent $(-.1381 + .0103 \cdot 31) \cdot 100 = 18.12$. Since we actually do have the data, we can see how well our prediction pans out by adding the 31 year olds back in to the graph.

Figure 5.7 plots the same regression line, fit to data on 18–68 years olds excluding 31 year olds. But it reintroduces some previously excluded data points, plotting them in red. The new data include 31 year olds, as well as folks aged 69–88.

With 31 year olds, we hit the mark almost perfectly, we predicted a turnout rate of 18.12 percent and the true rate was 18.11 percent. With 69 through 72 year olds, we did okay, though not as well. But our predictions start performing really poorly for the oldest individuals.

That's because the relationship between age and turnout seems to be quite different for the elderly. For younger people, turnout is increasing in age. But as people get old, turnout starts to drop precipitously with age, perhaps because of deteriorating health. As a result,

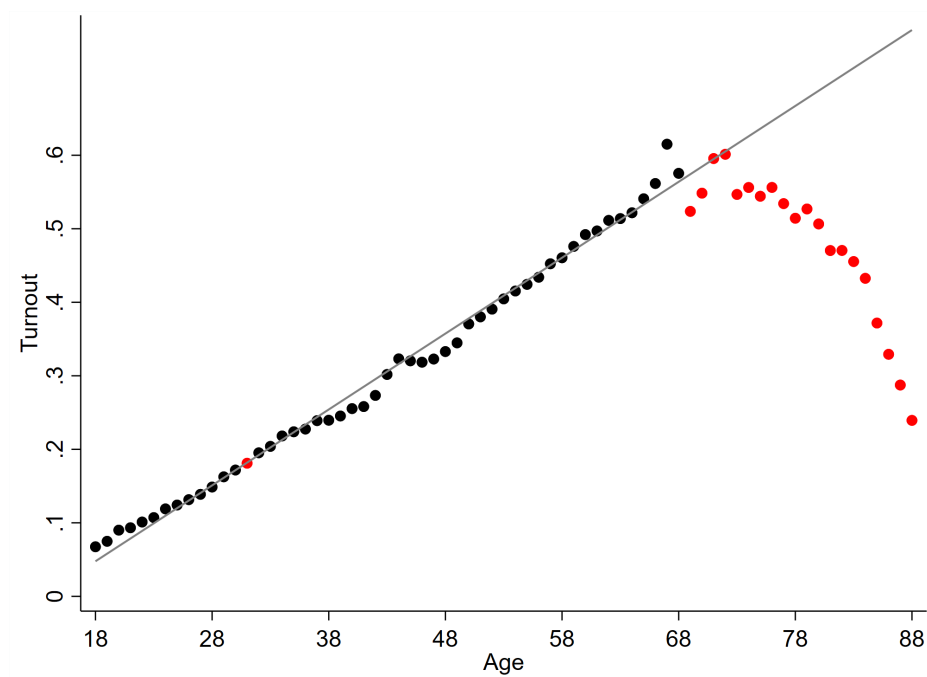


Figure 5.7: Using the regression line to predict voter turnout rate for out of sample ages.

trying to predict the difference in voter turnout between an 80 year old and an 88 year old using data on voter turnout of 18–68 year olds doesn't work very well. Just like with our prediction of a –14 percent turnout rate for infants, this illustrates what can go wrong when we try to extrapolate our predictions outside the range of the data that we used to generate the regression line on which those predictions are based.

Suppose we in fact wanted to analyze the relationship between age and turnout for everyone between the ages of 18 and 88. Clearly, just looking at the data, we can see that the relationship is not linear. How should we account for this non-linearity?

One approach would be to continue to fit a linear regression, but now using all the data. Even if the data themselves don't sit on a line, we can still find the line that minimizes the sum of squared errors. As you can see in Figure 5.8, the resulting line looks ok, but there is now a lot more error, since we are fitting a line to data that have a clearly non-linear relationship.

A second approach is to keep fitting regression lines, but use a different line for different parts of the data. For instance, we could find the line that minimizes the sum of squared errors for the data on people between 16–68, a second line that minimizes the sum of squared errors for the data on people ages 69–78, and a third line for that data on people ages 79–88. This would not be as parsimonious or easily communicated as running a single regression—instead of two parameters (α and β) we would have 6 parameters (a separate α and β for each regression line). But, as you can see in Figure 5.9, the payoff we get for that lack of parsimony is a tighter fit to the data (i.e., less error).

We hinted at a third way to deal with non-linearity back in Chapter 2. There's no reason that our regression equation has to only have one explanatory variable. If we know that there's a non-linear relationship between turnout and age, maybe we want to consider transforming the age variable into age-squared, age-cubed, and so on.

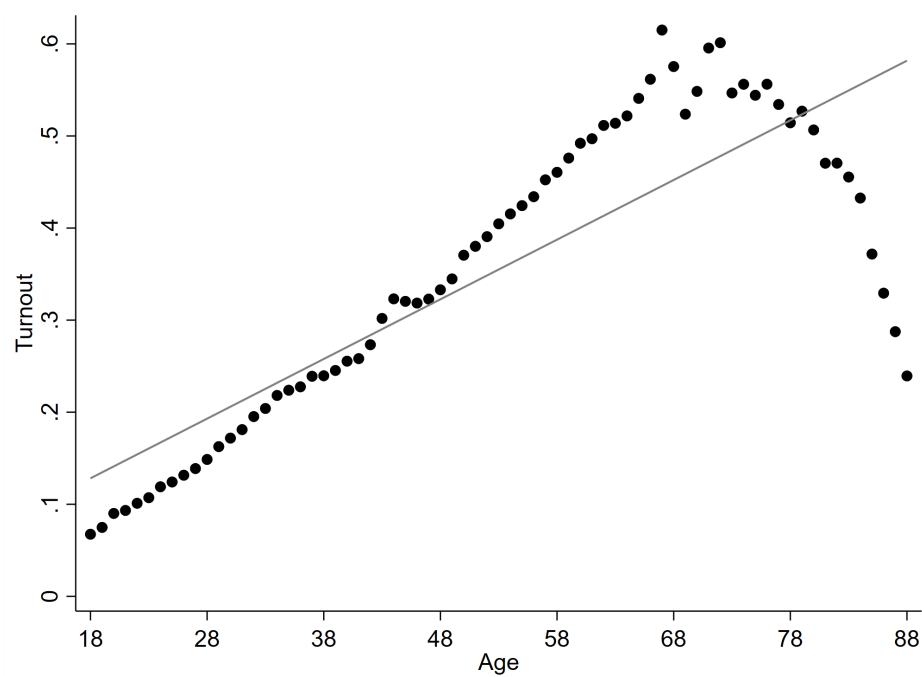


Figure 5.8: A regression line through voter turnout rate and age for all ages.

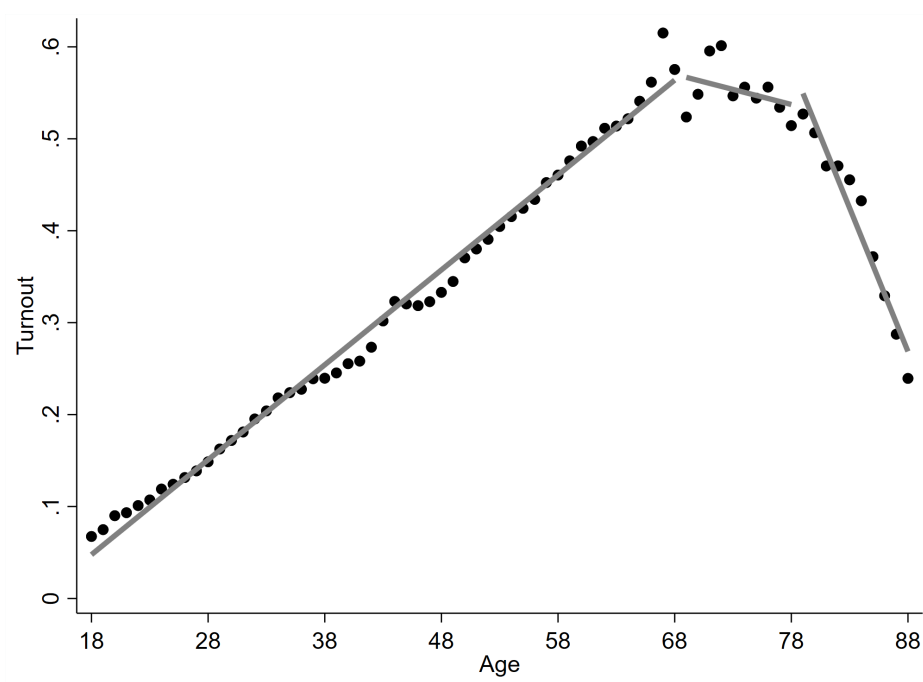


Figure 5.9: Separate regression lines through voter turnout rate and age for ages 16–68, 69–78, and 78–88.

When we took this approach in Chapter 2, we kept the regression simple. We just regressed the outcome variable against the explanatory variable squared. But we can do something more general than that. Instead of regressing voter turnout on just age or just age-squared, we can regress it on both. This allows us to fit a function that is more flexible than a line to our data. Of course, for each new variable we include, that's a new coefficient that we have to estimate when we minimize the sum of squared errors. But our computer can handle that.

In principle, we don't even have to restrict ourselves to different transformations of the age variable. We could also include other factors—average income or average voter registration status—which might further improve our predictions. We'll come back to that possibility in Chapter 10. For now, let's stick to transformations of our age variable.

With just one explanatory variable, it is easy to visualize what we are doing when we run a regression. We are just drawing a line through the data in a two-dimensional space. In particular, the line that minimizes the sum of the squared error.

With two explanatory variables, things are a little more abstract, but still manageable. Now we can think about finding a line going through our data in a three-dimensional space. Just picture adding a third axis coming out of the page toward you in our graphs. That axis will have the scale of the second explanatory variable (perhaps age-squared or income). Now the data forms a cloud in that three-dimensional space. Regression is still just drawing a line that minimizes the sum of squared errors, but now the line passes through that cloud of three-dimensional data points. Describing this line requires three parameters instead of two: the intercept (α), the slope with respect to changes in the first explanatory variable (we can call this β_1), and the slope with respect to changes in the second explanatory variable (we can call this β_2).

Once we go beyond two explanatory variables, it's hard to visualize the regression line, since most of us can't think in four or more dimensions. But you can analogize. You

understand what it means to find the line that minimizes the sum of squared errors with one or two explanatory variables. There is no reason we can't do the same with ten. Certainly your computer will have no trouble calculating the sum of squared errors and finding the OLS regression coefficients.

Let's see how this works in practice. We replicated the regression above, but included age-squared in addition to age as an explanatory variable. That is, we estimated the following equation:

$$\text{Predicted Turnout} = \alpha + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2$$

Once our computer calculates the associated regression coefficients, we can plug in any value of age and age-squared to get a predicted level of turnout. So, for instance, if we wanted to know the predicted turnout of 31 year olds, we'd plug in 31 for age and $31^2 = 961$ for age-squared.

And we don't have to stop at age and age-squared. Figure 5.10 shows the predicted turnout from different regressions, one with age and age-squared as explanatory variables (this is called a second-order polynomial); another with age, age-squared, and age-cubed as explanatory variables (third-order polynomial); another with a fourth-order polynomial; and finally one with a tenth-order polynomial!

The overall relationship between age and turnout is pretty complicated. As we've seen, it's approximately linear from 18 to 68, but then it takes a hard turn sometime after that. As a result, if we just include age, we don't do a great job fitting the data. Similarly, we see here that a regression with age and age-squared also doesn't do that well because the relationship in the data is poorly approximated by a quadratic curve. Our predictions get better and better as we include more and more terms, since we have more and more parameters that we can play around with to fit the data. By the time we get to a fourth-

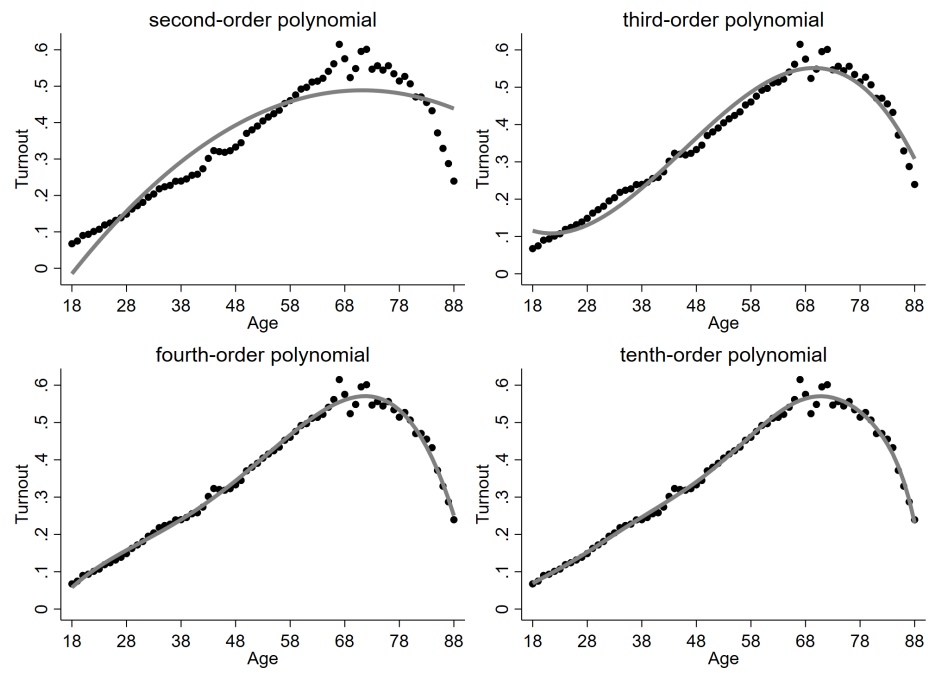


Figure 5.10: Fitting regressions through voter turnout rate and different polynomials of age.

order polynomial, the fit looks quite good.

Of course, the tenth-order polynomial does the best job approximating the data—the more explanatory variables included in the regression, the better the fit. But that doesn't necessarily mean that you want to include as many explanatory variables as possible. There are trade-offs that we have to be sensitive to.

For one, remember that part of our goal is to describe the data in a simple and parsimonious manner that is easy to understand and communicate. Describing the data with 11 parameters (α plus β_1 – β_{10} isn't much better in this regard than simply listing the turnout rate for each age group.

Furthermore, we often want to make *out-of-sample predictions*, forecasting voter turnout for age groups not actually observed in our data (like 90 year olds). Adding more and more terms often results in worse out-of-sample predictions. The reason is that, as the function we use becomes more and more flexible, it can read every little bump and hiccup in the data as meaningful, even when they aren't.

To illustrate this point, we repeated the analyses above, but ran the regressions only using data on people aged 18–78. Then we can see how well we do making out-of-sample predictions of the turnout rates for people with ages above 78. (These predictions are out-of-sample because we purposefully removed voters over the age of 78 from our data.) Figure 5.11 shows the results. The data used for fitting the regression are plotted in black. The data we are attempting to predict are plotted in red. And the gray curve represents the predicted values from the regression. As you can see, the fourth-order polynomial does well at predicting turnout for the oldest voters. But the tenth-order polynomial is a disaster!

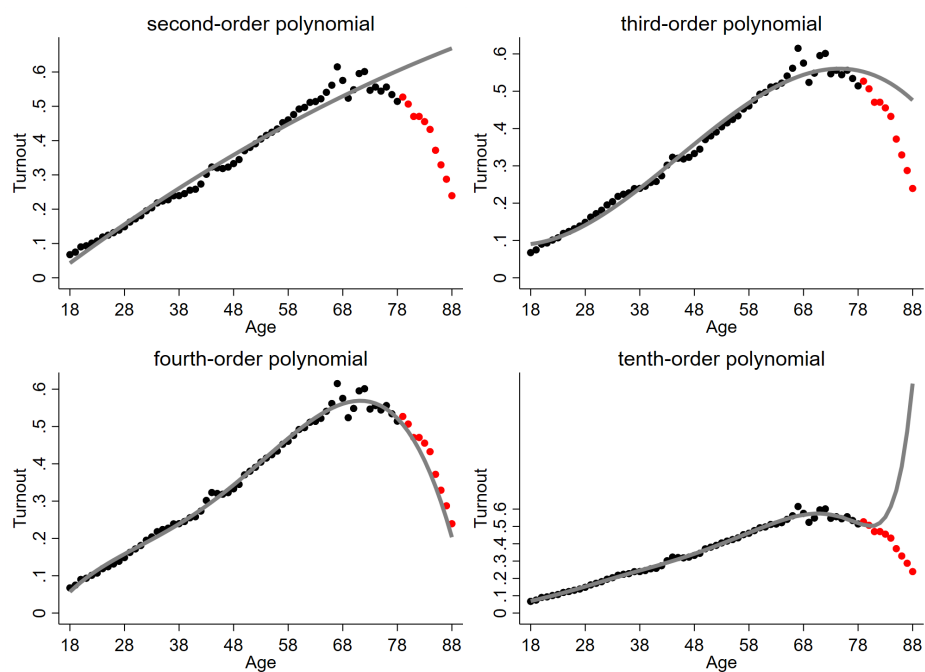


Figure 5.11: Using regressions through voter turnout rate and different polynomials of age to predict voter turnout for ages outside the range used for estimation.

The Problem of Overfitting: Predicting Election Results

The example we just saw, where the tenth order polynomial performed worse than the fourth order polynomial at out-of-sample prediction was an instance of a more general phenomenon called *overfitting*. To better understand overfitting, let's talk about a somewhat more realistic forecasting problem.

Americans are really interested in predicting the outcomes of upcoming presidential elections. When we tell people that we are political scientists, by far the most common question we get is “who’s going to win the next election?” We tend to disappoint with our answers, since election prediction is not what most political scientists spend their time on.

However, compared to most complex political phenomena, presidential elections are actually rather predictable. Even months before the election, we often have a pretty good idea of who is going to win based on how the economy is doing. And in the final weeks before Election Day, the average of polls usually gets within one or two percentage points of the final vote share. The journalist Nate Silver established himself as a giant of political data analytics by essentially averaging polls.

Of course, the fact that we can usually predict the vote share within one or two percentage points doesn’t mean we always know who’s going to win. Most presidential elections are highly competitive, and the Electoral College allows some candidates to win the election even while losing the popular vote. In close races, like in 2000 or 2016, given the available information on the morning of the election, there was probably no way an honest quantitative analyst could have been more than 90 percent sure that any particular candidate was going to win.

Although we said most political scientists don’t spend much time trying to predict election outcomes, some do. The academic journal *PS: Political Science & Politics* typically publishes a symposium before each presidential election with various attempts to predict

the outcome using quantitative data and analyses. Often, the goal of these analyses is to see how well they can predict the upcoming election results without using polling data. How well could we predicting vote share if we just knew the “fundamentals”, like economic growth and incumbency status?

To make such a prediction, a researcher might run a regression using historical data where each observation is an election, the outcome variable is the two-party vote share of the incumbent party in that election, and the various explanatory variables are features of that particular election like economic growth in the election year, whether the incumbent is seeking reelection, the number of war casualties over the past four years, etc. Having obtained the regression coefficients based on data from previous elections, the researcher can then plug in the values for explanatory variables from the current election and obtain a forecast for the upcoming two-party vote share. Because many other analysts are doing the exact same thing, the goal is often to find some new variable to include in the regression in order to improve its predictive power.

Mimicking this approach, we ran a regression predicting the incumbent’s vote share in presidential elections between 1948 and 2012. To be thorough, we include 10 different independent variables, all of which have been identified by political scientists as factors that might help us predict election results. Specifically, we include an indicator for whether the incumbent is a Democrat or Republican; an indicator for whether the incumbent is seeking reelection; GDP growth in years 1, 2, 3, and 4 of the most recent presidential term; an indicator for whether the country was involved in a major war at the time; a count of the number of consecutive terms in which the same party has been in power (many people expect that voters are more likely to replace a party that has been in power for a long time); the unemployment rate; and the change in the unemployment rate over the last 4 years.

We have good reasons to expect that these 10 variables should help us predict presidential election results, and at first glance, it looks like they do. The r^2 statistic from the

regression is .83, meaning that 83 percent of the variation in incumbent vote share appears to be accounted for by these variables. Furthermore, when we calculate predicted values from this regression, they only miss the actual vote share by an average of 1.7 percentage points.

The apparent success of our regression above, however, is misleading. It turns out, if we had simply generated 10 random variables (which we have done in computer simulations), and run the same regression using those meaningless numbers as our explanatory variables, we would have averaged an r^2 statistic of around .67 and an average error of 2.4 percentage points. This is almost as good as our predictions using real data, even though our 10 randomly generated variables were just noise, containing no information about the likely outcome of the election at all.

This is quite surprising. Why is it true? When you generate a bunch of entirely random variables, some of them are going to end up correlated with your outcome just by chance. In a regression, those meaningless variables will appear to predict the outcome. But, of course, they don't really. If you try to use the estimates generated by the relationship between those meaningless variables and past outcomes to predict future outcomes, you will fail miserably. Their predictive power is just an illusion created by chance.

This process whereby meaningless factors appear to predict things in the world is called *overfitting*. If we test enough explanatory variables, we're bound to find some that correlate with the outcome in our sample of data, even if those correlations are entirely the result of chance. This is why the tenth-order polynomial did such a bad job making out-of-sample predictions in the previous section. It was using meaningless correlations between high-order transformations of the age variable and voter turnout among one set of voters to try to predict turnout among another set of voters where, of course, those meaningless correlations did not continue to hold.

One way to try to assess and mitigate overfitting is by holding some data out of your

regression analysis and conducting out-of-sample tests—as we did with voters over the age of 78 in the previous section. In the context of predicting election outcomes, when generating a prediction for the vote share in 2012, we could leave the 2012 data out of the sample, run a regression using all the other elections, generate a predicted value for 2012 using those regression coefficients and the true values of the explanatory variables for 2012, and see how our predictions fare. In principle, we could do this for each year in our data set—remove one observation, run our regression, generate a predicted value for that observation, check our prediction against the truth, and repeat for each observation.

When we subject our regression using ten explanatory variables to prediction election outcomes to out-of-sample testing, it fares much worse than it first appeared. The average prediction error jumps up from 1.7 to 5.6 percentage points. We doubt that any campaign would hire a statistical consultant who could only promise to predict the election results within 5 or 6 percentage points, on average. Even more embarrassing, a naive prediction based on a simple average of the other elections in the sample gets within 4.6 percentage points, on average. In other words, the overfitted regression that we thought was giving us such accurate predictions is actually worse than a “null” regression that uses no explanatory variables.

Of course, when analysts are careful to avoid overfitting, they can generate useful predictions. A simple regression that uses only GDP growth in year 4 as an explanatory variable produces an out-of-sample prediction error of 3.8 percentage points, beating the null model. And if we included poll results as Nate Silver does, we would do even better. Nonetheless, it’s easy to trick yourself into thinking you’re generating good predictions when you’re not. Careful analysts only include variables in their regression that they believe are genuinely correlated with the outcome, they avoid having too many variables in their regression relative to the number of observations, and they validate their predictive strategy using out-of-sample testing.

How Regression Is Presented

Sometimes the outputs of a regression are presented graphically, as we have done thus far. But the most common form in which regression results are presented is in a table. For instance, Table 5.1 shows how the output of a regression of voter turnout on age might be presented.

DV = Voter Turnout	
Age	.0103 (.0066)
Constant	-.1381 (.0001)
R^2	.991
Root-MSE	.151
Observations	71

Table 5.1: Output of regression of average voter turnout on age.

You don't quite know what everything in this table means yet (we will come to standard errors in the next chapter), but most everything should be familiar. The number in the Constant row is the intercept, α . The number in the Age row is the slope of the regression line, β . We've also already discussed the idea of R-squared in Chapter 2. It is the amount of the variance in voter turnout that can be predicted from Age. And Root MSE is the square root of the mean squared error, so that gives you some sense of how far off, on average, our regression predictions are from the real data points.

A Brief Intellectual History of Regression

As far as historians of statistics can tell, regression was invented (or is it a discovery?) around the end of the 18th century. The first published instance of a linear regression is in the appendix to a brief book entitled *New Methods for the Determination of the Orbits of Comets* by the French mathematician Adrien-Marie Legendre. This was work with important implications for geodesy—the study of the measurement of the earth, which was a high-stakes problem, given the economic and military importance of navigation in the 18th century.

Legendre's status as the discoverer of regression was contested by a contemporary, the great German mathematician Carl Friedrich Gauss. In his 1809 *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*, Gauss staked his claim, writing, "Our principle, which we have made use of since 1795, has lately been published by Legendre." Legendre was not amused, and the two continued to snipe at one another over the matter throughout the early 19th century.

Neither Gauss nor Legendre referred to the method of drawing a line of best fit by minimizing the sum of squared errors as a *regression*. That term was coined by the late 18th century scholar, Francis Galton. Galton, a cousin of Charles Darwin's who was also married to Darwin's niece, was a polymath. He also came up with the idea for the modern fingerprinting system and was the first person to quantitatively document the wisdom-of-the-crowds phenomenon. More disturbingly, Galton was a eugenicist—a proponent of selective human breeding.

Galton's interest in eugenics led him to want to study evolution and heredity quantitatively. He started by measuring the easy things like height. In one analysis, he collected data on the heights of parents and their children. After plotting the data, he estimated the average relationship between these two variables using what we now call a regression line.

Galton's analysis was actually a little complicated. He compared the height of children to the average height of their parents after first adjusting the heights so that women's and men's heights were measured on the same scale. We don't want to go through all that. So, to get the idea, imagine an analysis like Galton's that studies just the heights of fathers and sons. The unit of analysis is a father-son pair, and the regression equation looks like this:

$$\text{Predicted Son's Height} = \alpha + \beta \cdot \text{Father's Height}.$$

When Galton estimated α and β using OLS, what do you think he found? We might have expected $\alpha = 0$ and $\beta = 1$. That would mean, on average, sons tend to be the same height as their fathers—i.e., we'd expect the son of a 5-foot-tall father to also be 5 feet tall, the son of a 6-foot-tall father to also be 6 feet tall, and so on. Instead, Galton was surprised to find $\alpha > 0$ and $\beta < 1$. Stop for a moment and think about why that might be.

Figure 5.12 demonstrates Galton's result graphically. The dashed black line shows the 45-degree line, that is the line with $\alpha = 0$ and $\beta = 1$. The thick gray line shows the best fitting regression line with $\alpha = 38.2$ and $\beta = .448$.

Let's start by interpreting these regression coefficients. The regression line lies above the 45-degree line for relatively short fathers and below for relatively tall fathers. This means that tall fathers tend to have sons that are taller than average but nonetheless shorter than they are. Similarly, short fathers tends to have sons that are shorter than average but nonetheless taller than themselves. Galton called this phenomenon, "regression to mediocrity." Today, we typically call this phenomenon *regression to the mean* or *reversion to the mean*, and we'll devote the entirety of Chapter 8 to understanding it. Since then, we've used the word *regression* to refer both to Galton's statistical technique and to the phenomenon that he discovered using it. So it's no coincidence that *OLS regression* and *regression to the mean* use the same word. They have a common intellectual history.

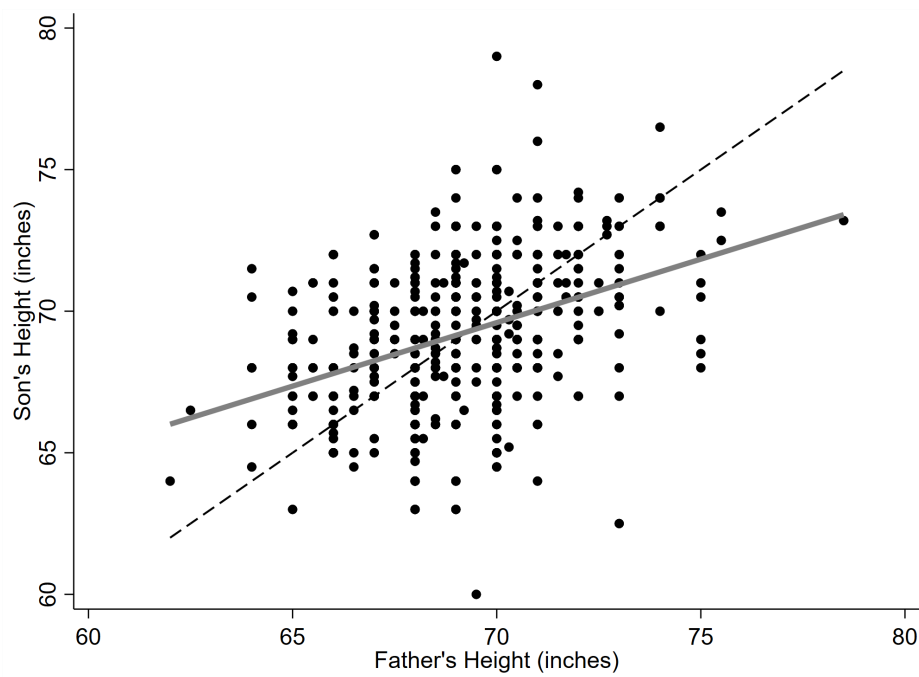


Figure 5.12: A regression line through son's height and father's height.

Key Terms

- **Line of best fit:** The line that comes closest to the data under some criterion.
- **Sum of squared error:** For a given line, calculate the error for each data point by finding its vertical distance from the line. The sum of squared error for that line is found by squaring each of the individual errors and adding them together.
- **Ordinary least squares regression (OLS):** The method for finding the line of best through data that minimizes the sum of squared errors.
- **Regression line:** The line of best fit through the data that one gets from OLS regression.
- **Conditional mean function:** A function that tells you the mean (average) of some variable conditional on the value of some other variables.
- **Dependent variable:** The variable associated with the outcome we are trying to describe, predict, or explain.
- **Independent or Explanatory variable:** A variable we are using to try to describe, predict or explain the dependent variable.
- **Out-of-sample prediction:** Using regression (or another statistical technique) to predict the outcome for observations that were not included in the original data you used to estimate the function you are using to do the prediction.
- **Overfitting:** Attempting to predict a dependent variable with too many independent variables, so that variables appear to predict the dependent variable in the data but have no actual relationship with it in the world.

Exercises

Download data from [INSERT LINK HERE] on years of schooling and earnings. The data set gives the average annual earnings for 41-50 year old men in the U.S. in 1980 at each level of schooling. One observation gives the average earnings (in thousands of dollars) for men with 8 years of schooling, another gives the average for those with 9 years of schooling, and so on.

1. Run a regression with earnings as the dependent variable and schooling as the sole independent variable. Interpret the estimated coefficients.
2. Suppose you wanted a parsimonious way to predict earnings using only years of schooling. What would you do?
3. Let's dig more deeply into whether the relationship between earnings and schooling is approximately linear.
 - (a) Start by making a scatter plot. Then plot the predicted values from your regression along with the raw data points as we did in Chapter 2. Does the estimated regression line look like it's fitting the data well?
 - (b) What if you run a fourth-order polynomial regression (i.e., include schooling, schooling², schooling³, and schooling⁴? Do those predictions meaningfully differ from the predictions coming from the linear regression?
 - (c) What if you run different regressions for different ranges of schooling? Do those lines look meaningfully different than the predictions you get from a single regression including all the data?
 - (d) Does all this make you think the simple linear approach was reasonable or unreasonable?

4. Similar to what we did with age and voter turnout, conduct some out-of-sample tests to evaluate your prediction strategy. Using only data for those with 12 years of schooling or less, see how well your different strategies from question 3 perform when predicting earnings for those with more than 12 years of schooling.

Notes

For more information on the early history of regression, see

Stephen M. Stigler. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap, Harvard.

And if you are generally interested in the history of probability and statistics (which you should be, it is very important), have a look at

Ian Hacking. 2006. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability and Statistical Inference*, 2nd Edition. Cambridge, UK: Cambridge University Press.

Ian Hacking. 1990. *The Taming of Chance*. Cambridge, UK: Cambridge University Press.

Chapter 6

Inferences about Relationships

What you'll learn

- All quantitative estimates are the sum of three terms: the true quantity of interest, bias, and noise.
- Statistical hypothesis testing allows analysts to assess whether an estimate was likely to have arisen from noise.
- Statistical significance and substantive significance are not the same and should not be conflated.

Chapters 4 and 5 articulated tools that allow us to describe a relationship between variables within a data set. We need variation in both variables and then we can describe the correlation between those variables using regression. But often we want to go further. We want to use the relationships between variables that we find in the data we have to make inferences about relationships that hold between those variables in the larger world (i.e., in data we don't have). For instance, once we've found that crime is higher on warm days in 2018, we'd like to know whether we are justified in concluding that this relationship is likely to hold in other years and isn't simply an artifact of the 2018 data. That is, we want to know whether an observed relationship reflects a genuine phenomenon in the world or whether it happens to be true in the sample of data that we looked at by chance. In this chapter, we discuss some tools that help us to adjudicate between these possibilities.

Estimation

To start thinking about this question, we need a common language to talk about the differences between the things we observe in our data set and the phenomena in the world that we'd like to learn about. To do so, we are going to use the following simple equation, which will come up so often throughout the book that, from now on, we are going to start calling it *our favorite equation*:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

We are going to define each of these terms carefully as we go. But before doing so, let's set the stage with a simple example.

Suppose we conduct a poll to learn which of two candidates (a Republican and a Democrat) is going to win an upcoming election. We can think of this as a prediction problem,

we are collecting data to forecast the future winner. But we can also think of this as pure description, we want to know the proportion of voters who support one candidate over the other.

In either case, a key challenge is that there are too many voters for us to ask all of them their opinions. Necessity forces us to poll a *sample* constituting only a small proportion of the total *population* of voters. Thus, we need to figure out what we can conclude about political views in this larger population from evidence generated by a poll of only a relatively small sample.

In our example, we are interested in learning the proportion of voters in the population who support the Republican. Let's call that proportion p . Since there are only two candidates, the proportion who support the Democrat is just $1-p$. Until we actually hold the election, we don't get to observe p so we have to try to estimate it.

Suppose we poll a random sample of 100 voters and ask them whether they will support the Republican or Democrat. We could estimate the number of voters in the population who support a Republican (which we can't observe) by calculating the proportion of people in our sample who support the Republican (which we can observe). Let's call our estimate from our sample $\hat{p}$. Our hope is that our estimate, $\hat{p}$, is close to the truth, p .

We can now introduce some important terminology with reference to this example. The true proportion of Republicans in the population (p) is called the ***estimand***—it is the unobserved quantity that we are trying to learn about with our data analysis. The process of sampling 100 voters and calculating the proportion who support the Republican is called the ***estimator***—it is the procedure we apply to generate a numerical result. The proportion of Republicans in our sample ($\hat{p}$) is called the ***estimate***—it is the numerical result arising from the application of our estimator.

With these definitions, we can take a first step towards making sense of our favorite

equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

The quantity we are interested in is the estimand. The quantity we observe in the data is the estimate. In an ideal world, the estimate would equal the estimand, so that our estimator would reveal to us the true quantity of interest. But our favorite equation says this isn't the case. Estimates differ from estimands because of bias and noise. To understand why, we need to learn more about those two troublesome quantities.

Why Do Estimates Differ From Estimands

Bias and noise are both important. But they differ, and their difference are often lost on people, leading to unclear thinking. So we'll take them in turn. But before discussing bias and noise in detail, an analogy might help.

Anthony likes to play the Scottish game of curling. In curling, two teams take turns sliding heavy granite stones down a long sheet of ice. As a stone slides, other team members sweep in front of the stone like crazy while running along the ice to try to smooth the path. We recommend watching a video, it's pretty fun. Anyway, the team with the stone closest to the center of a target on the far end of the sheet of ice (called the button) scores points.

Anthony's quite good at curling. He can more or less get his stone to go where he wants. But, despite his skill, sometimes his stones miss the button. Why is this? Well, there are all sorts of unobservable little factors that affect how the stone slides. Maybe the ice melted a bit funny part way down the sheet, causing a well-aimed stone to slide a little too far to one side or the other. Or maybe Anthony's hand gave an involuntary twitch as he released the stone. Or maybe Ethan slipped on the ice while trying to sweep and that accidentally caused the stone to slightly divert course. Anyway, for all of these reasons, Anthony's well

aimed stones might miss the button.

Ethan, by contrast, is terrible at curling. So when he's gone curling with Anthony, his stones frequently miss the target, typically to the left (let's not even talk about distance). He'd like to claim it is because of idiosyncratic factors, like with Anthony's misses. But if that were true, he wouldn't be more likely to miss left than right. No, the truth is, Ethan's technique is poor, so his stones are systematically improperly aimed.

There is, we think, a useful analogy between curling and data analysis. Think of the button as like the estimand, it is the truth you are aiming at. Think of your estimator as like the act of sliding the stone down the ice. And think of the outcome of one stone throw as like an estimate from one iteration of the estimator (i.e., running your procedure on one sample of data).

Your stone (estimate) might miss the button (estimand) for two reasons. First, like Anthony, you may have aimed well, but random factors may have moved the stone a little bit. These random factors are like noise. Since these factors are random, on average, they don't make Anthony miss more to the left or to the right. Indeed, on average, his stone's location is on the button. But that doesn't mean every individual stone lands on the button, it's just that his misses average each other out. This is just what noise does—estimates can equal the estimand on average, but because of noise, any given estimate may not equal the estimand.

Second, like Ethan, you may systematically aim too far to the left. There is still noise, so you might sometimes miss to the right. But, on average, your stone lands to the left of the button. These systematic errors are like bias. Unlike Anthony, Ethan's average stone misses the button. This is what bias does—when there is bias, even the average estimate doesn't equal the estimand, let alone any given estimate.

Ok, now that we have an analogy to help us see the difference between them, let's talk

about bias and noise in a bit more detail.

Bias

One reason an estimator might give you an estimate that differs from the estimand is because it is *biased*. Imagine applying your estimator over and over again an infinite number of times, each time to a new, independent sample of data. Doing so would generate an infinite number of estimates. Because of noise, some of those estimates will be bigger than the estimand (i.e., you'll get a larger share of Republicans in some of your samples than there are in the population) and some of those estimates will be smaller than the estimand (i.e., you'll get a smaller share of Republicans in some of your samples than there are in the population). But you would like the average of that infinite number of estimates to be equal to the estimand. That is, you don't want to predictably (or systematically) overestimate or underestimate the number of Republicans. You want to be aimed at the truth. We say that an estimator is *unbiased* if the average value of the estimates it generates would equal the estimand if we repeatedly applied the estimator to new, independent samples an infinite number of times.

There are lots of reasons a political poll might be biased. Suppose voters systematically lie to pollsters. Perhaps voters believe that pollsters are themselves likely to be Democrats and the voters want to please the pollsters, so some Republican voters report supporting the Democrat. Then our estimator will be biased in favor of Democrats—on average, reporting more voter support for Democrats than there really is. Or suppose Republicans are more likely to turn out to vote than Democrats, but equally likely to answer polls. Then poll respondents will differ from voters and the estimates from polls will be biased in favor of Republicans—on average, reporting more voter support for Republicans than there actually is. Finally, what if pollsters contact people by phone and phone owners are systematically different in their political leanings than the population as a whole? This will also lead to

bias. For lots of reasons, if we ran the poll an infinite number of times and averaged the estimates, that average might not equal the true proportion of Republicans in the population of voters, which is our estimand. Thus, the poll could be biased for any of these reasons.

In subsequent chapters, we will be very concerned with thinking about sources of bias. For the duration of this chapter, we are going to ignore these potential sources of bias to focus on the second potential problems with estimators, noise.

Noise

When you take a sample of the population, you inevitably introduce some noise into your estimate. When you ask 100 randomly selected people out of 100 million their opinions on political candidates, sometimes by chance you happen to talk to a disproportionate share of Republicans and sometimes you happen to talk to a disproportionate share of Democrats. As a result, even without bias, any individual estimate will not equal the estimand. Suppose your estimand is unbiased. If you applied it an infinite number of times, you would not over-estimate Republican or Democrat support, on average. But each individual estimate would not equal the estimand because of noise—that is, natural variability that results from sampling. This natural variability is sometimes referred to as *sampling variation*.

We have ways of quantifying the amount of sampling variation associated with an estimator. Think about repeatedly applying an estimator with new, independent samples of data an infinite number of times. The closer the various estimates would be to each other, the more *precise* is the estimator. Thus, a more precise estimator is one with less sampling variation.

What Makes for a Good Estimator?

In the end, we are trying to learn the true value of the estimand. Since our estimate can differ from the estimand because of bias or noise, what we really want is an estimator that is both unbiased and precise.

If our estimator is unbiased but imprecise, our estimates will typically differ from the estimand because there is so much noise. For instance, in our polling example, if we talk to just 1 voter at random, his or her opinion is an unbiased estimate of the average opinion in the electorate (if you did this an infinite number of times, p of those times you'd get a Republican and $1 - p$ of those times you'd get a Democrat). But the sampling variation associated with estimating voter opinion by asking just one person's opinion is huge—we will always estimate either 100% Republicans or 100% Democrats.

If our estimator is biased but precise, our estimates will typically differ from the estimand because they are very precisely estimating the wrong quantity. For instance, if we sample 10,000 voters, but only do so in Republican neighborhoods, we will get answers clustered very tightly around each other, but they will systematically overestimate the number of Republicans.

An estimator that is both precise and unbiased is called *efficient*.

Figure 6.1 illustrates that estimators can be unbiased, precise, neither, or both. The black diamonds show the estimand—the true value in the world we are interested in. The gray dots show various estimates that arise from repeated applications of a given estimator each time with an independent sample of data. If the gray dots are symmetrically distributed around the diamond (like Anthony's curling stones around the button), the estimator is unbiased. That is, the estimates it provides are right on average. If the gray dots are clustered tightly together, the estimator is precise. That is, there is very little noise so the estimator yields similar estimates with each iteration. All else equal, we would obviously

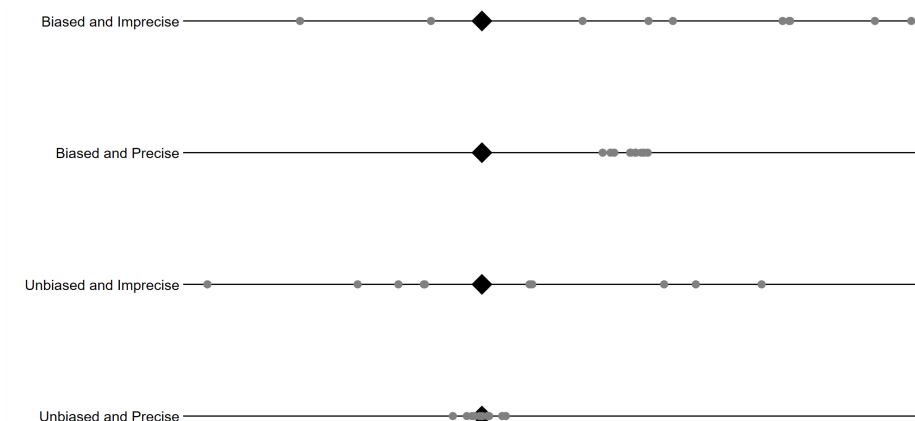


Figure 6.1: Understanding the difference between unbiased and precise.

like to have our estimator be less biased *and* more precise. However, sometimes there are trade offs between these goals, and we have to decide how much bias we're willing to accept for a certain gain in precision.

For a concrete example of the possible trade-offs between bias and precision, think about the weather forecast of tomorrow's temperature. What would it mean for a meteorologist to be unbiased? Their forecast might be too high on some days and too low on other days, but those errors would cancel out such that they're right on average. Unbiasedness is a good feature of a weather forecast. But it isn't everything. If, on a typical day, the weather forecast is systematically off by 30 degrees Fahrenheit, it is cold comfort to know it is equally likely to be 30 degrees over and 30 degrees under the true temperature. Clearly, unbiasedness is not the only thing you'd want from your local meteorologist. Some precision would be nice as well.

Notice, if an estimator is unbiased, we'd also like it to be as precise as possible. And as we've discussed, we might even allow for a little bias in exchange for a big gain in precision. But if an estimator is really biased, it's no longer obvious that precision is a good thing. For one thing, a precise biased estimator will never be anywhere close to the truth. Whereas

with less precision, you might sometimes make good predictions despite the bias, albeit by accident. (Ethan would probably be better off if there was an earthquake just as he released his curling stone because at least sometimes his stone would stay in play.) Furthermore, precision might give you a false sense of confidence. Beware a precise estimate with an unknown bias.

Quantifying Precision

Remember the motivating question for this chapter: when we estimate something from a sample of data, how confident should we be in drawing inferences about the larger population? As we've seen, if our estimator is biased, we should certainly be worried. But even if our estimator isn't biased, we still have to be worried that our estimates do not reflect the true relationship in the larger population (the estimand) because of noise. In order to know how worried we should be about this possibility, we need to quantify the precision of an estimator. We do so through a statistic called the *standard error*.

In Chapter 2, we learned about the idea of a standard deviation, which is a measure of how spread out a distribution is. Imagine that we repeated our estimator an infinite number of times, each time with a newly drawn set of data. That would give us a distribution of estimates—each time we draw the data and run our estimator we get a different estimate because of the noise. That imagined distribution is called the *sampling distribution*. The standard deviation of that imagined sampling distribution is called the *standard error*. The standard error, if we knew what it was, would give us a sense of how far any given estimate will be from the average estimate. If the estimator is unbiased, the average estimate equals the estimand. So, for an unbiased estimator, the standard error tells us approximately how far a typical estimate is from the estimand, which is the true value we are trying to learn about.

If the standard error is large, then the estimates would be very spread out and the estimator is relatively imprecise (i.e., there is a lot of sampling variation). If the standard error is small, then the estimates would be very close together and the estimator is relatively precise (i.e., there is little sampling variation).

We can provide some intuitions about when an estimator is precise versus imprecise (i.e., when the standard error will be small versus large). In our polling example, the standard error is approximately equal to $\sqrt{\frac{p(1-p)}{N}}$, where N refers to the sample size—i.e., the number of people polled. While we aren't going to show you how to derive this formula (a topic for a more advanced class or book), we can learn some things about what makes an estimator more or less precise by thinking about the formula.

Let's start with understanding the numerator, $p(1-p)$. Notice, this term is maximized at $p = \frac{1}{2}$ and decreases as p gets larger or smaller. So, suppose the true proportion of Republicans in the population, p , is either very large (close to 1) or very small (close to 0). This makes $p(1-p)$ very small and, therefore, makes the standard error small. Why? When p is very large or very small, there is little possibility of sampling error. If 99% of voters are Republicans, when you collect your sample of, say, 1000 voters, it will be very unlikely that you find many Democrats. By contrast, if p is close to one-half, the standard error is large. This reflects the fact that there is lots of room for sampling error. You could easily find a 55-45 or a 45-55 split in your sample of data drawn from a 50-50 population. The closer p is to one-half, the more natural variation there is in our outcome of interest, making the standard error is larger.

Now consider the denominator. It tells us that as the size of our sample increases, our standard error goes down. This makes sense. The problem of imprecision comes from the fact that our sample might not accurately reflect the whole population. When the sample is large, it will more closely approximate the population. We can more accurately estimate the opinions of a million people by talking to 10,000 people than by talking to 10 people.

But the formula actually tells us something a little more subtle than that. It tells us that the standard error shrinks in proportion to $\sqrt{N}$. Suppose the true proportion of Republicans in the population is $p = 0.5$. Then, if we took a poll of 1000 voters, we'd have a standard error of $\sqrt{\frac{.5 \cdot .5}{1000}} \approx .016$. Suppose we conducted a much larger poll of 10,000 people. Then our standard error is $\sqrt{\frac{.5 \cdot .5}{10,000}} = .005$. So increasing the sample size by a factor of 10 only improves the accuracy of the poll approximately 3-fold. If we further increased the sample size to 100,000, we'd get another roughly 3-fold improvement in accuracy ($\sqrt{\frac{.5 \cdot .5}{100,000}} = .0016$). In other words, the formula for the standard error tells us that there are diminishing returns to bigger and bigger sample sizes. The standard error of a survey with 100,000 respondents is negligible, despite the fact that 100,000 is still less than one-one thousandth of the entire population of interest.

One tricky thing you might have noticed is that we need to know p in order to calculate the standard error. But we don't know p , that's why we're doing the poll to begin with. In practice, we approximate the standard error by substituting $\hat{p}$, our estimate of p , into the formula. Of course, this approximation would run into problems if you had a really small N or a value of p that was really close to 0 or 1. Suppose you talked to 5 people and found that none of them were Republicans. Blindly applying the procedures above, you would wrongly conclude that nobody is a Republican and that your standard error is 0. Of course that's wrong, and that's because with small samples and with extreme values of p , your approximation using $\hat{p}$ is misleading.

We should also point out that although there is a nice formula for the standard error in our polling example, this won't always be the case. Fortunately, our computers can often produce reasonably reliable approximations of standard errors, even in more complicated circumstances.

Statistical Inference and Hypothesis Testing

Now we can finally turn to this chapter's motivating question: How do we make inferences about populations using estimates from samples? Let's stick with our polling example. As we emphasized, when we conduct a poll, even if we think it is unbiased, we also want it to be precise because we want to know that it is giving us an estimate that's close to the truth. How can we assess this? How well does a sample of, say, 1,000 voters estimate the views of the 140 million Americans that will determine who wins an upcoming presidential election? Let's see.

Confidence intervals

An important mathematical fact, called the *Law of Large Numbers*, tells us that as our sample size gets really big, the noise will essentially disappear. But how big is big enough? The standard error will help us figure this out.

Another important mathematical fact, called the *Central Limit Theorem*, tells us that if our poll is indeed unbiased, then if we were to repeatedly run our poll, approximately 95% of our estimates ($\hat{p}$, Republicans in our sample) will end up being within approximately 2 standard errors of our estimand (p , Republicans in the population). Therefore, pollsters will often report what they call the *margin of error*, which is simply twice the standard error.

Researchers and pollsters also sometimes report what they call the *95% confidence interval*. This is the interval that ranges from the estimate ($\hat{p}$) minus two times the standard error up to the estimate plus two times the standard error.

The 95% confidence interval is a source of some confusion. Often, people will casually say that we're 95% confident that the true value lies within the 95% confidence interval. But that's not quite right. The correct statement is a lot clunkier. Technically, we can say

that if we repeated our estimator an infinite number of times, the true estimand would be inside the 95% confidence interval 95% of the time.

To get a picture in our head of how confidence intervals work, let's go back to our curling analogy for a second. Suppose Anthony pushes an infinite number of stones down an infinite number of ice sheets. Think of the spot on the ice where the exact center of his stone comes to rest as the estimate. That estimate is extremely unlikely to be sitting exactly on the button—your estimate is almost never exactly equal to the true value of the estimand. But the stone is wider than that one spot. So we might ask how often the button will be touching some part of the stone. That will depend on how wide the stone is. (Of course, there is a regulation width in curling, but allow us a little poetic license. We aren't Scottish. And this isn't the Olympics.) We could find the exact width of a stone such that the button would be touching some part of the stone on 95% of Anthony's throws. This is like the 95% confidence interval. It isn't that, on any one throw, we're 95% confident that the button is covered by some part of the stone. It is that, on 95% of throws the button is covered by some part of the stone.

This analogy can also help us think about confidence intervals other than the 95% confidence interval. Sometimes we want to be more confident than 95%. So we might be interested in the 99% confidence interval. Ask yourself whether the 99% confidence interval is wider or narrower than the 95% confidence interval. If Anthony wants 99% of his throws to end with some part of the stone touching the button, he is going to need to use a wider stone. So the 99% confidence interval is wider than the 95% confidence interval. We have to admit a larger range of possible Republicans if we want to be sure our estimate is within that range 99% of the time rather than just 95% of the time.

Hypothesis testing

How do confidence intervals help us know if we should have confidence that our estimate reasonably approximates the truth? For that, we need to think about *hypothesis testing*. Often, when we conduct statistical analyses, we'd like to test a particular hypothesis. In the example of an electoral poll, we might like to know which candidate is going to win the election. Suppose we conducted an unbiased poll of 1000 voters and this yielded an estimate of the Republican candidate's vote share of .532 or 53.2 percent. How confident should we be that the Republican is actually going to win the election—i.e., that at least 50% of voters prefer the Republican? Hypothesis testing provides a way for us to say something about that.

One way of thinking about this question is as follows. We have some evidence from our poll that the Republican candidate is more popular than the Democrat. But we want to know how good that evidence is. That is, we want to know how likely it is that we could have observed such evidence *even if the Republican is not more popular than the Democrat*. So we test how likely it is that we would have observed the evidence we observe if the two candidates are actually equally popular. We often call this the *null hypothesis*.

To understand the hypothesis test, start by assuming the null hypothesis is true—i.e., the two candidates are exactly equally popular, so $p = 0.5$. Now ask how likely would it be for us to obtain a poll result at least as favorable for the Republican as the one we actually found, $\hat{p} = 0.532$.

We already have the information we need to answer this question. With a true vote share of $p = 0.5$ and a poll of 1000 voters, the standard error of our estimator is approximately 1.6 percentage points ($\sqrt{\frac{.5 \cdot .5}{1000}} \approx .016$). Our estimate of .532 is 2 standard errors above the null hypothesis ($.5 + 2 \cdot .016 = .532$). We didn't choose these numbers by accident.

As we said earlier, the Central Limit Theorem tells us that 95% of estimates from a

unbiased poll of the sort we ran will fall within 2 standard errors of the truth, meaning that only 5% of estimates will fall more than 2 standard errors from the truth. Moreover, in half of those unfortunate cases, the estimate will be 2 standard errors *below* the truth (i.e., results showing the Democrat way ahead). So if the null is true, the probability that we'd get a poll result as favorable for the Republican as the one we got is about 2.5%, or 1 in 40.

We obviously picked numbers that would make this calculation straightforward, but your computer can do this calculation for any poll result. Statisticians are in the business of developing methods for conducting these calculations. If you want to know some more statistics lingo, the analysis we just did is called a one-sided z-test. You don't need to know about z-tests to understand the rest of this book, but if you want to learn about them, you can consult virtually any statistics book (or Wikipedia). More generally, the important thing is that hypothesis testing is a strategy for assessing the probability of getting a result as extreme as yours under the null.

Statistical significance

We just saw that if the null is true, the probability we would have gotten a result as favorable to the Republican as what we found is only .025. We often call this probability a *p-value*. If our p-value is really low, then we might conclude that the null is unlikely to be true. Thus, we have some statistically compelling evidence that the Republican is indeed favored by voters—if the true vote share were evenly split, it's unlikely that the poll result would be this favorable to the Republican (and even more so if the true vote share favored the Democrat).

A common strategy is to pre-specify a particular threshold (most commonly, .05), and if the p-value is below that threshold, then we say we reject the null hypothesis and conclude that we have *statistically significant* evidence for the hypothesis we were testing against the null hypothesis.

Of course, hypothesis testing does not provide certain conclusions. Quantitative analysts often select a significance threshold of .05, which means that there's a 5% chance of obtaining a statistically significant result even if the null hypothesis is true. But hypothesis testing provides one way of thinking quantitatively about whether a pattern or result you have detected in your data set is likely to reflect a genuine phenomenon rather than simply being the product of random noise.

One common error is to assume that the p-value tells you the probability that the null hypothesis is true. It doesn't. It tells you the probability of getting an estimate as extreme as the one you got if the null is true. Those two numbers are typically different. Indeed, to calculate the former quantity, you'd have to have a lot more information (e.g., how likely you thought it was that the null was true before you saw the evidence). We'll discuss these issues in Part IV.

Statistical Inference about Relationships

So far we've developed our ideas about bias, noise, and hypothesis testing in a simple setting where we are just trying to learn about the share of voters who support the Republican candidate. But all these concepts and tools of statistical inference can be applied to much more interesting problems, including estimating relationships, like correlations.

Suppose we ran a regression to estimate the relationship between some outcome variable Y and some explanatory variable X (think about crime as Y and temperature as X). The previous chapter discussed how we can utilize linear regression to estimate a coefficient that describes the relationship between two variables in a data set. But now let's motivate this in a different way, having to do with inference.

Suppose our data set consists of information on the income and education of a random sample of 1,000 workers, but we are actually interested in the average relationship between

income and education in the entire population of all workers. So we are trying to make inferences about a correlation in the population (our estimand) based on a correlation in the data (our estimate). How do we do this?

Start with the following equation, which describes the average relationship between income and education, in the population.

$$\text{Income}_i = \alpha + \beta \cdot \text{Education}_i + \varepsilon_i.$$

Income_i is person i 's income, Education_i is person i 's years education, and ε_i is an error term specific to person i reflecting all the idiosyncratic facts about person i besides their education that matter for their income. The parameters α and β take whatever values that minimize the sum of squared ε_i 's across the population. For example, β is the average extent to which income increases with each additional year of education in the population. These parameters α and β are features of the world. We don't know them. But, since we would like to know, on average, how income changes with education, β is our estimand.

Although we don't know what β is, we can estimate it applying linear regression to our data on 1000 workers. Let's call the estimate from that regression $\hat{\beta}$. It is the regression coefficient and it reflects the correlation between education and income in our sample.

Unfortunately, β and $\hat{\beta}$ are not the same thing. The former is an estimand and the latter is an estimate, which, as we know from our favorite equation, may differ from the estimand because of both bias and noise. Let's assume that our sample of workers was randomly selected from the population, so there is no bias. (We'll talk more on random sampling and unbiasedness in Chapter 11). But there is still noise. So if we want to know how close $\hat{\beta}$ is likely to be to the true β , we're going to need to think about the standard error.

Just like $\hat{p}$, our estimate of the proportion of Republicans in the population, had a

standard error, so too does our estimate of the relationship between income and education, $\hat{\beta}$. The standard error tells us how far, on average, our estimate would be from the truth if we repeated our estimator an infinite number of times with independent samples of data. Just as with the poll result, there are formulas for calculating this standard error. For now, you don't need to worry about the formula because a computer will calculate the standard error for you. Thinking more technically about standard errors is a topic for a more advanced statistics book.

Once you've estimated the standard error associated with a regression coefficient, you can do the same kinds of things that you did with the standard error of the poll result. For instance, you can construct a 95% confidence interval—if you repeated the procedure an infinite number of times, the estimated confidence interval would contain the truth 95% of the time. You can also conduct hypothesis tests and compute p-values. All of this can help you assess how noisy or precise your estimate of the true relationship is.

One common question people are interested in is whether there is compelling evidence that there is any true relationship at all. That is, suppose you find a positive $\hat{\beta}$; income and education are positively correlated in your sample. Should you be confident, based on that evidence, that they are positively correlated in the larger population?

You can start to answer that question by testing the null hypothesis that the true relationship between income and education is in fact zero. To do so, you ask how likely it is that you would have gotten an estimate as big as $\hat{\beta}$ if there was in fact no correlation between income and education in the population. If you obtain a small p-value and reject the null hypothesis, then you have good reason to believe that there is a relationship between income and education in the population.

One reason statistical inference of this sort is so useful is that we're bound, from time to time, to find relationships in our sample that do not reflect genuine relationships in the larger population. This is the nature of noisy data. So we need to check whether we have

good reason to believe our findings aren't just the result of noise.

What If We Have Data for the Whole Population?

Sometimes we have data for an entire population of interest. For instance, suppose we want to know the correlation between participation in varsity athletics and GPA for University of Chicago students. It's conceivable that we could convince the university to give us the relevant data for every single student, in which case we wouldn't need to estimate the correlation in the population by calculating it in a sample. We could perfectly measure the true estimand, the correlation between athletics and GPA for the whole population of University of Chicago students.

Here's a tricky question. Does it still make sense to think about standard errors, confidence intervals, and statistical significance when we have data for the entire population? One argument is that these tools are irrelevant because there was no sampling. So there is no noise. The estimate is the estimand. As such, there's no need to think about statistical inference.

But we still think there are good reasons to pay attention to the concept of noise and the associated measures of uncertainty even if we have data for an entire population. Let's talk about why.

Suppose we found a small, positive correlation between playing a varsity sport and GPA. It still seems reasonable to ask whether that difference arose for a reason or whether it arose just by coincidence.

What would it mean for a correlation to arise by coincidence? Suppose that there's no good reason to think there should be an on average difference between the GPAs of athletes and non-athletes. The admissions standards are the same for both kinds of students, athletic participation has no effect on GPA, academic performance has no effect on athletic

participation, and so on. Nonetheless, there are all kinds of idiosyncratic differences between students that lead their GPAs to be different from one another. And with any finite number of students, there's bound to be at least a slight difference between the GPAs of athletes and non-athletes, even if there's no good reason for the difference.

To assess whether this observed correlation arose by coincidence or for a reason, we can't collect more data. We already have all the data there is to have on University of Chicago varsity athletes. However, it turns out that the tools of statistical inference and hypothesis testing still provide a useful way to think about whether an observed pattern was a coincidence or not.

One way to think about the problem is that, although we have data on all the actual students at the university, these actual students are just a small sample of a much larger hypothetical population of students that could have been at the university. We can start with the null hypothesis that the true correlation in that larger hypothetical population is zero, and we can ask how likely it is that we would have observed a correlation as large as the one we observed among the actual students by chance. Of course, doing this requires a metaphysical leap from actual populations to hypothetical populations. But engaging in a little bit of metaphysics is a price we think is probably worth paying to preserve our ability to think about whether some observed relationship reflects a genuine, predictable pattern or whether it was a fluke.

Substantive vs. Statistical Significance

Statistical hypothesis testing is often helpful and informative because we want to know whether an observed phenomenon is likely to have arisen purely by chance (e.g., due to sampling variation). However, *statistical* significance (i.e., a low p-value indicating that a result was unlikely to have arisen by chance) is not the same thing as *substantive* signif-

icance, and we must be careful not to conflate these two concepts. Often, we don't just want to know whether some phenomenon *exists* or not, which is the question of statistical significance. We want to know how big or small it is because that will tell us whether the phenomenon is *important* or not. For example, Coca-Cola probably already knows that their marketing has some effect on sales. But that doesn't help them know how much to spend on marketing. For that, they want to know how big these effects are. Let us give you examples that illustrate the two ways quantitative analysts can draw misleading conclusions by emphasizing statistical significance over substantive significance.

In 2012, six researchers published a study in *Nature* showing that people were more likely to vote in the 2010 U.S. midterm elections if their Facebook page displayed a banner indicating which of their friends voted. The study was notable for several reasons. Facebook allowed the researchers to randomize the experience of 61 million voting-age Facebook users in the U.S. on Election Day. And indeed, the experimental intervention appears to have increased turnout—the estimated effect is highly statistically significant ($p = .02$). The researchers concluded that “strong ties are instrumental for spreading both online and real-world behaviour in human social networks.” And the study received significant press coverage for demonstrating how important social pressure is for voting.

What most observers failed to notice—and the authors of the study chose to downplay—was that the estimated effect of the Facebook banners on voter turnout was less than 0.4 percentage points. This is a substantively small effect arguably of little relevance for campaigning or understanding elections. The fact that 0.4 percent of eligible voters can be persuaded to vote through a Facebook banner reporting their friends' voting choices does not tell us that strong ties are instrumental for spreading behavior. Of course, with a sample size of 61 million, almost any non-zero estimate will be statistically significant. That's not a bad thing. Big sample sizes mean that our estimates are quite precise, so we will more reliably detect genuine effects. However, we can't just assume that any statistically

significant result is also substantively significant.

We've seen that statistically significant results can be substantively insignificant. The opposite is also true. In a 2011 study by Samuel Berlinski and Torun Dewan published in the *Quarterly Journal of Political Science*, the authors estimate the effects of the Second Reform Act of 1867 on elections in the United Kingdom. Despite the fact that the Second Reform Act roughly doubled the size of the eligible electorate and brought working class voters to the polls for the first time, the authors report that there was little effect on election outcomes. In their abstract, the authors report that "there is no evidence relating Liberal [one of the major British parties] electoral support to changes in the franchise rules."

Is this really the right way to interpret the evidence? When the authors say there is no evidence, what they mean is that their estimates of the effect of the Reform Act are not statistically significant. So they can't say, with 95% confidence, that those results didn't arise by chance. But, while not statistically significant, the evidence from the study actually suggests that the Second Reform Act had very important electoral consequences. The numerical estimates indicate that the Reform Act's doubling of the electorate increased the Liberal Party's vote share by 8 percentage points, a substantively large effect implying that the new, working-class voters enfranchised by the reform were much more likely to support the Liberal Party than were wealthier, previously enfranchised voters. However, although the estimate is substantively large, it is also imprecise and therefore not statistically significant. Fixating on this statistical insignificance, the authors conclude that the Second Reform Act must have had little effect. But the evidence actually indicates that our best guess is it had a big effect. It's just that that guess is uncertain.

Although statistical significance is useful and informative, it is often misused and misunderstood. A theme throughout this book is that clear thinking and quantitative evidence are complements, not substitutes. Just because we're doing statistics doesn't mean we can stop thinking hard about the questions we really want to answer. We should utilize sta-

tistical inference when possible. But we should also always remind ourselves not to forget about making substantive inferences from the evidence.

Key Terms

- **Population:** The units in the world we are trying to learn about.
- **Sample:** A subset of the population on which we have data.
- **Estimand:** The unobserved quantity we are trying to learn about with our data analysis.
- **Estimator:** The procedure we apply to data to generate a numerical result.
- **Estimate:** The numerical result arising from the application of our estimator to a specific set of data.
- **Bias:** Differences between your estimand and your estimate that arise for *systematic* reasons—i.e., for reasons that will persist on average over many different samples of data.
- **Noise:** Differences between your estimand and your estimate that arise due to *idiosyncratic* facts about your sample.
- **Unbiasedness:** An estimate/estimator is unbiased if by repeating our estimation procedure over and over again an infinite number of times, the average value of our estimates would equal the estimand.
- **Precision:** An estimate/estimator is precise if by repeating our estimation procedure over and over again, the various estimates would be close to each other. The more similar the hypothetical estimates from repeating the estimator, the more precise the estimate.

- **Efficiency:** An estimate/estimator is efficient if by repeating our estimation procedure over and over again, the various estimates would be close to each other and close to the estimand. The more similar the hypothetical estimates from repeating the estimator, the more precise the estimate.
- **Sampling Distribution:** The distribution of estimates that we would get if we repeated of our estimator an infinite number of times, each time with a new sample of data.
- **Standard Error:** The standard deviation of the sampling distribution. If the estimator is unbiased, the standard error gives us a sense of how far, on average, our estimate would be from the estimand if we repeated our procedure over and over with independent samples of data.
- **95% confidence interval:** If we applied the estimator an infinite number of times, each time on a new sample of data, the estimand would be contained in the 95% confidence interval (newly calculated each time) 95% of the time. Importantly, it is not true that we are 95% confident that the true estimand lies in the 95% confidence interval.
- **Hypothesis testing:** Statistical techniques for assessing how confident we should be that some feature or the data reflects a real feature of the world, rather than arising from noise.
- **Null hypothesis:** The hypothesis that some feature of the data is the result of noise.
- **Statistical significance:** We say that we have statistically significant evidence for some hypothesis when we can reject the null hypothesis at some pre-specified level of confidence (typically, 95% confidence).

- **p-value:** The probability of finding a relationship as strong or stronger than the relationship found in the data if the null hypothesis is true. We use p-values to assess statistical significance. For instance, if the p-value is less than .05, then we have statistically significant evidence (at the 95% confidence level) that the relationship is real. Importantly, the p-value does not tell the probability the null hypothesis is true.

Exercises

1. Historically, the average temperature in Chicago is about 50 degrees Fahrenheit.
 - (a) Simple Steve is a local meteorologists in Chicago. He decides that he will stop using his complicated meteorological models and, instead, simply predict that it will be 50 degrees every day. Is the problem with Steve's forecasts bias, imprecision, or both?
 - (b) Simple Steve is not the worst of the meteorologists in Chicago. Wacky Wanda decides to flip a fair coin every day, and if it comes up heads, she predicts a 150-degree heat wave, and if the coin comes up tails, she predicts a -50-degree frozen tundra. Is the problem with Wanda's forecasts bias, imprecision, or both?
 - (c) Genius Georgia decides to really mess with the citizens of Chicago. She has a meteorological model that, miraculously, correctly predicts the weather every single day. But Genius Georgia decides that she'll always report that the temperature is going to be 5 degrees colder than what her model says. Is the problem with Georgia's forecasts bias, imprecision, or both?
 - (d) Finally, Evil Elvis decides to combine Wanda and Georgia's approach. He takes Georgia's forecast, flips a coin, and either adds or subtracts 50 degrees, depending on whether he gets heads or tails. Is the problem with Elvis's forecasts bias, imprecision, or both?

2. Let's return to the analysis of schooling and earnings from last chapter's exercises. When you regress earnings on schooling, in addition to giving you estimated coefficients, your computer also probably gave you some other numbers that you didn't understand until you read this chapter. For the coefficient associated with years of schooling, you should have obtained an estimate of 1.16, indicating that each additional year of schooling corresponds with increased earnings of about 1,160 dollars. What are the estimated standard error, p-value, and 95% confidence interval associated with that coefficient? Provide a substantive interpretation of each one.

Notes

If you are interested in the history of how scientists and statisticians came to settle on the widely accepted 5% threshold for statistical significance, which is surprisingly interesting, see:

Michael Cowles and Caroline Davis. 1982. "On the Origins of the .05 Level of Statistical Significance." *American Psychologist* 37(5):553–558.

The paper on Facebook and turnout that we mention is:

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415): 295–298.

The paper on the British Second Reform Act is:

Samuel Berlinski and Torun Dewan. 2011. "The political consequences of franchise extension: Evidence from the second reform act." *Quarterly Journal of Political Science* 6(34):329–376.

Chapter 7

Over-Comparing, Under-Reporting

What you'll learn

- If analysts make lots of comparisons, but report only the statistically significant ones, there will be lots of false-positive results and over-estimates.
- These false-positives can be the result of nefarious researcher behavior (p-hacking). But they can also arise in a community of entirely honest researchers (p-screening).
- There's no easy solution, but analysts and consumers have some tools at their disposal to reduce the risk of being misled.

In 2008 and 2010, Paul the Octopus made headlines for his apparent prowess in predicting the outcome of soccer matches. Before matches between two national teams, Paul's keepers would present to him two boxes of food, each marked with the flag of one of the competitor countries. The keepers interpreted the box Paul ate from first as his forecast for who would win the match. Paul was surprisingly accurate, and journalists and gamblers eagerly awaited his predictions.

Paul was the subject of much fascination and some scorn. According to Nick Collins of *The Telegraph*, an Argentinian chef was so angry after Paul correctly forecast Germany's defeats of Argentina that "he threatened to cook Paul in retribution." Gamblers were betting on the accuracy of Paul's predictions before he had even made them. Collins reported that "William Hill, the bookmaker, says it has taken so many bets on whether Paul will call the final between Spain and Holland correctly that it had to cut odds from evens to 10/11."

A skeptic might point out that, although octopuses are impressively intelligent, there's no way that Paul could actually have had special insight into the outcome of soccer matches. Even experts have a hard time calling games in a sport where, as far as we can tell, basically no one ever scores. And Paul presumably knew nothing about the teams playing or even about soccer in general. (You'd think he was American, rather than German.) Was Paul's success dumb luck?

As discussed in the previous chapter, we have tools for assessing whether an observed pattern is likely to be the result of dumb luck or, more scientifically, noise. We can conduct a hypothesis test and calculate a p-value.

How does Paul fare in such a hypothesis test? Paul made 14 predictions over the course of his career, and he was correct in 12 of those 14 games. That's pretty good. Suppose the null hypothesis that this was dumb luck is true—i.e., Paul was picking in a completely random fashion, with each box equally likely to be selected. In other words, imagine that

Paul's predictions were the equivalent of flipping a coin. To figure out whether it is plausible that Paul's record emerged from dumb luck alone, we want to know how likely is it that Paul would have guessed correctly at least 12 times if he was in fact just flipping coins.

This problem is simple enough that you can compute the p-value by hand. The basic idea is this. Assume Paul is guessing at random. Calculate how likely it is that he'd get exactly 12 correct, how likely it is that he'd get exactly 13 correct, and how likely it is that he'd get exactly 14 correct. The sum of those three probabilities is the probability Paul would have done at least as well as he did by dumb luck.

Before we go on with the Paul story, let's pause and learn how to calculate these probabilities. Doing so will help make sure we are all thinking clearly about what dumb luck really means.

It will help to start with a simplification of the problem. Our null hypothesis is that Paul the Octopus is guessing at random. That is, that Paul predicting the winner of a game is analogous to a person flipping a coin and having it land on heads. So let's think about flipping a coin. Suppose you flip a coin 3 times (we'll get to Paul's 14 in a bit). Table 7.1 shows all the things that could happen.

	Three Heads	Two Heads	One Head	Zero Heads
Ways It Can Happen	HHH	HHT HTH THH	HTT THT TTH	TTT

Table 7.1: Ways to flip three coins

Equivalently, if Paul forecast three games, then zero, one, two, or three of his predictions could be correct.

What is the probability that you get, say, exactly two heads? Well, there are eight total things that could happen, and if we're flipping a fair coin, they're all equally likely. Of

those eight, three involve getting two heads. So the probability you get exactly two heads when you flip three coins is $\frac{3}{8}$. Analogously, if Paul was forecasting three games at random, the probability he would get exactly two correct is $\frac{3}{8}$.

But that isn't quite the quantity we want to know. We want to know the probability you get *at least* two heads or that Paul correctly forecasts at least two games?

Well, in addition to getting two heads, you could also get three heads. There is one way for this to happen, so the probability you get three heads is $\frac{1}{8}$. Hence, the probability you get at least two heads is $\frac{3}{8} + \frac{1}{8} = \frac{1}{2}$. Analogously, if Paul had called three games, and he was just guessing randomly, the probability he'd get at least two right is one-half.

But Paul didn't just forecast three games. He forecast 14. Making that table for fourteen coin flips would be pretty boring. So let's think about how to solve this problem a little more generally.

Suppose you flipped a coin n times. How likely is it that you get exactly k heads? Let's start by calculating the probability that each of the first k coins lands on Heads and the remainder land on tails. The probability that the first k land heads is $\frac{1}{2}^k$. The probability that the remainder land tails is $\frac{1}{2}^{n-k}$. So the probability that the first k flips land heads and the remaining $n - k$ flips land tails is $\frac{1}{2}^k \times \frac{1}{2}^{n-k}$.

Of course, that's only one way to get exactly k heads. The k heads don't have to be the first k flips. They can any group of k out of the n coins. There are $\frac{n!}{k!(n-k)!}$ different ways to get exactly k heads when flipping n coins. So the overall probability of getting exactly k heads out of n coin flips is

$$\frac{1}{2}^k \times \frac{1}{2}^{n-k} \times \frac{n!}{k!(n-k)!}.$$

The exclamation points above mean *factorial*. We refer to the expression $n!$ as "n factorial" and it's defined as the product of n and every positive number less than n . So for example, $3! = 3 \times 2 \times 1 = 6$.

Let's see that this confirms our finding from before in our three coin example. If we flip three coins, what is the probability we get exactly two heads?

$$\frac{1}{2}^2 \times \frac{1}{2}^{3-2} \times \frac{3!}{2!(3-2)!} = \frac{1}{4} \times \frac{1}{2} \times \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = \frac{3}{8}.$$

And now we can calculate the probability that Paul the Octopus would correctly predict twelve or more games out of fourteen, if he was picking at random. The probability he gets exactly twelve right is:

$$\frac{1}{2}^{12} \times \frac{1}{2}^{14-12} \times \frac{14!}{12!(14-12)!} \approx .00555$$

The probability he gets exactly thirteen right is:

$$\frac{1}{2}^{13} \times \frac{1}{2}^{14-13} \times \frac{14!}{13!(14-13)!} \approx .00085$$

The probability he gets all fourteen right is:

$$\frac{1}{2}^{14} \times \frac{1}{2}^{14-14} \times \frac{14!}{14!(14-14)!} \approx .00006$$

So the probability Paul calls twelve or more games correctly is approximately $.00555 + .00085 + .00006 \approx .0065$, around 1 in 155. In other words, if Paul had no special insight into soccer, it's highly unlikely that he would have been as accurate as he was. That's, of course, precisely why everyone was obsessed with Paul. And it seems like perhaps they were right to be. Using the standard statistical hypothesis testing approach that we introduced in Chapter 6, we can reject the null hypothesis that Paul is just guessing at random and conclude that we have statistically significant evidence that Paul is indeed an expert soccer forecaster.

The analysis above is pretty similar to what two mathematicians, Chris Budd and David Spiegelhalter, did when they were interviewed about Paul back in 2010. But if we look at Paul's games a little more closely, we can see that this analysis may be overly generous to Paul's putative psychic powers.

Paul lived in Germany, and he was primarily asked to predict the outcome of games in which Germany was competing. In fact, 13 of the 14 games involved Germany. Furthermore, Paul had a strong tendency to pick Germany. Maybe he liked that flag because he'd been seeing it for years. Maybe the German box happened to be his favorite box for reasons unbeknownst to us. Maybe Paul's handlers subconsciously trained Paul to pick Germany. Who knows? It also turns out that Germany is good at soccer—they win most of the time. So maybe Paul's success isn't so shocking given those two facts. Let's redo the analysis above with this information in mind.

Paul predicted the outcome in 13 games involving Germany and he picked Germany to win 11 of those games. Germany in fact won 9 of them. Our null hypothesis is again that Paul's predictions were dumb luck, in the sense of having no special insight into soccer. But this time, instead of imagining that he was equally likely to choose either box (i.e., that he flipped a fair coin), imagine he was predisposed to pick the German box (i.e., he flipped a weighted coin). Let's assume his predisposition meant that his probability of picking Germany was $\frac{11}{13}$ in any game where Germany played, since that's the frequency with which Paul in fact selected Germany. If Germany won 9 of 13 games, and Paul selected Germany with a probability of $\frac{11}{13}$ each time, how likely is it that he would have been correct 11 or more times just by pure chance? This p-value could also be computed by hand, but it is complicated to do so. So instead we ran a simple simulation on our computer to approximate it. With these tweaked assumptions, the chances that Paul gets 11 or more games right out of 13 is about .03 or 1 in 33—still unlikely but much more likely than 1 in 155.

Now, what do we think? It still looks pretty unlikely that Paul's success is attributable

purely to dumb luck. Even if he was predisposed to predict the strong German team, there was only a 3% chance that Paul would be as successful as he was. Therefore, a traditional hypothesis test with a .05 threshold would still lead us to reject the null. We continue to have statistically significant evidence that Paul is good at predicting soccer matches.

You won't be surprised to learn that we're still skeptical. But why? For one, Paul is not the only octopus out there. What if there were actually 10 octopuses scattered around Germany, each trying to predict the outcomes of soccer matches? The world, of course, would only ever hear about the most successful one. If this is right, then we still haven't tested the right hypothesis to figure out how likely it is that Paul's accuracy was due to dumb luck. If there really were 10 octopuses trying to predict soccer matches, and Paul just happened to be the one who did well and therefore became famous, instead of asking how likely it is that Paul would be so accurate by luck, we have to ask how likely it is that one of the 10 octopuses would be that accurate by chance. Because if it had turned out that Paulina the Octopus was right 12 out of 14 times instead of Paul, then we'd be talking about Paulina and we'd never have heard of Paul.

Figuring out how likely it is that some octopus out of the 10 would have been as accurate as Paul is relatively straightforward. But to do the calculation, we need to understand one more fact about p-values. As we've already learned, the p-value is the probability of observing an outcome at least as extreme as the one you observed if the null hypothesis is true. So, if the null hypothesis is true, how often will you observe an outcome as extreme as an outcome with a p-value of .05? Exactly 5% of the time. And if the null hypothesis is true, how often will you observe an outcome as extreme as an outcome with a p-value of .2? Exactly 20% of the time. And so on for each and every p-value. This is just a restatement of the definition of the p-value.

But from this obvious fact, we learn something important. When the null hypothesis is true, we should observe p-values of .05 or lower 5% of the time, p-values of .2 or lower 20%

of the time, p-values of .5 or lower 50% of the time, and so on. Hence, it must be that, when the null hypothesis is true, you are equally likely to find each p-value. (The technical jargon for this is that p-values are “uniformly distributed under the null”.)

So, what’s the probability that at least one of our German octopuses would generate a record of prediction with a p-value at least as good as Paul’s record of .03 by dumb luck alone? We just saw that the probability that any one octopus generates a p-value of .03 or lower by dumb luck alone is .03. Therefore, the probability that any one octopus generates a p-value higher than .03 is .97. If there are two octopuses and they’re making their guesses independently, the probability that neither of them generates a p-value better than .03 is therefore $.97^2$. So the probability that at least one of them generates a p-value of .03 or better is $1 - .97^2$ (i.e. one minus the probability that both generate a p-value worse than .03). If 10 octopuses are taking random guesses, the probability that at least one generates a p-value as good as Paul’s is $1 - .97^{10} \approx .26$. In other words, if 10 German octopuses went through the same ridiculous prediction exercise as Paul, there’s a 1 in 4 chance that at least one of them would have accumulated a record of predictive accuracy at least as glorious as Paul’s, even if none of the octopuses were in fact soccer experts. This should make us much more skeptical of Paul’s abilities.

We don’t know how many German octopuses were in the soccer forecasting business. But we do know that lots of other animals got in on the action. No joke, Leon the Porcupine, Petty the Hippopotamus, Anton the Tamarin, and Mani the Parakeet all predicted soccer matches around the same time as Paul. And those are just the ones that made the news. Presumably, there were dozens more that we never heard about. And this discussion only includes soccer. What about all the other sports? And what about all the non-athletic things there are to predict. If Judy the Badger were good at predicting college football games, Steve the Cat were good at predicting congressional elections, and Fran the Otter were good at predicting stock market shifts, they would be celebrities too. But they turned

out to be no better than chance, so we never heard about them.

Budd and Spiegelhalter, the mathematicians, were quick to point this out themselves. Spiegelhalter notes that “if someone flips a coin and gets the same result 9 or 10 times, it is not remarkable in itself, but it will seem remarkable to the person flipping the coin.” In other words, if enough people flip coins, one of them is bound to flip a bunch of heads in a row. And despite the fact that somebody was bound to have a lucky streak of heads by chance, that person might wrongly conclude that they have an unfair coin or that they are a particularly skilled coin flipper. Unfortunately, as we’ll see, this problem applies to much more serious situations than coin flipping and soccer forecasting, with far reaching implications.

The Problem of Over-Comparing and Under-Reporting

Statistical testing and p-values are clearly useful. When we find patterns in the world, we want to know if they reflect genuine phenomena or if they could have easily been produced by random chance.

But there is a problem which the story of Paul the Octopus highlights. Neither the public nor the broader scientific community gets to see all the statistical tests that were (or could have been) conducted. Often, the only results reported and published are the statistically significant ones. It’s just not that interesting to write about Mary the Octopus who is about as good as a coin flip at predicting soccer matches. But, as we’ve seen, if there are 20 different animals out there predicting things, we’d expect 1 of them to have a p-value less than .05 by pure luck, even if none of them actually have any special insight into soccer. Only that 1 in 20 case will be written up or reported in the news. So if we base our beliefs purely off of what gets reported, we will have systematically misguided beliefs.

Making a lot of comparisons while selectively reporting only the interesting or statisti-

cally significant ones is a dangerous combination. But it is wide spread. And, because of it, when we hear about a new, exciting scientific result, it is much harder to know how likely it is to reflect a genuine phenomenon than the simple logic of hypothesis testing against the null seems to suggest.

This problem of over-comparing and under-reporting doesn't just affect how confident you should be that a phenomenon is genuine. It also creates bias in the published record of estimates. To see why, let's go back to our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

Suppose, for the sake of argument, that all the studies we are talking about are really well designed. There is no bias. The only reason estimates differ from the true quantity of interest in the world (the estimand) is because of noise.

But let's also suppose, in the spirit of under-reporting (i.e., not reporting every result) that we only hear about the results of studies where the evidence is strong enough to reject the null hypothesis that the true estimand is zero (i.e., that our estimate was the result of noise).

Which kinds of estimates are statistically distinguishable from zero? If the null is true, so that there is no genuine relationship in the world, we are much more likely to get estimates that are small in magnitude (i.e., estimates close to zero) than estimate that are large in magnitude. Thus, the only estimates that will be statistically significant are those that are large enough in magnitude.

This means that, for any given true estimand, the estimates that end up being reported will be those where the noise happened to be large enough in magnitude to push the magnitude of the estimate far enough away from zero to make it statistically significant. So if lots of comparisons get made (over-comparing) and only the statistically significant ones

get reported (under-reporting), not only will our p-values be wrong, but the collection of estimates that we learn about from published studies will systematically over-estimate the magnitude of the true estimand.

Distressingly, even though we started by assuming no bias in our estimate, we've learned that the process of over-comparing and under-reporting introduces bias, not into any one estimate, but into the overall distribution of estimates reported in a scientific literature. This troubling phenomenon is sometimes referred to as *publication bias*.

Figure 7.1 illustrates how this works. In the top figure, we see 50 unbiased but imprecise estimates. The true estimand is 1, and since our estimates are unbiased, the average of all estimates is also 1.

We calculate the standard error and find that the 95% confidence interval is from -2 to 2. That is, the probability an estimate would have arisen even if there was no relationship (i.e., the true estimand was 0) is less than 5% only if that estimate is greater than 2 in magnitude. So only the estimates larger than 2 are deemed statistically significant (we don't have any estimates less than -2). The statistically significant estimates are in the lower figure.

Suppose that only these statistically significant estimates are ever reported. Now, of course, the reported estimates are systematically greater than the estimand. So if we based our beliefs about the true value of the estimand on these published estimates, we'd be systematically over-estimating the truth. This is publication bias.

The problems associated with over-comparing and under-reporting can arise in a variety of ways. Let's consider a couple.

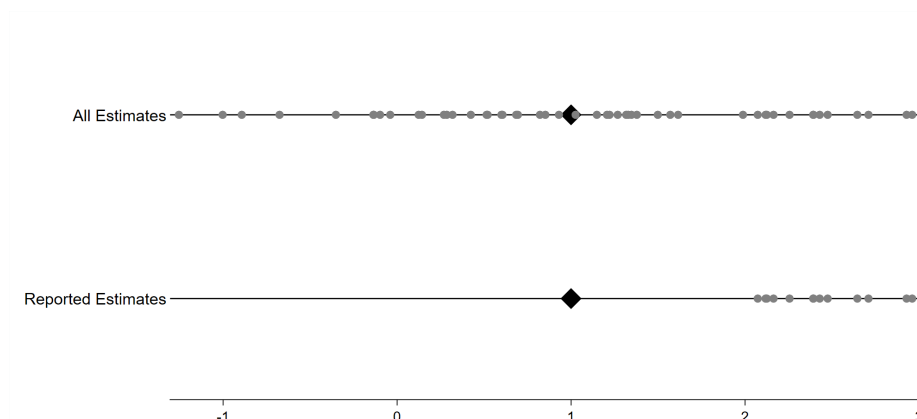


Figure 7.1: Only reporting statistically significant estimates creates bias.

p-hacking

One way we might end with over-comparing and under-reporting is through bad or dishonest behavior by individual analysts. For instance, suppose a scientist does an experiment and doesn't quite get statistically significant evidence for the expected or desired result. That scientist might reason that something probably wasn't quite right in the initial attempt, and so try a slight tweak on the experiment. Indeed, the scientist can keep trying similar experiments until one comes up statistically significant. Because of noise, if we try the experiment enough times, eventually we will get a result, even if there is no real phenomenon being studied. That's the problem of over-comparing. And, of course, if an unscrupulous scientist only reports the results from the one experiment that yielded a statistically significant result, we have the problem of under-reporting and, thus, publication bias.

Or maybe the analyst has some flexibility in how a particular statistical test is implemented. Suppose you are asked to do an analysis at work about the relationship between productivity and having a standing desk. Should you pool the entire work force together or run separate regressions for different age groups? Should you include higher-order polyno-

mials of age? Should you separate women and men in the analysis? How about people with different job responsibilities, medical histories, and so on? As you can see, there are lots of reasonable ways to think about doing the analysis. If you keep trying out different ones, you'll eventually find a statistically significant result, even if there is in fact no real relationship between productivity and standing. Searching over specifications is, thus, another way of over comparing.

Yet another way is by trying out lots of different outcomes. Suppose you want to evaluate the efficacy of some new pill for heart disease. You might run an excellent clinical trial that generates no bias at all. But perhaps you collected data on lots of outcomes for the experimental subjects: mortality, heart attacks, strokes, cholesterol, days of hospitalization, ability to exercise, and so on. You can then test whether the pill has a statistically significant effect on each of these outcomes. If you have enough different outcomes, you're likely to find a statistically significant result on one of them just due to noise, even if the medicine doesn't actually do anything. And, if you lack the proper ethics, you might just report the results for that one outcome in the hope of convincing doctors to prescribe your new pill.

The scientific community now calls this kind of behavior, where one plays around with the data or tests until a p-value below a particular threshold emerges, *p-hacking*. As we've seen, it can take many different forms, and you have to work hard to avoid it as a quantitative analyst and to detect it as a consumer of quantitative evidence.

Fun fact. The term p-hacking was coined by Joseph Simmons, Leif Nelson, and Uri Simonsohn in a clever study where they showed, among other things, that by using standard methods in social science, they could provide statistically significant evidence that listening to the song "When I'm Sixty-Four" by the Beatles makes subjects older!

p-screening

Of course, p-hacking is a big concern. But it need not be the case that any individual is acting in a dishonest or negligent way for the problem of over-comparing and under-reporting to occur. It can happen even if absolutely everyone behaves in a completely honest and responsible manner!

Imagine that 20 scientists around the country all have the same scientific hunch. Let's suppose it's about the efficacy of a potential new cancer drug. In truth, the hunch is false—the drug doesn't work. But there is no way for the scientists to know this at the outset. So, as scientists do, they design studies to test the drug. Indeed, all 20 of their labs, independently and unaware of the others, run the same, high-quality experiment, but on different samples. They each recruit a large sample of patients with the relevant type of cancer. At random, they assign half of them to receive the new drug. The other half receive a placebo sugar pill. At the end of the study period, they assess whether the group that received the drug was more likely to go into remission than the group that received the placebo.

Nineteen out of the twenty labs find no statistically significant evidence—the remission rates among those who received the drug and those who received the placebo are indistinguishable—and conclude that the drug doesn't work. Such null results aren't considered very exciting; “another drug doesn't cure cancer,” isn't a great headline. So scientific journals are reluctant to accept papers reporting null results for publication. As a consequence, these labs may not bother to write up their results, instead just moving on to more promising lines of research. This is sometimes called *the file drawer problem* because statistically insignificant results get locked away in a file drawer. Even if they do write up their findings, they might have trouble finding a journal interested in publishing them. In either case, the scientific community and the public fail to learn about these 19 null results.

One (un)lucky lab out of the twenty finds statistically significant evidence suggesting the drug works. We know the drug doesn't work (though the scientists don't), so we know this is sheer chance. It just so happens that, for reasons having nothing to do with the drug, the people assigned to receive the drug in this experiment also had higher remission rates than the people assigned to receive the placebo. These things happen. The estimate can differ from the estimand, even absent bias, because of noise.

Since the other studies were never written up, as far as the scientist in charge of this one lab is aware, all existing evidence points toward this new drug working. So this one lab, appropriately, writes a scientific paper on their findings. Because the result is surprising and noteworthy, it is likely to be well published and reported on by the scientific press. And, indeed, if you look at this one study, it looks great. The lab ran a good, unbiased experiment. They made only the one appropriate comparison about the one appropriate outcome. There was no p-hacking. And the data support their hypothesis. So everyone believes this result even though it is in fact completely wrong and, if we had access to all the data (i.e., from the 19 "failed" experiments), we'd see that the preponderance of the evidence points in exactly the opposite direction.

We call this problem *p-screening*. (No one else calls it that, we've just made it up. We hope you'll help us get it into common usage.) The issue here isn't that some individual researcher is p-hacking their way to a statistically significant result. The issue is that the scientific community, through its publication standards, screens out studies with p-values that are above some threshold. Under p-hacking we don't see null results because dishonest researchers hide them. Under p-screening we don't see null results because honest researchers can't publish such results. But either way the outcome is the same. The results we do see suffer from publication bias due to lots of comparisons being made, but only the statistically significant ones being reported.

Yikes! Because of p-screening, publication bias due to over-comparing and under-

reporting can happen in a scientific (or other) community, even when everyone behaves just as they should. This should make you worry that it is going on all the time. In fact, stop and ask yourself, for how many of the things you believe might this story characterize the state of knowledge. Once you start thinking clearly about the problem, you'll start to see its potential everywhere.

Are Most Scientific “Facts” False?

Over-comparing and under-reporting gives rise to publication bias. And these practices are pretty deeply baked into a lot of scientific practice and culture. This realization has led to something of an existential crisis in many scientific fields, with practitioners wondering whether many widely accepted scientific facts are actually false, the artifacts of over-comparing and under-reporting.

This is a real concern. It is surely the case that many things that we believe are true are in fact false because of publication bias. But certainly not everything. And analysts have started to think more clearly about how we might diagnose when a scientific consensus or literature is or is not likely to suffer from severe publication bias. To see how, we are going to talk through a couple examples of the problem and various attempts to address it.

ESP

In 2010, Cornell psychologist Daryl Bem made huge news by publishing a study in the *Journal of Personality and Social Psychology*, a prestigious academic psychology journals, claiming that human beings have extrasensory perception (ESP). Often, academic researchers and quantitative analysts are the ones debunking claims about the paranormal, but in this case, a respected, tenured Ivy League professor was the source of the outlandish claim.

In Bem's experiment, students were asked to predict which virtual curtain on their

computer screen (left or right) had an object of interest hiding behind it. Bem claims to have found statistically significant evidence that his subjects were better than chance at predicting the future and identifying the correct curtain.

This is a very exciting finding if you are journal editor who cares about notoriety or a science journalist who cares about readership. The result is cool. The scientist in question works at a reputable university. The article is published in a major scholarly journal. There is no reason to think the data are faked. The study provides scientific evidence of a, to say the least, surprising phenomenon. What journalist with blood pumping through their veins could resist the story.

This study and all the media attention it received notwithstanding, we're fairly confident that people don't have ESP. So what is going on?

There are, of course, the normal concerns with statistical hypothesis testing. If an analyst uses a significance threshold of .05, there's a 5 percent chance of finding support for a result (i.e., rejecting the null even), even if the result is false (i.e., the null is true). And as we'll see in Part IV, if you already have good reasons to believe that people do not have ESP, then you shouldn't shift those beliefs much in response to this one study.

But we have other concerns based on the themes of this chapter. This is a case where many researchers might be conducting experiments, but only the one with statistically significant evidence of an unlikely phenomenon gets published. Presumably, nobody is going to publish a paper that reports that people are no better than chance at guessing the correct curtain. That's what we all already believe. So we should worry a lot about publication bias due to p-screening.

We should probably also worry that Bem himself engaged in too many comparisons and too little reporting. Bem reported the results of 9 different experiments, carried out over the course of 10 years. These experiments are all inexpensive and easy to conduct. Since Bem

was, by all accounts, committed to the study of ESP, it doesn't seem far-fetched to imagine that he might have conducted many more ESP experiments over this 10 year period. And, if so, the 9 experiments which were reported on might well have been the ones with the strongest confirmatory evidence of ESP.

There are also some signs of over-comparing and under-reporting within the study itself. For example, Bem doesn't find evidence of ESP in general; he only finds it when the object behind the curtain is erotic in nature. For other kinds of objects, he finds no evidence of the paranormal. Of course! Wouldn't it make sense for us to have evolved ESP that allows humans to detect erotic activity around the corner but not other kinds of activity? Furthermore, in some tests, he only finds effects for women not men, or only for those who are easily bored. Given all of the different tests conducted by Bem, it would be surprising if he *hadn't* stumbled upon a few statistically significant results, just by chance.

Reassuringly, the community of psychologists remained skeptical and responded quite quickly to Bem's study. Several follow-up studies tried and failed to replicate the findings. Disappointingly, however, the *Journal of Personality and Social Psychology* initially refused to publish the replication studies debunking Bem's claim. The editor justified this decision on the grounds that the journal has a long standing policy of refusing to publish *mere replication*, an activity that is apparently insufficiently important or creative to grace the pages of his journal. Fortunately, the journal eventually had a change of heart and published a meta-analysis of replication attempts, which strongly suggested that the original result was unreliable. This illustrates one important corrective to the problem of over-comparing and under-reporting: vigilance and a commitment to investigating whether findings replicate within a scientific community.

Get out the vote

Political campaigns engage in lots of activities—phone calls, direct mail, door-to-door canvassing—to try to get out the vote. Since the 1990s, scholars have teamed up with campaigns to conduct experiments to learn about the efficacy of these efforts. In such studies, some people are randomly assigned to treatment (e.g., getting a direct mailing with information about the date of the election or the location of the polling place) and other people are randomly assigned to control (e.g., no extra information). We can learn about the average effect of the get-out-the-vote efforts on voter turnout by comparing the turnout rates in the two groups.

In the published record, the average effect of a get-out-the-vote intervention appears to be about a 3.5 percentage point increase in voter turnout. Moreover, almost no published paper reports an effect of less than 1 percentage point. So, if a campaign consulted the published record, it would conclude that get-out-the-vote efforts are quite effective.

But there have been many more get-out-the-vote experiments than there are scientific papers published on the topic. Which means that some of those experiments did not result in publications. Why not?

If our fears about over-comparing and under-reporting are right, we might expect that the answer is p-screening—experiments that yielded no statistically significant evidence of an effect did not result in a paper. If this is true, then there is publication bias. So we should expect that the true average effect of get-out-the-vote interventions is much smaller than one would conclude just from reading the published findings.

Three political scientists, Don Green, Mary McGrath, and Peter Aronow decided to investigate this possibility quantitatively. They managed to get their hands on the data from over 200 experiments, done by a variety of scholars over many years. Some of those experiments had resulted in published papers. Others had not. They did an analysis to

find the average effect of get-out-the-vote interventions across all 200 of these interventions. The answer: half a percentage point! Dramatically less than the 3.5 percentage point average effect shown in the published record. The unpublished record is, indeed, much less supportive of the efficacy of get-out-the-vote than is the published record.

The efficacy of get-out-the vote is one of the most rigorously studied topics in the social sciences. And so, candidates or campaigns wanting to figure out the best way to allocate scarce resources, might naturally turn to published studies to inform their decision. However, we now know that doing this would lead them to overestimate how effective get-out-the-vote is by a factor of 7. p-screening can have serious consequences.

p-hacking forensics

It's always hard to know for sure if p-hacking took place in an individual study. And it is a good practice to be charitable, assuming that most people are attempting to behave in an above-board and honest manner most of the time. That said, clear thinking can help us get a sense of how widespread the p-hacking problem is. The clearest evidence comes from observing the p-values in actual published scientific literatures. Doing so doesn't tell us whether any individual study is p-hacked. But it can help us sniff out whether an overall literature looks like it has a bunch of p-hacking going on.

Here's how it works. You start by thinking about what the distribution of p-values in a literature would look like in several different possible states of the world:

1. If there is no real relationship in the world and there is no p-hacking;
2. If there is a real relationship in the world and there is no p-hacking;
3. If there is no real relationship in the world and there is p-hacking;
4. If there is a real relationship in the world and there is p-hacking.

You are then going to compare the distribution of reported p-values in a literature to these to try to figure out which state of the world you are most likely to be in. For what follows we are going to assume there is still p-screening going on (so there are no p-values greater than .05). We just want to figure out if the literature is also p-hacked. But everything we are going to say is true even without p-screening.

Start with case 1: there is no real relationship in the world and there is no p-hacking. If there is no real relationship out there in the world, that means the null hypothesis is true. And, as we've already discussed, if the null hypothesis is true, then all p-values are equally likely to emerge in any given study. So, if there is no p-hacking or p-screening, the observed distribution of p-values in published studies should look approximately uniform—that is, between 0 and .05 different p-values should appear with approximately equal frequency.

There are two reasons we might see a deviation from this uniformity. The first is that there is a real relationship in the world. The second is that there is p-hacking.

This brings us to case 2: there is a real relationship in the world (i.e., the null hypothesis is false) and there is no p-hacking. If we are actually studying a real relationship in the world, we should be more likely to detect a statistically significant relationship than if there is no real relationship in the world. So, in case 2 where there is a real relationship and no p-hacking, we expect a distribution of p-values in the published record that is skewed toward low p-values relative to the uniform distribution we expected in case 1. That is, reflecting the fact that we are detecting a real relationship, there should be more low p-values in case 2 than in case 1. So, if we see a distribution skewed toward low p-values, that is suggestive evidence that the literature is detecting a real effect in the world.

The other reason we might see a deviation from case 1 is because of p-hacking. This takes us to case 3: there is no real relationship in the world, but there is p-hacking. When there is no real relationship in the world, really small p-values are especially unlikely. Larger p-values are more likely. So, if we p-hack, most of the studies we report will have relatively

large p-values (i.e., very close to .05). That is, unlike in case 2, where the distribution was skewed toward low p-values, in case 3, we should expect the distribution of p-values to be skewed toward high p-values, reflecting the fact that we are manipulating things until we “detect” a relationship that isn’t real.

Of course, case 4 combines cases 2 and 3. If there is a true relationship, that skews things toward low p-values. If we also p-hack, that skews things back toward high p-values. So it is hard to know what to expect in this case. But, nonetheless, just with the distinctions between cases 1, 2, and 3, we can make some progress diagnosing p-hacking in a literature.

The simulations shown in Figure 7.2, conducted in a 2014 study by Simonsohn, Nelson, and Simmons who first proposed examining the distribution of p-values in order to assess p-hacking, demonstrate the logic we just explained. The top half of the figure shows the distribution of p-values that emerge if there is no p-hacking. The bottom panel shows the distribution of p-values that emerge when there is p-hacking. As we move from left to right, the actual size of the relationship being studied increases, from no real relationship to a quite strong relationship.

So, start with the top row, where there is no p-hacking. In the first column, we are in our case 1, no real relationship and no p-hacking. And we see that all p-values are equally likely. As we move to the right, the relationship becomes real, so we are in case 3, a real relationship and no p-hacking. As expected, the distribution of p-values becomes skewed toward lower p-values, which are more likely because the effect is real.

If we go to the bottom row, in the first column we are in case 2, no real relationship and p-hacking. As expected, now the distribution is skewed toward higher p-values, which are more likely because low p-values are hard to find when there is no real relationship. As we move to the right, we are in case 4, real relationship and p-hacking. As expected, now the distribution of p-values can go either way, depending on whether the impact of the real effect or the impact of p-hacking is bigger. In the second column, the distribution is still

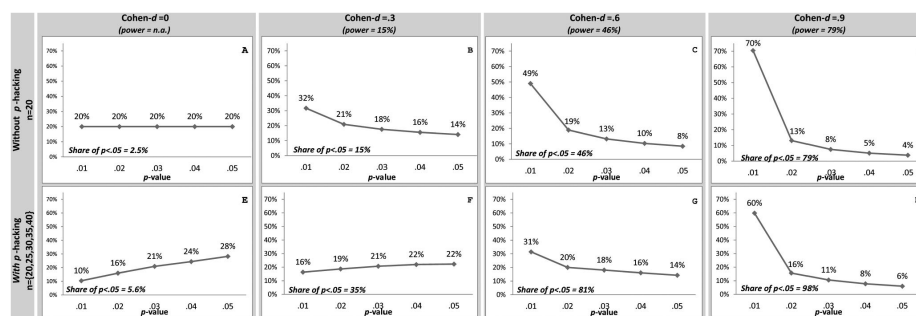


Figure 7.2: p-hacking distorts the distribution of p-values in a literature.

skewed toward high p-values. But by the fourth column, where the relationship is strong, we are back to a distribution skewed toward low p-values.

Sadly, many academic literatures exhibit the worst case scenario shown in the bottom left corner of Figure 7.2. Simonsohn, Nelson, and Simmons examine papers in a prominent psychology journal. They then subdivided the papers according to whether they did or did not have any words that might be a sign of over-comparing. For instance, one of their words of concern is “excluded”, as in “I excluded this variable (or group, or outcome) from my analysis because it didn’t give the result I wanted”. Another is “transformed,” as in “I transformed age into age², age³, age⁴, and so on until the results supported my hypothesis”. In the right panel of Figure 7.3, the blue line shows the distribution of p-values for studies without those words. Reassuringly, we see a distribution skewed toward low p-values, indicating genuine effects. In the left panel, the blue line shows the distribution of p-values for studies with those words. Disturbingly, we see a distribution skewed toward high p-values, indicating lots of p-hacking. So, while these forensics don’t tell us exactly which papers are or are not p-hacked, they allow us to look at the distribution of p-values in a literature and ask how worried we should be that any scientific consensus based on those studies is biased by p-hacking.

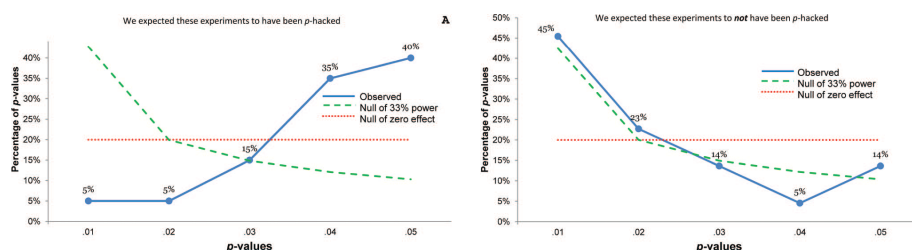


Figure 7.3: Using the distributino of p-values to diagnose p-hacking.

Potential Solutions

The problem of over-comparing and under-reporting is pervasive. And it isn't confined to scientific studies. Your reasons for trusting your investment advisor, doctor, tarot card reader, or what have you are all subject to concerns about over-comparing and under-reporting. But it is a particularly insidious problem for science. And so scientists have started thinking about how they might change scientific practice to reduce the problem. While most of these don't apply outside of the scientific context, it is still interesting and worthwhile to give them some thought.

Change the significance threshold

Maybe we can solve the problem by using a more stringent significance thresholds for p-values. Maybe the conventional .05 threshold makes it too easy to hunt around until you find a statistically significant result. In 2017, 72 researchers across various fields published a letter in *Nature Human Behavior* urging the scientific community to adopt a dramatically lower significance threshold of $p < .005$.

On the one hand, a lower significance threshold would certainly make it harder to conjure up a statistically significant result by over-comparing. On the other hand, lowering the significance threshold would likely increase incentives for p-hacking by making statistically

significant results rarer and, thus, more valuable. It might even increase complacency about these issues, leading us all to think a little less critically. And of course, while a threshold of .005 means fewer false positives (i.e., rejecting the null hypothesis when it is true), that comes at the cost of more false negatives (i.e., failing to reject the null hypothesis when it is false). It's not obvious where we should draw the line to balance that trade-off. The answer probably depends on the particular question.

Adjust p-values for multiple testing

The p-value is supposed to tell us the probability of obtaining a result like the one you found if the null hypothesis is true. As we've seen, if researchers engage in over-comparing and under-reporting, the p-value doesn't reflect this probability. It is too low.

If we know how many tests were run, we can try to correct the p-value. As we discussed in the case of Paul the Octopus, if researchers conduct 10 independent tests but only report their lowest p-value of .03, the true p-value is more like $1 - (1 - .03)^{10} \approx .263$. Correcting p-values to account for the number of tests run is a good way for researchers to be transparent and for consumers of quantitative information to better assess the state of the evidence. Unfortunately, it also isn't a panacea. The kind of simple calculation done above only works if the tests are truly independent. If the tests are closely related to each other—for example, if we're testing the same hypothesis with the same data but using slightly different variables in a regression—if may be much less clear how to adjust the p-values correctly.

Pre-registration

Researchers can pre-commit to exactly the tests they are going to do before they ever see the data. To do so, they pre-register their study, writing down and publishing exactly what they plan to test for and how before actually doing the study. As long as they pre-register

a reasonable number of tests, this prevents them from over-comparing. It also makes it harder to under-report. People will be suspicious if a scientific paper only includes the results from 3 out of 10 promised tests. Moreover, some scientific journals are even now willing to accept scientific studies for publication based only on the pre-registered plan—committing to report the results regardless of what the researcher finds, which also helps with under-reporting.

We definitely think pre-registration is a useful tool for mitigating the problem of over-comparing and under-reporting. But let's think a little more about the merits and limitations of this approach.

In a typical drug trial, one randomly assigned subset of individuals receives the drug while the others are randomly assigned into the control group, receiving a placebo. After some period of time, the researchers compare to see whether the treatment and control groups differ from one another on some relevant health outcome. Often, researchers test for an effect of the treatment on a variety of different outcomes—e.g., blood pressure, weight, arthritis, stroke, etc. If a drug appears to reduce weight but has no effect on other outcomes, the trial is declared a success, and the new product will be marketed for weight loss.

As we've already discussed, testing the impact of the drug on lots of outcomes is one way to run into the problem of over-comparing and under-reporting. How do we know if the drug really reduces weight or if weight happened to be the 1 outcome out of 20 for which the scientists found a statistically significant relationship by pure chance (i.e., a false positive). Surely, if the researchers could test enough different outcomes, they'd find at least one of them with a statistically significant difference between the treatment and control group, just because of sampling variation.

The National Heart, Lung, and Blood Institute (NHLBI) has funded many clinical trials of new drugs and dietary supplements since 1970, and in 2000, they came up with a clever way to use pre-registration to combat this problem. They required the developers of a drug

or supplement to announce beforehand the goals of the product. Under these new rules, a clinical drug trial is only declared a success if the researchers show a statistically significant effect on the pre-registered outcome of interest.

A 2015 study by Robert Kaplan and Veronica Irvin shows that after the NHLBI started requiring pre-registration, the rate of successful trials dropped from 57 percent to 8 percent. This suggests that many of the “successful” trials prior to pre-registration were the result of over-comparing rather than any genuine effect of the drug or supplement. This was a big success for pre-registration.

Now, think about the 8 percent success rate after pre-registration. Kaplan and Irvin use a significance threshold of .05, which means that even if none of the drugs or supplements were effective, we’d expect 5 percent to generate statistically significant evidence of efficacy just by chance. So the 8 percent success rate is not much higher than what we’d expect even if none of the drugs work. That means that even after we observe a successful, pre-registered trial, we still should not be that confident that the drug is effective. In fact, our best guess is that there’s still a 5 in 8 chance that a positive, pre-registered result is false positive.

Replication

One way to know if an estimated effect is genuine is to replicate it on new, independently generated data. Replication isn’t foolproof. But, as we saw with the ESP studies, it can help.

Suppose we do just one comparison and use a significance threshold of .05. The likelihood of finding statistically significant evidence of a relationship, even if none exists, is .05. But if we run the study twice, using independent data each time, the probability of finding statistically significant evidence in *both* studies if no real relationship exists is

$.05 \times .05 = .0025$. If we do it a third time, the probability we find evidence for the non-existent relationship in all three studies is $.05^3 = .000125$, pretty unlikely. By replicating, we reduce the chance that we reach a spurious positive conclusion.

Of course, there is the opposite risk. Failure to reject the null hypothesis is not proof that the null is true. So, if we only believe results if they replicate multiple times, we might sometimes wrongly reject real effects, especially if we conduct replications on sparse or noisy data where effects are hard to detect. Ideally, we'd get lots of data and replicate on large samples.

Sometimes such replication is feasible and sometimes it's not. If a drug trial shows an effect on weight that wasn't expected, the researchers can recruit a new pool of subjects and see if the new treatment group shows similar weight loss relative to the new control group. But if we discover a phenomenon regarding 20th-century gubernatorial elections, the behavior of the moons of Venus, or the behavior of leaders in the run-up to World War I, there's just one sample of them. There's no way to go collect more data. In that case, we may want to think less literally and more conceptually about replication. We do this not by directly replicating the existing findings, but rather by asking something like, "if this phenomenon is genuine, what are some other hypotheses that should also be true?"

Anthony has a paper with Pablo Montagnes that illustrates this approach. A prominent study by Andrew Healy, Neil Malhotra and Cecilia Mo, published in the *Proceedings of the National Academy of Sciences*, claims that the outcome of college football games affects who wins elections. Specifically, the incumbent party reportedly performs better in congressional and gubernatorial elections in the home counties of teams that won prior to the election. This kind of finding makes some people worry about democracy. (Not us, but that is a topic for another day.)

Anyway, the study is in many ways good. Football wins and losses seem pretty random, so there isn't lots of reason to be concerned about bias. But this is exactly the kind of

setting where you might worry about a false positive resulting from over-comparing and under-reporting. For instance, there is almost certainly a p-screening problem—would a prestigious scientific journal have published a paper showing that college football games do *not* appear to have any influence on elections? Moreover, there are lots of kinds of sports that might have been used to predict incumbent success. Where is the paper on basketball or curling losses and elections? So we shouldn't leap to conclusions just because one published paper presents evidence supporting the claim that losses in one particular sport are associated with election outcomes.

Anthony and Pablo couldn't conduct a purely independent replication because there's no way to re-run decades of football games and elections. Instead, they tested additional hypotheses that seem like they should also hold if football losses really do affect elections. For instance, if football losses really affect elections, you might expect the relationship to be particularly strong in places that care a lot about college football. If voters really blame incumbent politicians for bad football outcomes, you might expect the impact of a football loss on the incumbent's party to be bigger when the incumbent is actually seeking reelection relative to when some new candidate from the same party is running. And so on.

Here are some examples of what Anthony and Pablo found. It turns out that the estimated effect of football games is *smaller* in counties where more people follow college football than in counties where fewer people follow college football, no greater when an incumbent actually runs for reelection as opposed to just the incumbent's party running a candidate, and just as strong outside the home county of the team as inside the home county of the team. Furthermore, the relationship between football losses and electoral outcomes does not appear to exist for NFL games, despite the fact that NFL teams have the same kind of regional support as college football teams and NFL games are roughly 10 times more popular than college football games.

After testing multiple, independent hypotheses that seem like they should hold if the

relationship between football and elections is real, none of which receive support in the data, it seems unlikely that college football games really do influence elections. This isn't a classic replication. But it shows how looking at the evidence for additional hypotheses that are related to the original hypothesis can help shed light on the strength of the evidence for surprising results.

Related to the idea of independent replication is the use of hold-out samples, which we discussed in Chapter 5 when we talked about over-fitting. Suppose you have a large sample of data, and want to explore it for relationships. It might be a smart idea to hold out a randomly selected subset of that data from the exploration. For example, you could randomly select half the observations and use them as an exploratory data set. Then, after you've found a few interesting relationships, you could test whether those relationships also appear in the hold-out sample of data that you have not yet analyzed. If over-comparing produced a false positive in your initial analysis, we would expect that the same relationship is unlikely to appear in the hold-out sample. But if you've identified a real phenomenon, then we should expect it to persist in the hold-out data.

Test important, rather than cute, hypotheses

If you read a study that would have never been published had the researchers found the opposite of what they found (e.g., failed to reject the null), you should be particularly worried about over-comparing and under-reporting. But if a study reports about questions for which we care about the answer intrinsically, regardless of what that answer turns out to be, a lot of the problems of over-comparing and under-reporting disappear. In particular, if the findings can be published regardless of the result, we can stop worrying about the file drawer problem and p-screening. Happily, many important questions fall into this latter category. By contrast, a lot of cute and fun questions fall into the former category. Unfortunately, the science press seems to love studies on those cute questions. (We're

looking at you Shankar Vendatam and Stephen Dubner.)

Don't obsess over statistical significance

The threshold .05 is just an arbitrary number. Substantively important effects may be statistically insignificant and statistically significant results may be substantively unimportant. Statistical hypothesis testing is a useful tool, but it can be abused. We should use p-values when appropriate. But they're not the end-all-be-all for assessing the believability of a quantitative finding. You can't just calculate. You have to keep thinking clearly. In Part IV of this book, we'll talk about how to incorporate empirical evidence with our prior beliefs in order to think clearly about what our beliefs should be after seeing some new evidence evidence in order to make the best possible decisions.

Beyond Science

We have focused on the ways in which over-comparing and under-reporting create deep challenges for the scientific enterprise. But, as the story of Paul the Octopus illustrates, the problem is broader than that. Indeed, we suspect you run into it on a regular basis, often without noticing.

Suppose someone is trying to sell you something—perhaps a car, financial advice, or a subscription to a dating app. The salesperson might tell you, “this car was rated number 1 in customer satisfaction 5 years after purchase!” Sounds great. But you might want to ask yourself how many measures of satisfaction they looked at before finding one on which this car was rated number 1. Did they look at reliability, repair record, safety, longevity, gas mileage, and resale value in addition to customer satisfaction? Did they also look 1 year from purchase, 2 years from purchase, 3 years from purchase, and so on? If so, they made a lot of comparisons, and told you about the one that puts the car in the best possible light.

This is not an unbiased estimate of the car's quality.

Similarly, your financial advisor might tell you, "this mutual fund outperformed the S&P 500 for 7 of the last 8 years." That sounds good. But how did it do relative to the Dow Jones or a broad market index? How did it do over the last 9 years, 10 years, 15 years? Did the advisor choose to compare to the S&P over the last 8 years because that was the natural comparison or because it was the comparison that made the mutual fund look best?

In general, then, you need to think clearly about the problem of over-comparing and under-reporting not just when formally thinking about hypothesis testing and statistical significance. Whenever you are offered a piece of evidence, you should be asking yourself whether this particular comparison is the natural one or the first one you would have thought to look at. If not, you might pause to contemplate how many plausible comparisons there are and, thus, how difficult it would have been to come up with some comparison that made whatever point the speaker was trying to make.

In the spirit of appreciating the fact that this problem is, indeed, everywhere, let us leave you with one final story that illustrates yet another way in which over-comparing and under-reporting frequently rears its head—identifying superstars.

Superstars

We like to admire and study people who are really successful. We've already seen one reason why this can produce misleading inferences, correlation requires variation. Another reason that we shouldn't be so quick to admire and study superstars is that there may not be anything particularly special about them beyond good luck.

Bill Miller majored in economics in college, served as a military intelligence officer, dabbled in a Ph.D. program in philosophy, and worked as treasurer for a steel and cement company before taking a position at Legg Mason Capital Management as director of research

in 1981, at the age of 31. Miller was clearly a smart guy with a promising career ahead of him. The next year, he began running the Legg Mason Value Trust mutual fund. For the first decade or so, the fund's performance was mediocre, slightly underperforming the market average. But Miller eventually hit his stride, scoring some big returns in the late 1990s and early 2000s. By 2006, fellow investors and reporters noticed that Legg Mason Value Trust had outperformed the market for 15 years running, an unprecedented streak that launched Bill Miller into the upper echelons of finance stardom.

Inevitably, everyone wanted to know Miller's secret. What made him such a successful investor? Perhaps surprisingly, Miller didn't achieve his success by developing intimate knowledge of niche industries or technical trading algorithms. His fund primarily invested in a small number of already well known companies like Google, Amazon, eBay, JP Morgan, and Aetna. When describing his investment philosophy in a 2006 letter to investors, Miller reported that he simply looks for the "best value." He further speculated on what separates his fund from so many competitors: "We differ from many value investors in being willing to analyze stocks that look expensive to see if they really are. Most, in fact, are, but some are not."

Miller makes it sound easy. He just invests in companies that are undervalued. But before we conclude that Bill Miller is a genius investor, let's consider the possibility that Miller was simply lucky, like Paul the Octopus.

There is an idea in finance called the efficient-market hypothesis. More or less it says that no fund or investment strategy should be able to systematically beat the market average over the long run. Loosely, the logic goes like this. If some genius investor came up with an investment strategy that predictably beat the market, other investors would mimic that strategy. This would change the prices of the assets traded under that strategy. Investors would keep doing this until that strategy no longer beat the market. For instance, if a company's stock price fully reflects all available information about the value of that

company, which we'd expect in a large market with lots of people trading on the best available information, there should be no way to systematically predict whether the stock price will go up or down without insider knowledge.

If the efficient-market hypothesis is right, then Miller and the other fund managers and stock pickers are just doing the equivalent of flipping coins. And we know that if enough people flip coins, a few of them will a long string of heads, by sheer luck. So, to assess whether Miller is indeed a genius, we need to ask how likely it is that he just happens to be the one guy who hit a long string of heads by luck.

To get started, let's imagine that beating the market is really just like flipping heads with a fair coin. This is our null hypothesis. Then we want to ask, if our null hypothesis is true, how likely is it that someone would flip 15 heads in a row.

The chances that a given investor in a given 15-year period beats the market every single year by chance is really low. The probability some investor beats the market by luck in one year is $\frac{1}{2}$. The probability an investor beats the market by luck two years in a row is $\frac{1}{2} \times \frac{1}{2}$. Extending this logic, the probability a given investor beats the market 15 years in a row by sheer luck is $\frac{1}{2}^{15}$ or about 1 in 30,000. So maybe Miller is a genius; if he was just flipping coins there's only a 1 in 30,000 chance that he would be so successful. But maybe not. Let's make sure we are thinking clearly about a few things.

There are lots of investors out there, and if any one of them beat the market 15 years in a row, they would have been just as famous as Miller, and we'd be asking them about their investment strategy instead of his. Therefore, the relevant question is not how likely it is that one particular fund manager, Bill Miller, would beat the market 15 years in a row by chance. We should be asking how likely it is that some fund manger, who turned out to be Bill Miller, would beat the market 15 years in a row by chance.

Notice, this is just like publication bias resulting or the problem of Paul the Octopus.

With publication bias, we only hear about the few studies out of many with statistically significant results. With Paul, we only heard about the one animal out of many who correctly forecast a lot of soccer games. And, similarly, we only hear about the few investors who have really long winning streaks. In all three cases, if we only think about the studies, animals, or investors we get to hear about, we over-estimate how likely it is that there success reflects a real phenomenon in the world.

In any given year, there are at least 24,000 professional funds trading, and presumably each will continue trading if it beats the market. So let's assume (as our null hypothesis) that there are 24,000 fund managers, none with any special insight. That means that whether each of them beats the market in any given year is a 50-50 proposition. So figuring out how likely it is that one of them beats the market 15 years in a row is just like figuring out how likely someone is to flip 15 heads in a row, if 24,000 people are all flipping 15 fair coins.

Doing the same kind of calculation we did for Paul the Octopus, the answer is about .52.¹ It is very unlikely that any particular investor will beat the market 15 years in a row by sheer luck. But when you consider the thousands of investors out there, it's actually more likely than not that one of them will beat the market 15 years in a row, even if none of them has any special insight and they're all just flipping coins.

These calculations look even worse for Miller if we consider that a 15-year streak would have seemed just as impressive had it started in another year. Once we consider all the funds and all the possible 15-year periods, it seems extremely likely that some fund manager would have such a streak at some point just by chance. These calculations, combined with our knowledge of the efficient-market hypothesis, should make us skeptical of anyone who

¹The probability that one investor gets 15 years in a row right is $.5^{15}$. So the probability that one investor doesn't get 15 years in a row right is $1 - .5^{15}$. So the probability that none of 24,000 get 15 years in a row right is $(1 - .5^{15})^{24,000}$. So the probability that at least one investor does get 15 years in a row right is $1 - ((1 - .5^{15})^{24,000})$, which is approximately .52.

claims to know the secrets to beating the market. The sheer number of traders and funds means that there are bound to be some exceptionally good track records. And those are the ones we hear about. So, before handing over your life savings to an investment manager, you should ask whether you would invest the same money betting on Paul the Octopus' soccer picks? If not, let us recommend that you consider low cost index funds.

What do you think happened to Bill Miller after his flurry of press coverage in the mid-2000s? The streak ended in 2006. His fund lost 55% of its value during the 2008 financial crisis, the fund continued to trail the market for several more years, and he eventually stepped down from his post in 2012. Looking across the full 30-year period in which Miller managed Legg Mason Value Trust, it actually *underperformed* the market. Alas, his historic winning streak still earns him regular appearances on cable news programs where he pontificates on market conditions and hot stock picks. In 2017, his new fund, Miller Opportunity Trust, was once again making news for impressive returns in 2017. The secret? A big bet on Apple, the most highly valued company in the world.

Key Terms

- **Publication bias:** The phenomenon whereby published results are systematically over-estimates because there is a bias toward publishing statistically significant results.
- **p-hacking** Searching over lots of different ways to run an experiment, make a comparison, or specify a statistical model until you find one that yields a statistically significant result and then only reporting that one.
- **p-screening:** A social process whereby a community of researchers, through its publication standards, screens out studies with p-values above some threshold, giving rise to publication bias.

Exercises

1. In late April of 2020, the National Institutes of Health announced the results of a study on the use of the drug remdesivir to treat COVID-19. Some COVID-19 patients were randomly given remdesivir others were given a placebo. The study found statistically significant evidence that treatment with remdesivir reduced recovery time, as measured by the number of days it took for a patient to be discharged from the hospital after being put on the drug. The study was double blind (neither the patients nor doctors knew whether a subject had been put on the real drug or the placebo). The study size was reasonably large (hundreds of patients). And treatment assignment was random.
 - (a) Based on the lessons of this chapter, identify two more pieces of information that would help you assess how confident you should be in the efficacy of remdesivir.
 - (b) It turns out, the study was pre-registered. The pre-registration plan identifies 28 outcomes that the scientists are going to measure. How does this change your beliefs about whether the findings identify a real effect in the world? Why?
 - (c) The pre-registration plan also identifies one outcome as the primary outcome of interest. That primary outcome of interest was the one reported, how long it took for a patient to be discharged from the hospital. Does this affect your answer to the previous question? Why or why not?
 - (d) But wait, there is one final twist. The pre-registration plan was actually revised during the course of the study. It turns out, length of hospitalization was not listed as the main outcome of interest until a revision on April 16, 2020. Prior to that, the primary outcome of interest was listed as a patient's score on an 8 point scale measuring disease severity. This is reflected in the April 2, 2020 version of the plan. The researchers, in a statement, explained that they had not

seen the data coming out of the study prior to changing their primary outcome of interest. Reflect on how all of this affect your views on the study findings?

2. Download "VoterSurveyData2016.csv" and "README.txt" from [INSERT LINK HERE], and read the README file. Suppose we want to know whether prior exposure to Donald Trump before he was a politician affected political behavior in the 2016 U.S. presidential election. To proxy for exposure to Trump, a survey asked people whether they watched *The Apprentice*, a television show starring Trump, and whether they watched *Home Alone 2*, a movie featuring a cameo by Trump.

- (a) Using the data available, try to find at least three interesting, statistically significant relationships suggesting that prior exposure to Trump corresponded to political behaviors in the 2016 presidential election.

If you're struggling to find statistically significant relationships, think about all the different things you can test for. You could use having seen *The Apprentice*, *Home Alone 2*, or both as your measure of prior Trump exposure. You can use support for Trump, support for Clinton, or voter turnout in 2016 as your outcome of interest. You can subset the data to look specifically at subgroups of interest (e.g., women, black voters, southerners, rich people, young people, etc.).

- (b) Once you've found three statistically significant relationships, interpret them substantively and think about what they mean. Did you learn something interesting about American electoral behavior?
 - (c) Think about what you just did in question 2. How does it relate to this chapter? What concerns does this give you about your findings?
 - (d) The data you just analyzed is real survey data from the 2016 Cooperative Congressional Election Study. We randomly selected 1,000 respondents and shared a subset of their responses with you. However, we lied above when we said that

respondents were asked whether they watched The Apprentice or Home Alone 2. We made those variables up. (Sorry, we won't lie to you again.) Furthermore, the values for those variables were actually generated completely at random. Explain why you were nonetheless able to find a relationship between those variables and political behavior. Would you expect that relationship to continue to hold if we provided data on another 1000 respondents and again randomly generated the data.

Notes

For more on p-hacking, see

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant *Psychological Science* 22(11):1359-1366.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2014. P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology* 143(2):534-547.

To see the ESP study and to read about some of the failed replications, see

Bem, Daryl J. 2011. Feeling the Future: Experimental Evidence for Anomalous Retroactive Influence on Cognition and Affect. *Journal of Personality and Social Psychology* 100(3):407-425.

Galak, Jeff, Robyn A. LeBoeuf, Leif D. Nelson, and Joseph P. Simmons. 2012. Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology* 130(6):933-948.

For more on p-screening in the context of get-out-the-vote studies, see

Green, Donald P., Mary C. McGrath, and Peter M. Aronow. 2013. Field Experiments and the Study of Voter Turnout. *Journal of Elections, Public Opinion and Parties* 23(1):27-48.

The essay by 72 researchers on lowering our threshold for statistical significance is

Benjamin, Daniel J. et al. 2017. Redefine Statistical Significance. *Nature Human Behavior* 2:6-10.

The study on the frequency of null results in NHLBI studies before and after pre-registration is

Robert M. Kaplan and Veronica L. Irvin. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLoS One* 10(8).

To read more about whether college football outcomes influence elections, see

Andrew J. Healy, Neil Malhotra, and Cecilia Hyunjung Mo. 2010. “Irrelevant Events Affect Voters’ Evaluations of Government Performance.” *Proceedings in the National Academy of Sciences* 107(29):12804–12809.

Anthony Fowler and B. Pablo Montagnes. 2015. “College Football, Elections, and False-Positive Results in Observational Research.” *Proceedings in the National Academy of Sciences* 112(45):13800–13804.

If you would like to see the full history of revisions of the pre-analysis plan for the remdesivir study mentioned in Exercise 1, you can find it here <https://clinicaltrials.gov/ct2/history/NCT04280705>.

Chapter 8

Reversion to the Mean

What you'll learn

- Lots of things tend to revert toward the mean, meaning that extreme observations are often followed by less extreme observations.
- This phenomenon will arise for virtually any outcome that is a function of both signal (i.e., something real in the world) and noise.
- If you don't think clearly about mean reversion, it is easy to misinterpret evidence.
- One place where we shouldn't expect mean reversion is in stock prices.

In Chapter 7 we saw that, because of over-comparing and under-reporting, it makes sense to be skeptical of new, exciting, scientific findings. And so, when such findings are initially reported, often the first question that gets asked is “does the result replicate?” That is, if we were to run a new, independent study designed similarly to the original study, would we find a similar effect? The instinct to ask this question is solid: we are worried that some reported results reflect the vagaries of chance, rather than real phenomena in the world. So, before hanging our hats on new findings, we want to see that they hold up in multiple studies. And, in fact, researchers often fail to replicate hyped results in follow-up studies. Indeed, in some areas this is so common that people have started talking about a *replication crisis* undermining confidence in entire fields of scientific enquiry.

Jonathan Schooler, a prominent psychologist at UC Santa Barbara, famously noticed such a pattern of replication failures in some of his own most influential studies. Interestingly, the effects Schooler estimated typically didn’t disappear entirely when replicated. But they did get systematically smaller. And for some reason, the more he replicated his studies, the smaller and smaller the results got. He asked around and found that many colleagues had experienced the same thing. Replicated results were often smaller than the original findings.

One potential explanation for this pattern is that once you’ve done a study and subjects are aware of the results, they change their behavior. This phenomenon, whereby subjects change their behavior because they know they’re being studied, is sometimes called the *Hawthorne effect*.¹ People also use the term *demand effect* to refer to situations where the subjects behave differently because they know what the experimenters are looking for and are trying to please them.

¹Fun fact. The term *Hawthorne effect* comes from a study of the relationship between working conditions and productivity at the Hawthorne Works factory outside Chicago. But it turns out that the data were analyzed badly. The economists Steven Levitt and John List reanalyzed the original data and showed that what looked to the original researchers like a Hawthorne effect was more likely attributable to other factors, such as differences across days of the week rather than the subjects changing their behavior in response to being studied.

Schooler and his colleagues quickly ruled out the Hawthorne effect and other similar explanations because they found the same pattern in studies of birds, who presumably have no idea what's being studied and don't care one way or the other about pleasing human researchers. So what else could explain the peculiar pattern of disappearing effects?

Schooler started referring to this phenomenon as *cosmic habituation*. He wondered whether there is some unknown force in the universe that causes effects to shrink every time they're studied. One analogy he gives is to the habituation of human sensory perception. When something first touches your arm, you are acutely aware of it. However, over time, you habituate, and your sensation of being touched diminishes. Maybe the universe is like that. The first time we observe some phenomenon, there is an acute effect. But, over time, the universe habituates to our studies and we observe the effect less and less. In other words, scientists actually alter reality every time they study it. Spooky.

Schooler's theory of cosmic habituation has received significant media attention, including an episode of the popular radio show and podcast *Radio Lab* and an article in *The New Yorker* entitled "The Truth Wears Off."² But before we follow Schooler down the path of hypothesizing new cosmic forces, let's see if thinking a little more clearly can help us resolve this puzzle a bit less mystically.

Francis Galton and Regression to Mediocrity

As we described back in Chapter 5, Francis Galton made a similarly eerie discovery in the 1860s. He collected data on the size of parents and their children. He did this for humans. He also did it for plants, collecting data on the size and weight of the seeds of parent and child sweet-peas.

Galton drew scatter plots of these kinds of data, putting the parents' size on the hor-

²<http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off>

horizontal axis and the children's size on the vertical axis. Then he plotted a regression line through the data. You can see an example of one of Galton's plots, for the heights of parents (adjusting for biological sex) and their children, in Figure 8.1.

Originally, Galton expected the regression line would be a 45-degree line—i.e., that its intercept would be zero and its slope one. That seems a reasonable guess. It would be true if, on average, children are the same size as their parents (again, adjusting for biological sex).

As it turns out, however, this guess is not correct for humans or for sweet-peas. Let's look at Galton's figure. In that figure, the 45-degree line is labeled “mid-parents.” The phrase refers to Galton's measure of the average height of a child's parents, after first adjusting to put female and male heights on the same scale. The (unfortunately, unlabeled) x-axis corresponds to parent's height. Then, the line tracing out parent's height must be the 45-degree line. The line labeled “children” is the regression line running through data which takes as its x-value the parents' height and as the y-value the child's height.

As you can see in the figure, Galton's regression shows a positive y-intercept—at the lowest value on the x-axis (parents' height), the regression line lies above the 45-degree line. And Galton's regression line has a slope that is positive, but distinctly less than one—the regression line is increasing more slowly than the 45-degree line.

What does this regression imply about the relationship between parents' height and children's height. The fact that the slope is positive means that, on average, the taller the parents, the taller their children. As you can see in the figure, the fact that the y-intercept is positive means that particularly short parents tend to have children who are taller than they are. On the left-hand side of the x-axis (where parents are short), the regression line lies above the 45-degree line. But because its slope is less than one, the regression line increases less quickly than the 45-degree line. And, indeed, the two lines cross in the middle. As such, on the right-hand side of the x-axis (where parents are tall), the regression line lies

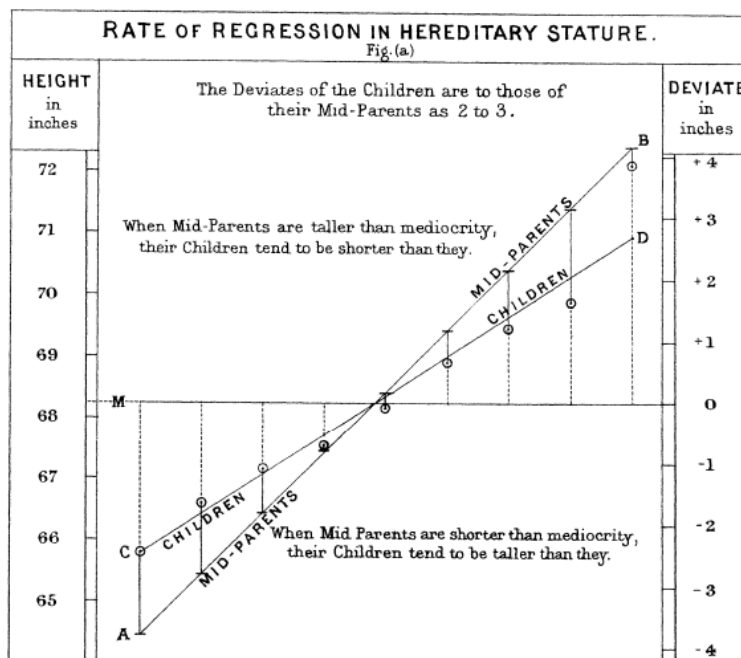


Figure 8.1: Galton's original illustration of mean reversion.

below the 45-degree line. Particularly tall parents tend to have children who are shorter than they are.

As we mentioned earlier, Galton referred to this phenomenon as “regression to mediocrity.” And it is because of this word choice that we now use the word regression to refer to the practice of fitting lines to data. It is also why some people refer to the phenomenon whereby things tend to revert toward the mean (which is the subject of this chapter) as regression to the mean. However, to avoid confusing the two concepts, we will refer to the latter by its other common name, *reversion to the mean*.

Galton's findings sound a lot like cosmic habituation. Perhaps there is some unseen force in the universe that pushes size within families toward the average. Maybe when the universe sees tall parents or big sweet-pea seeds, it restores order by making their offspring

smaller. Or perhaps when Galton measured the heights of the parents and the diameter of sweet-pea seeds, he somehow made their off-spring shrink! Galton was probably perplexed by his unexpected findings. But he wasn't so quick to jump to supernatural conclusions.

Eventually, Galton realized what was going on. Size is determined by many factors. For the purposes of seeing Galton's thought as clearly and simply as possible, let's think about sweet-pea seeds. And let's imagine that a sweet-pea seed's size is influenced by just two things: (1) the genes it inherits from its parent and (2) the amount of direct sunlight it gets while growing. The larger the genes a seed get from its mother, the larger it will be, all else equal. And the more sunlight it gets, the larger it will be, all else equal.

Under this simple model, let's think about how a seed can end up especially large or especially small. Suppose you find a really large sweet-pea seed. It could be large because it got really large genes from its mother. It could also be large because it happened to grow in a year with uncommonly good sun. Or it could be some combination of the two. Odds are, if a seed is *really* big, it had both factors working in its favor: a parent with large genes and excellent sun.

So what should we expect if these really big seeds are planted and produce offspring of their own? They'll pass along their larger-than-average genes. But, most likely, the child seed won't experience the same outstanding sunlight as its parent did. On average, it will grow in average sun. So the child will be larger than average because of the genes it inherited. But the child will probably be smaller than its parent because its parent got particularly lucky with respect to sun exposure. The same holds for the children of very small seeds. They get their parents' small genes. But they are likely to experience better sunlight than their parents did and, thus, be larger than their parents.

So, if this simple model were correct, we would observe exactly Galton's pattern. Larger plants tend to have larger children (the regression line has a positive slope). But size reverts toward the mean—really small parents tend to have children who are smaller than average

but larger than they are, and really large parents tend to have children who are larger than average but smaller than they are (the slope of the regression line is less than 1).

Obviously, seed size is more complicated than this. Many things influence it other than genes and sun. But the example makes the point. We should expect reversion to the mean if size is partly determined by genes that are systematically transmitted from parent to child and also partly determined by idiosyncratic or random factors that are uncorrelated across generations (like sun exposure). The same goes for human height, as we see in Galton's plot.

More generally, there will be mean reversion for any outcome that's partly a function of systematic factors (which we sometimes call *signal*) and partly a function of random or idiosyncratic factors (which we sometimes call *noise*). Imagine an outcome observed over and over, where with each observation, the outcome reflects a combination of a systematic signal (e.g., the genes) and random noise (e.g., sunlight). Extreme outcomes typically arise because of extreme values of both the signal and the noise. In other iterations, while the signal stays fixed, the noise takes a new, random value. And, in expectation, the value of the noise will be average. So extreme values in one iteration are expected to revert toward the mean in other iterations.

Many phenomena in the world have this signal and noise structure. So we should expect reversion to the mean to pop up a lot. As such, thinking clearly about evidence requires understanding reversion to the mean, expecting it, and accounting for it when possible. In what follows, we first delve a little more deeply into the nature of mean reversion, to make sure we are clear on exactly what is going on. We then consider a variety of different real-world settings to understand when we should and should not expect mean reversion to appear.

Mean reversion is not a gravitational force

One common misconception about mean reversion is that it reflects something like a gravitational pull. That is, that the world is full of outliers and that ineluctably, over time, things are being pulled toward the mean. This isn't right.

To see if you are thinking clearly about this, try answering each of the following questions:

1. John Junior is exceptionally tall. If you had to guess, would you guess that John Junior's son, John III, is: (a) Shorter than John Junior. (b) The same height as John Junior. (c) Taller than John Junior.	2. John Junior is exceptionally tall. If you had to guess, would you guess that John Junior's father, John Senior, is: (a) Shorter than John Junior. (b) The same height as John Junior. (c) Taller than John Junior.

Before we turn to the answer, let's start by noting that you should have given the same answer to both questions. Understanding why is essential.

For many people, once they learn about mean reversion, question 1 is pretty intuitive. John III is probably shorter than John Junior. John Junior is particularly tall. So he probably has tall genes (signal). And he also probably had idiosyncratic things happen that led him to grow particularly tall (noise). His son, John III, will likely inherit his tall genes. But, if you had to guess, you'd guess the idiosyncratic other factors will likely be more average. This is the logic of mean reversion as explained by Galton.

But, in our experience, question 2 is often a bit more vexing. Many folks, including one of Anthony's professors at MIT (a world-renowned expert on the analysis of quantitative

data), reason as follows. John Junior is really tall. And there is mean reversion in the world. So John Junior's height is probably closer to the mean than was his father's. Factoring in mean reversion, for John Junior to be so tall, his father must have been a virtual giant! Hence, they reason, while the answer to question 1 is (a), the answer to question 2 must be (c).

It's okay if you thought something along those lines. The argument has a certain appeal. But it is wrong, and it's important that you see why. The answer to both questions 1 and 2 is (a). And the logic for John Junior's father is identical to the logic for John Junior's son. It is the logic of mean reversion. Here's how it goes, one more time.

Suppose you observe some outcome made up of signal and noise and that outcome is surprisingly large. (The argument, of course, works for surprisingly small too.) Then suppose you are going to observe another outcome, where the signal is the same as your first observation, but there will be a new draw of the noise. Since the first observation is so large, it probably reflects a large value of the signal and a large value of the noise. The new observation will again have a large signal value, but the noise is likely to be smaller. So that new observation will likely be smaller.

Importantly, in making this argument, we said nothing about which outcome was determined first in time. We just talked about the order in which you observed them. Mean reversion is not a gravitational force pulling things towards the average over time. For the logic of mean reversion, it makes no difference which came first temporally. So, if John Junior is very tall, and his son has the same signal (genes) but different noise, then his son is probably shorter than him. And, if John Junior is very tall, and his father has the same signal (genes) but different noise, then his father is also probably shorter than him.

The easiest place to see this in the real world is in data from athletic competitions, where we observe the same competitor doing the same task over and over again. And, what we see in those settings, is that mean reversion characterizes the data, moving forward or

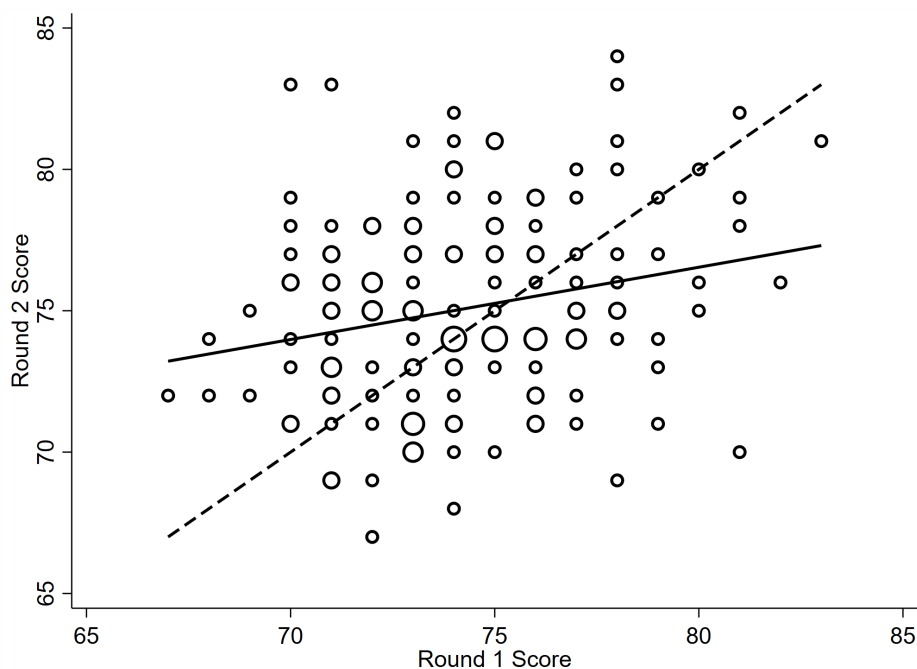


Figure 8.2: Scores across rounds in the 2013 US Open golf tournament.

backward in time.

Figure 8.2 is a scatter plot of scores from the first two rounds of the 2013 U.S. Open golf tournament. A player's score from round 1 is on the horizontal axis and from round 2 is on the vertical axis. In cases where multiple players had identical scores, the size of the circle is increased to reflect the number of players. The 45-degree line is dashed, and the actual regression line that best fits the data is solid. What do you see here?

On average, there is a positive correlation between scores in the two rounds, the regression line is sloping upwards. The players who did better in round 1 also tended to do better in round 2. That makes sense, some players are better than others (that's the signal). But the slope of the line is far less than 1, the regression line is shallower than the 45-degree line. In general, if your round 1 score was worse than average, your round 2 score was better than your round 1 score. And if your round 1 score was better than average, your

round 2 score was worse than your round 1 score. It's exactly the same as the pattern with the heights of parents and children or the size of mother sweet-peas and their offspring discovered by Galton. And we can assure you that we didn't cherry pick this example. It's a virtual guarantee that you'll see this same pattern for almost any golf tournament.

A golf commentator might look at these data and tell a story. Maybe the players who had a really good first round succumbed to the pressure. They choked. And that explains why they did worse in round 2. And maybe the players who had a bad first round realized they had to put in more practice, change strategy, or really focus. And they accordingly improved their scores.

That's possible. But it could also just be mean reversion. Golf scores are a function of both skill (signal) and luck (noise). The players with the best score in a given round are probably better than the average player in the field. But they probably also had some luck on their side. A few putts went in that could have just as easily lipped out. Their few bad shots got fortunate bounces, keeping them out of trouble. And so on. In their other rounds, they're still better-than-average players, so we expect their scores to be better than average. But they probably won't have the same good luck, so their score will probably be worse than in that exceptional round.

Notice, nothing in the previous paragraph said anything about which round came first. This reflects the fact that reversion to the mean isn't some gravitational force pulling things towards the mean over time. And this realization allows us to probe which story—the commentator's or mean reversion—is more likely to be true.

Think about a player who has a particularly good score in round 2. Should we expect their round 1 score to be better or worse than their round 2 score? The old temptation was to think that because of reversion to the mean, in order for them to have a good score in round 2, they must have had a really good score in round 1, allowing them to still have a good round 2 score despite reversion to the mean. But we now know better.

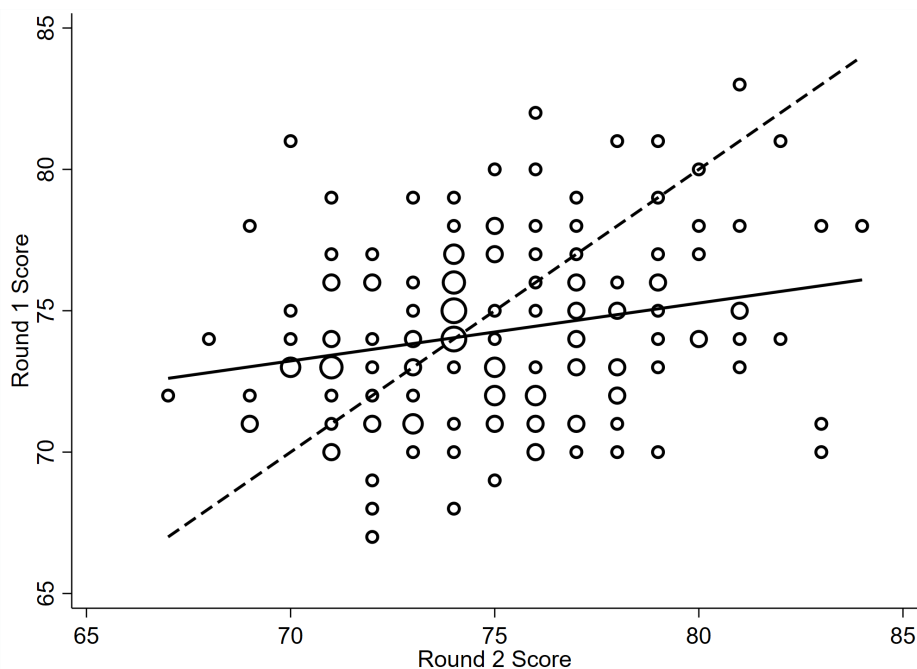


Figure 8.3: Scores across rounds in the 2013 US Open golf tournament with flipped axes.

The logic of mean reversion has nothing to do with time. The score in each round of golf is a combination of signal and noise. If a player had a particularly good round in some round, we should expect a different round by that player (with the same signal but different noise) to be lower, regardless of which round came first. And if a player had a particularly bad round in some round, we should expect a different round that player to be better, regardless of which round came first.

Figure 8.3 shows the same graph you saw before but with the axes flipped, so that round 2 is on the horizontal axis and round 1 is on the vertical axis. The overall pattern is almost exactly the same. People who had particularly good scores in round 2 were better than average in round 1 but still worse in round 1 than in round 2. Just like with John Junior and John Senior reversion to the mean is perfectly happy to work backward in time.

Seeing both versions of the graph demonstrates why the commentator's explanation is

unlikely to be right. It is hard to believe that a good score in round 2 causes players to feel extra pressure which somehow goes backward in time, making their round 1 score worse. And yet, we see the same pattern whether we look moving forward or backward in time. The explanation is mean reversion.

Seeking Help

Mean reversion can be particularly problematic for clear thinking in settings where we seek help when things have gone unexpectedly wrong. Why is this?

If things have gone unexpectedly wrong, that suggests there is some underlying expectation of how things should be going—perhaps formed based on long past experience—and that we have deviated from that expectation in a bad direction. We might think of the expectation of how things should be going as reflecting the signal. And deviations from that expectation might reflect noise.

Let's think about that in a couple of concrete settings.

Suppose you are a relatively healthy person. You might think of good health as reflecting the true underlying signal for you. Even most fundamentally healthy people feel unwell from time to time—because of the flu, a back ache, or what have you—for idiosyncratic reasons that perhaps do not reflect any fundamental change to their underlying state of healthiness. And most healthy people also have days when they feel particularly vigorous and well. Thus, we might think of a day when you feel fine as reflecting your true signal, with very little noise. Days when you feel sick can be thought of as days with particularly negative values for the noise. And days when you feel like you could go out and conquer the world can be thought of as days with particularly positive values for the noise.

Or maybe an example about school will speak more directly to our readers. You have some underlying level of academic skill that reflects how strong a student you are in any

given area. That is your true signal. But some days you do way better than normal on a test—perhaps reflecting a particularly lucky draw of questions or a particularly good night’s sleep. And some days you do way worse than normal on a test—perhaps reflecting an unlucky draw of questions or a late night. These idiosyncratic features constitute noise.

Now, what does this have to do with mean reversion? Ask yourself, on which days is a person likely to seek help from, say, a chiropractor? Probably on days when they wake up with a back ache that feels worse than normal. That is, on days when the noise is particularly negative. And ask yourself which students seek out help from a test prep company? Probably students whose performance on an important standardized test was worse than they had anticipated based on their sense of their underlying ability. If this is right, then mean reversion tells us that, even if chiropractic adjustments or test prep strategies don’t actually help at all, we should expect people who seek out this kind of assistance to see improvements. And, if they aren’t thinking clearly about mean reversion, they are likely to give too much credit to the chiropractor or the test prep company. Mean reversion can be a good business model.

This kind of problem is pervasive. We’ve actually already seen an example in Chapter 1. In that chapter, we discussed broken-windows policing. Recall, in New York City, the police targeted a new policing strategy at the highest crime precincts and found that, after the change of strategy, crime went down in those precincts. But this is just what we’d expect from mean reversion, even if broken-windows doesn’t work at all. The highest-crime precincts will tend to get better and the lowest-crime precincts will tend to get worse, regardless of any policy change. Because of this mean reversion, policing strategies that target the highest-crime precincts will look to the naive observer as if they are really effective, even if they aren’t.

And, actually, reversion to the mean was lurking underneath another of our examples in Chapter 1. Remember when Ethan’s son’s doctors recommended that he try out a gluten-

free diet because he was underweight. Their idea was that, if his weight percentile increased once he went on the gluten free diet, that would be evidence of gluten intolerance. But reversion to the mean says that we should have expected Abe's weight percentile to increase even with no intervention. Month to month, a baby's weight is a function of both signal (e.g., health, genetics) and noise (e.g., random features of the environment, idiosyncrasies in the growth trajectory). If a kid has off-the-charts low weight in one month, he probably had extremely low values on the noise terms that month. Over time, we should expect more average values on the noise, so his weight percentile should increase. It would be a mistake to interpret this, on its own, as evidence that some change in behavior (e.g., going gluten free) explains the weight increase.

Once you think about it clearly, you'll see that all kinds of interventions and treatments will look like they work even if they do nothing. People typically seek out help when things are at their worst. We'd expect things to improve, even without an intervention, because of reversion to the mean.

There are many expensive medical treatments for which the best available evidence is not so different from the evidence for broken-windows policing or SAT prep. For instance, there are no randomized trials validating the efficacy of a variety of kinds of surgery. Consider a patient who goes to the surgeon with joint pain of some sort. The doctor recommends surgery. At the end of a recovery period, the patient says she feels better. Now, the doctor may have all sorts of knowledge about body mechanics and physiology that provide some good reasons for believing the surgery should help. But we should at least entertain the possibility that there also is mean reversion at work—that is, that many patients would have experienced at least some significant improvement without surgery.

Indeed, once a randomized trial is run, researchers sometimes find that a common surgery does not in fact provide the hypothesized benefits. For example, in a 2002 study of arthroscopic surgery for osteoarthritis of the knee, researchers found that the commonly

prescribed surgeries had no detectable effect on knee pain. Yes, patients reported less knee pain two weeks after surgery, but other subjects who simply received skin incisions and were told that they received the surgery reported the same reduction in knee pain. That's right, doctors actually gave sham surgeries to some of their patients, and those deceived patients were no worse off than those who received the real surgery. Why did all the patients seem to feel better? Presumably, you only go under the knife for knee pain when it's especially severe, and it looks like most of those patients would have felt better in a few weeks even without the surgery.

Mean Reversion, the Placebo Effect, and Cosmic Habituation

We've just seen that, if we fail to think clearly about mean reversion, we are likely to misinterpret the extent to which certain kinds of interventions, including medical interventions, are actually responsible for improved outcomes. But, it turns out, mean reversion can even create problems when we try to do careful scientific studies. Let's see why this is the case in settings. First, we will consider the much discussed *placebo effect*. Then we will revisit the problem of cosmic habituation with which we opened this chapter.

The placebo effect

Few phenomena in medicine are cited more than the so-called *placebo effect*. Many people suspect that the knowledge that we've undergone treatment somehow activates the body's own healing powers, independent of the direct effects of the treatment itself. For this reason, medical researchers are careful to compare the effectiveness of new drugs to placebos. They want to account for the possibility that believing you're receiving a treatment will heal you all on its own. So they use things like sugar pills, so that experimental subjects don't know whether they are getting a real treatment or not.

Why do medical researchers, and others, think that there's a placebo effect? One source of evidence comes from medical trials themselves. In such experiments there is a treatment group that gets the drug and a control group that gets a placebo pill. And often, in such studies, both groups' health improves. The improvement of the control group is taken as evidence for the placebo effect.

But now you can see that this kind of evidence for the placebo effect is problematic. The people in the control group (and the treatment group) entered the medical study because they were unwell. We might expect them to tend to get better even in the absence of any treatment. This is not evidence that their minds are healing their bodies. It is mean reversion.

If you actually wanted to test for a placebo effect, you'd want to divide the pool of experimental subjects into a group that got a placebo pill and a group that got no treatment at all. (Of course, you could also have a group that got the real medicine. Let's not forget why we are here in the first place.) Few studies explicitly test for the effect of a placebo treatment relative to no treatment. The ones that do typically find no evidence of a placebo effect. Moreover, those studies that do find evidence for a placebo effect typically concern purely subjective outcomes. So people may perceive themselves to be feeling better after taking a placebo, even if they aren't objectively healthier.

For example, in 2011, a team of researchers from Harvard Medical School published a paper in the *New England Journal of Medicine* on the effects of real medical treatments vs. placebo treatments vs. no treatment for asthmatic patients. Interestingly, both the real treatment (an albuterol inhaler) and the placebo treatments (a placebo inhaler or acupuncture) led patients to report that they felt better. But when the scientists actually measured their lung capacity, only the real treatment had an effect, the placebo treatments were no better than doing nothing. So to the extent that there's evidence for a placebo effect, it's evidence of mind over mind, not mind over matter.

In short, once we think clearly about reversion to the mean, we see that there's little compelling evidence of a placebo effect in medicine. Yet somehow, almost everyone believes in the placebo effect because they're bad at recognizing and correctly interpreting reversion to the mean. Even some of the greatest medical researchers have fallen victim to this confusion. Vitamin C is widely believed to provide significant health benefits, despite little evidence. (To be clear, a small amount of Vitamin C is necessary to avoid scurvy, but for almost everyone in the developed world who naturally consumes some Vitamin C, there is little evidence that additional Vitamin C is beneficial.) An important source of this widely held superstition is Linus Pauling, a world-famous chemist and two-time Nobel Prize winner who strongly advocated for Vitamin C. Some believe that Pauling knew that vitamin C had little effect, but he believed in the placebo effect, and he thought that telling people that a Vitamin C supplement, or a glass of orange juice, would heal them was a cheap and easy way to get the placebo effect working, inducing the human body to somehow cure itself.

Cosmic habituation, redux

If you are paying particularly close attention, you might have noticed a connection between the talk of signal and noise in this chapter and our favorite equation. Recall, our favorite equation says that an estimate from data is made up of three things: the true estimand, bias, and noise:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

Imagine a study that is really well designed to learn about an estimand so that there is no bias. The estimate generated by that study will be made up of the true estimand and noise from things like sampling variation. The true estimand stays constant across different studies of the same phenomenon. But the resulting estimates nonetheless vary from study to study because of the noise. So, if we imagine multiple studies of the same phenomenon,

we can think of the true estimand as the signal. And we can think of the noise as, well, the noise.

Given this, we should expect repeated scientific studies of the same phenomenon to exhibit mean reversion. If the first study found a particularly large relationship, it is probably the case that the true relationship in the world is big and it is probably the case that the sampling variation happened to create noise in the positive direction in that study. In the next study, we should expect to find a smaller estimate because, while the true relationship (i.e., the estimand) is still probably large, the noise from the sampling variation will probably not be as large the next time around.

With this realization, we are now, finally, ready to return to cosmic habituation and see why it probably isn't best explained by a mystical force whereby the universe accustoms itself to the activity of scientists here on earth.

If it isn't mystical forces, why do estimated effects tend to get smaller when replicated? Part of the answer is reversion to the mean. But that's not the whole answer. To really understand what is going on with cosmic habituation, you need to combine our new understanding of reversion to the mean with our discussion from Chapter 7 of over-comparing and under-reporting. Let's see why.

Imagine several scientists, independently studying some phenomenon—say whether giving people time to think leads them to make better decisions, which is one of Jonathan Schooler's areas of interest. Each scientist does a study. One finds that people who are given time to think make dramatically worse decisions. Another finds that people who are given time to think make slightly worse decisions. A third finds no relationship. And a fourth finds that people who are given time to think make better decisions. Given the sample size of the various studies, only the large negative finding is statistically significant.

What generates these different estimates? Presumably there is some true effect of giving

people time to think on the quality of their decisions. This is the estimand in our favorite equation. We can also think of it as the *signal* that is common across each of these studies. But then there are lots of idiosyncratic features, the *noise*, that affect the observed relationship (the estimate) in any given study. For instance, even though these are experiments, in any one of them, by happenstance, it could occur that the people given time to think turn out to be intrinsically much worse decision makers than the people not given time to think. This large, negative noise term would lead that study to find a particularly large negative effect. In another study the people given time to think might happen to be intrinsically slightly better decision makers than the people not given time to think. This positive noise term would lead that study to find a more favorable relationship between thinking time and decision making. So, we can see, the results of these studies are made up of both signal and noise. As such, we should expect replications to experience reversion to the mean.

Now let's think about which of the findings are most likely to be replicated. From our earlier discussion of p-screening and publication bias, if we had to guess, we'd guess that only one of these studies is notable enough to the original researcher or the scientific community to warrant an independent replication—the one with the large, statistically significant negative relationship between time to think and the quality of decision making. This study has two things going for it over the other studies. First, it has a statistically significant finding. Second, that finding is pretty surprising—who would have thought that thinking things through leads to worse decisions?

Imagine that this is what happens. One large, surprising finding gets published. It was found in a well designed study, so people think it is probably true. But, because it is an important result, scientists will also want to see if it will replicate. What should we expect them to find? Well, we just saw that a finding is more likely to get published and warrant replication (both because of its surprise value and because it is more likely to pass the statistical significance threshold) when the estimated effect size is particularly large in

magnitude. But we also know that particularly large estimates are probably the result of both large values of the signal *and* large values of the noise. So, because of reversion to the mean, when we go to replicate this study, we should expect to find a smaller (in magnitude) estimated effect size (as, indeed, the other three, unpublished studies had found). That is, because of a combination of over-comparing, under-reporting, and reversion to the mean, we should expect to see cosmic habituation!

Figure 8.4 is our favorite illustration of this phenomenon. The figure describes the changing evidence on the link between particular genes and particular diseases. The different colors represent different hypothesized gene-disease linkages. Each curve shows the sign and size of the estimated relationship, taking into account all the available data at any given point in time. In this particular graph, a value of 1 on the vertical axis means that the evidence shows no relationship at all between the gene and the disease. A value below 1 means that the evidence shows a negative relationship between the gene and the disease. And a value above 1 means that the evidence shows a positive relationship between the gene and the disease. The further from 1, the larger (in magnitude) the estimated relationship.

The data points on the far left of the plot are the estimated relationships between genes and diseases from the very first study on the topic. Moving to the right, the next data point shows the estimated relationship taking into account the data from both the first and second study. This continues as we move to the right, until we get to the most recent published study.

What you can see in Figure 8.4 looks like cosmic habituation. The first published study in each of these literatures finds a large relationship between a gene and a disease. This is the kind of study that gets published in a prestigious journal and covered in the press. But as scientists engage in replication, reversion to the mean kicks in. The magnitude of the estimated effects is typically smaller in subsequent studies. As we add more and more information, we get closer to the truth, which is quite far from the over-estimate reported

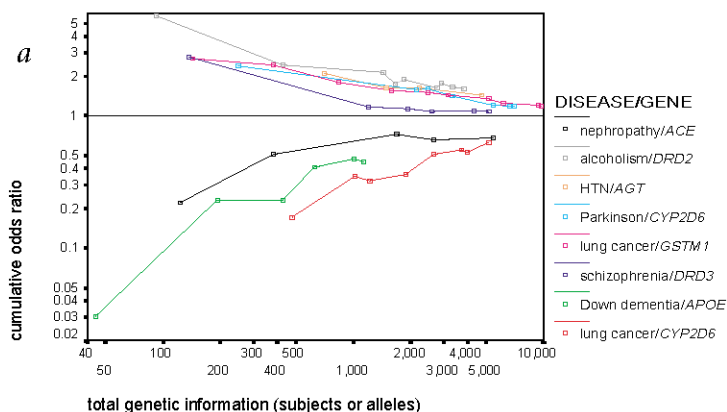


Figure 8.4: The estimated effect size shrinks as more data is accumulated in genetic studies.

in the initial study. As we see, by the end, the evidence suggests at most a very weak relationship between the gene and disease in question. The follow up newspaper story is rarely written.

Where should we *not* expect reversion to the mean?

Once you understand reversion to the mean, you should start to worry about it a lot. Commentators, analysts, and casual observers constantly misunderstand it, introducing complex theories to explain patterns that reflect a simple statistical phenomenon. The preceding discussion makes it sound like we should expect reversion to the mean almost everywhere, and to a close approximation, that's right. We should see reversion to the mean for any variable that is influenced by signal and noise. So instead of listing all the instances of reversion to the mean, let's think about situations where we shouldn't expect it.

Obviously, if the signal is much greater than the noise, then we wouldn't expect to see much reversion to the mean. Suppose we repeated the golf analysis above, but instead

of plotting scores from two different rounds within one tournament, we plotted average scores from two different seasons on the PGA Tour. An average score from an entire season contains a lot more information about a player's ability. Much of the good and bad luck that constitutes the noise across rounds averages out, so there will be less reversion to the mean in that picture (but still some).

But there are some situations where we should expect no reversion to the mean at all. Let's think about the stock market. Should we expect to find reversion to the mean in stock prices? Maybe we could beat the market and become billionaires by exploiting reversion to the mean.

Reversion to the mean would seem to predict that, on average, companies with low stock prices should rebound in the future, and companies with high prices should drop. It also seems to predict that increases should be followed by decreases and vice versa. Could we exploit this information by buying the stocks that just dropped and selling the stocks that just increased?

The answer to this question is almost surely no. Suppose it were the case that there was reversion to the mean in stock prices. A clever hedge fund would realize this, follow the strategy described above, and make a lot of money. But with enough investors following this strategy, the reversion to the mean would go away, because the prices of the low stocks would increase and the prices of the high stocks would decrease in response to the market's buying and selling decisions. The efficient-market hypothesis, discussed briefly in Chapter 7, says that with enough traders looking for these kinds of opportunities, we shouldn't be able to predict changes in stock prices, and therefore, we shouldn't expect reversion to the mean.

Reversion to the mean is quite prevalent in the business world. We see it for corporate revenues and profits, despite the desire by startups and venture capitalists to project future revenues by making linear or sometimes exponential projections from past revenues. So

why don't we see reversion to the mean for stock prices? The main reason is that stock prices already reflect beliefs about the future, while revenues don't. The price of the stock is driven by investors' beliefs in the long-term value of the firm. And if there were reversion to the mean, it would suggest that investors are bad at forming their beliefs. If there were a stock that we could expect to increase in price, investors would all buy it, driving the price up, erasing our expectation of a change.

The stock market is just one example of a general phenomenon. There should be no reversion to the mean when it comes to beliefs. It wouldn't make sense to say something like "Today, I believe that Republicans have a 60% chance of winning the House in the next election, but come Election Day, I expect my belief to be lower." That doesn't make sense because your belief is just your belief about the future—nothing else. If you expect that your belief will be 40% on Election Day, then your belief should be 40% today.

Key Words

- **Hawthorne effect:** The phenomenon whereby subjects change their behavior because they know they are being studied.
- **Demand effect:** A specific instance of a Hawthorne effect in which research subjects change their behavior to try to please the researcher.
- **Signal:** The systematic component of an outcome that is persistent across observations.
- **Noise:** Random components of an outcome that change from observation to observation.
- **Reversion to the mean:** The phenomenon whereby, if one observation of an outcome made up of signal and noise is particularly large (respectively, small) other

observations will typically be smaller (respectively, larger).

Exercises

1. Early on in every baseball season, someone appears to be on pace to break the home-run record, but they almost never do it. Let's think about why.

Suppose you hit a phenomenal number of home runs in the first 20 games of the season, and you're on pace to break the record.

- (a) In the next 20 games, are you more likely to hit an above average or below average number of home runs? Why?
- (b) In the next 20 games, are you likely to hit fewer home runs, the same number of home runs, or more home runs than in the first 20 games? Why?
- (c) A commentator notices that players on pace after the first 20 games almost never breaks the record. The commentator argues that this shows that players lose their nerve when they start thinking about setting the record. What data might you want to collect, and how might you want to analyze it in order to see if this interpretation is right?

Notes

The study comparing real knee surgeries to sham surgeries is:

J. Bruce Moseley, Kimberly O'Malley, Nancy J. Petersen, Terri J. Menke, Baruch A. Brody, David H. Kuykendall, John C. Hollingsworth, Carol M. Ashton, and Nelda P. Wray. 2002. "A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee." *New England Journal of Medicine* 347(2):81–88.

The article showing that, among asthma patients, the placebo effect appears to exist for subjective measures but not for objective measures is:

Michael E. Wechsler, John M. Kelley, Ingrid O.E. Boyd, Stefanie Dutile, Gautham Marigowda, Irving Kirsch, Elliot Israel, and Ted J. Kaptchuk. 2011. “Active Albuterol or Placebo, Sham Acupuncture, or No Intervention in Asthma.” *New England Journal of Medicine* 365(2):119–126.

Figure 8.4 is taken from

John P.A. Ioannidis, Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. “Replication Validity of Genetic Association Tests.” *Nature Genetics* 29:306–309.

Part III

Is the Relationship Causal

Chapter 9

Why Correlation Doesn't Imply Causation

What You'll Learn

- Correlation does not necessarily imply causation, and vice versa.
- Experts across domains frequently mistake correlation for causation.
- There are two key reasons why an observed correlation might be a biased estimate of a causal relationship: confounders and reverse causation.
- If we think clearly, we can sometimes sign this bias.
- There is an important distinction between confounders and mechanisms.

As we discussed in Chapters 2 and 3, information about correlations and information about causal relationships are useful for different purposes. Knowledge of correlations, on its own, can help us describe the world and forecast the presence of certain features of the world based on the presence of other features of the world. Knowledge of a causal relationship is particularly valuable for decision making because it can tell us how the actions we take affect the world. Remember our definition of causation from Chapter 3.

A causal effect is a change in some feature of the world that would result from a change to some other feature of the world.

So when we say that some action has a causal effect on some outcome, we're asserting that the outcome would be different in a counterfactual world in which the action was different. Knowing the effects of our actions allows us to anticipate and weigh their costs and benefits.

From a pragmatic perspective, this distinction is why the maxim "correlation doesn't imply causation" is so important. If we mistakenly take knowledge of a "mere" correlation as implying knowledge of a causal relationship, we might end up making big mistakes, taking actions because of a misguided belief about how those actions will affect outcomes we care about. To see what we mean, let's talk through an example where high-stakes decisions are being made about how to deploy resources, and where we can disentangle correlation from causation with some confidence. The example concerns the topic of charter schools in the United States.

In his heart-rending movie, *Waiting for Superman*, David Guggenheim tells the story of several young children from poor families. In each case, a child is enrolled in a substandard public school. And, in each case, the parents are working hard to get their child into a charter school (or, in some cases, a magnet school).

Charter schools are operated at public expense but independently of the public school

system. Some charter schools are run by not-for-profit organizations and others by for-profit corporations. The idea behind the charter-school movement is to encourage innovation and choice. Charter schools are free from some of the constraints (e.g., union contracts, legacy curricula) that public schools face. Hence, the argument goes, they can innovate in curriculum, teacher incentives, and the like in ways that regular public schools cannot. And, because they have to compete for students, they will be motivated to come up with new and better approaches to education. Whether charter schools in fact succeed in improving educational outcomes is a matter of heated debate.

In many areas with poorly performing public schools, more kids want to attend charter schools than can be accommodated. By law, when a charter school is oversubscribed it must admit students by random lottery. Families apply to the school, and after that, luck determines which kids get the coveted spots.

During the course of the movie we are told a slew of facts about the performance of the charter schools to which the students are seeking admission. Compared to public schools with socio-economically similar populations of students, these charter schools have better test scores, higher graduation rates, less crime, and so on. Indeed, the charter schools featured in the movie perform much better than public schools with respect to virtually every measurable outcome.

As the movie powerfully illustrates, the odds are stacked against the children. Some of the charter schools have hundreds of applicants for a couple dozen spots. So the odds of winning the lottery are exceptionally bad. As the movie ends, we discover that few of these students were admitted to the school of their choice. Instead, they will be enrolled in “failure factories” where, we are left to believe, in all likelihood, their potential will be wasted.

But is this the right inference? Does getting into a charter school really have a big impact on a child’s future? There is more at stake here than our feelings about the kids

in the movie. Over the past several decades, charter schools have emerged as one of the dominant approaches to school reform in the United States. Expansion of charter schools as an alternative to traditional public schools has bipartisan support—it was, for instance, pushed aggressively by both the Bush and Obama administrations. The share of public school students attending charter schools has risen from below 1 percent in 1999 to over 6 percent today. And critics raise concerns about the possibility that this expansion has led to a decline in resources available to traditional public schools; perhaps harming the students enrolled in those schools. So we'd really like to know whether charter schools have a positive effect on academic outcomes for their students.

Here's what we know. There is definitely a correlation between charter school attendance and academic performance. Within a city, on average, low-income students who go to charter schools have better educational outcomes than low-income students who go to traditional public schools.

As an example, consider the Preuss School, a charter school created by the University of California at San Diego. The Preuss school serves low-income middle and high school from all over San Diego County. By all accounts, it is a remarkable school, sending almost 100% of its students, who come from families with no history of college attendance, on to higher education. The school has garnered praise from many sources, including being named the “top transformative high school” in the country by *Newsweek* magazine multiple times.

And, indeed, it is certainly the case that students at the Preuss School perform much better than their peers at the San Diego public schools. For instance, look at the data in Figure 9.1 showing the difference in the percentage of students who pass standardized tests in Math and English at Preuss versus the San Diego City Schools (SDCS).

Just like in the stories from *Waiting for Superman*, from these data, it sure looks like the Preuss School is having a huge impact on the academic performance of its students.

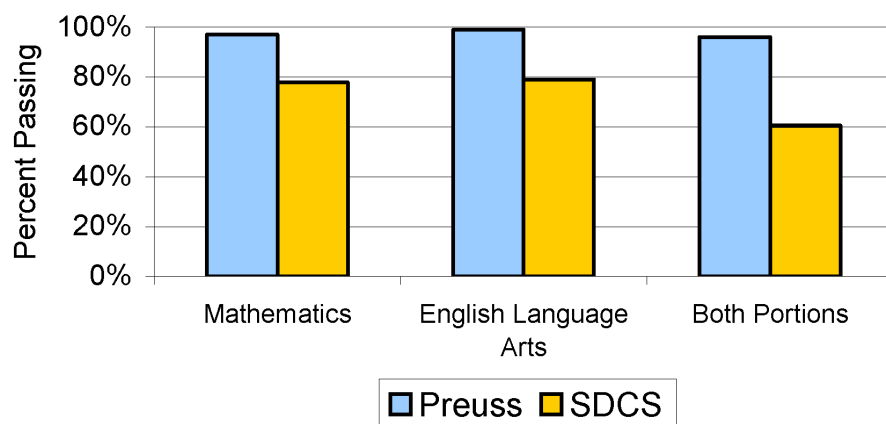


Figure 9.1: Standardized test scores at Preuss School and San Diego City Schools. (source: McClure et al.)

But before we jump to conclusions, let's think about this a little more clearly. These data show a positive correlation between going to Preuss and academic performance. But does the fact that students going to Preuss (or other charter schools) perform better academically than students at public schools imply that going to a charter school is *causing* them to perform better? That is, in the counterfactual world in which some other kids go to Preuss and these kids go to the public schools, would those kids perform better academically and the current Preuss students perform worse? Does the correlation imply causation? That's what we need to know if we are going to figure out whether investing resources in charter schools is a good decision.

Of course, it could well be that charter schools are genuinely causing students to perform better, meaning that in the counterfactual world where the charter school students had attended regular public schools they would have performed worse and their replacements from the public school would have performed better. But another possible explanation, as *Waiting for Superman* so eloquently illustrates, is that the students who choose to attend these schools are themselves different from the average public school student in

important ways. That is, perhaps the explanation for the correlation isn't a better school, but better students. If that is the case, then, in the counterfactual world in which these better students had not gone to a charter school, can we be so sure that they wouldn't still have outperformed their peers?

Ask yourself, under what circumstances are economically disadvantaged parents likely to sign up their child for the lottery to get into a charter school? Two circumstances occur to us. First, if the parents believe their child is gifted, they might be particularly motivated to get their child into a school with a good reputation. Second, if the parents themselves are particularly invested in their child's education, they might be more likely to do the work necessary to get that child into the lottery.

Natural talent and parental involvement are themselves pretty important determinants of student achievement. Suppose that the pool of students in the charter school lotteries (and, hence, at those schools) are, on average, more talented and come from families more interested in education than the population at large. Then, even if the charter schools themselves have no effect on the performance of their students, those students would nonetheless outperform the general population simply by virtue of their greater ability and more supportive family environment. Put differently, if all children went to the exact same school, the children who are currently at the charter schools would still be above average because they are more talented and have more dedicated parents

Remember the question we care about: can sending a child to a charter school be expected to improve the performance of that child relative to what he or she would have achieved at their local public school? The discussion above shows that we can't know the answer to this crucial question by comparing the performance of students currently in charter schools versus traditional public schools. We'd be comparing a group of particularly talented, ambitious kids from highly dedicated families to the general population. How would we know whether differences between these groups arose because of the effects of

charter schools, because of baseline differences in academic performance between groups, or both? In simpler terms, we'd be comparing apples to oranges.

To determine whether the charter schools are actually a cause of their students' excellent outcomes, we need to make a comparison that comes closer to capturing the counterfactual nature of causality. To do this, we need a way to compare apples to apples. The ideal question we'd like to answer is something like this: if everything else about two children were identical, would the child who went to the charter school do better than the child who went elsewhere? We obviously can't answer this question. But we can get closer by trying to answer something like this: if everything else about two groups of children was identical, would the kids who went to charter schools do better on average than kids who went to public schools

To take a shot at this latter question, we have to move beyond just comparing charter-school kids to all other public-school kids. We do so by narrowing our focus to just the children who tried to get into charter schools. All of those children were promising enough or had families dedicated enough to apply to a charter school. But, as a result of the lottery, some lucky students get into the charter school and others do not. Since the lottery was random, the pool of lottery winners and the pool of lottery losers should have the same characteristics, on average (i.e., if we ran the lottery over and over again). So we can learn a lot more about the actual effect of attending a charter school by comparing the academic performance of those who won the admissions lottery to those who entered but lost the lottery. If the positive correlation between charter school attendance and academic performance is still there in this narrower comparison, we will feel much more secure in giving it a causal interpretation because now we're comparing apples to apples.

This comparison has been done for many charter schools. Let's start by looking at what happens when you make that comparison for the Preuss School. We don't have data on this comparison for the same standardized test as above, but we do for another important

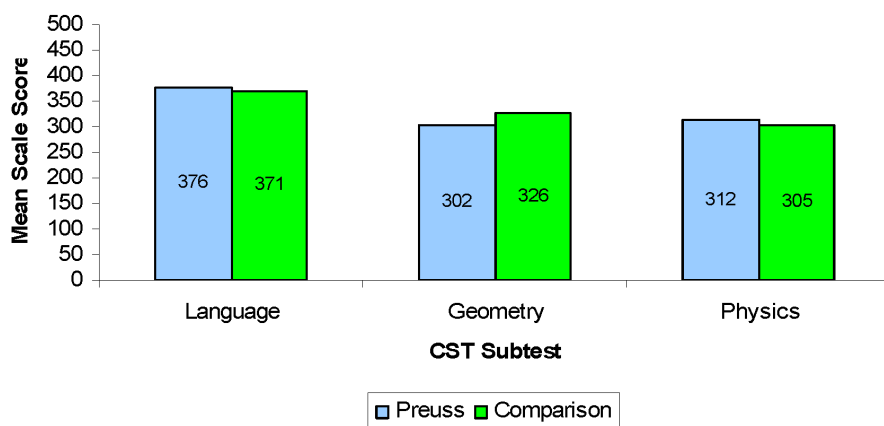


Figure 9.2: Comparing standardized test scores of kids who did and did not win the Preuss School admission lottery. (source: McClure et al.)

standardized test. It is shown in Figure 9.2.

This comparison shows the importance of comparing apples to apples. Yes, on average the students at the Preuss School outperform the students in the San Diego City Schools as a whole. But when lottery-winning students are compared to lottery-losing students, the correlation disappears—there is no performance difference.

A similar finding emerges in similar studies of many other charter schools. To be sure, different studies find different things. Researchers in Boston find that kids admitted to charter schools run by the well-known Knowledge is Power Program do better than kids who applied but lost the lottery. But the following, from another study of school choice programs in San Diego, is more typical of the literature:¹

In the vast majority of cases, we found no evidence that winners and losers of a given lottery fared differently in these achievement tests one to three years after the admissions lottery was conducted. We interpret this to mean that winning a lottery neither helps nor hurts achievement growth. (page xii)

¹Julian R. Betts. Does School Choice Work?: Effects on Student Integration and Achievement

Think about what this means. When we make an apples-to-apples comparison, the high-performing charter schools appear to have little to no causal effect on student performance. Most of the perceived good effects of these schools come from the fact that the children who enter the charter school lotteries are already academically different from the average student. Those students would have done better than average anyway. Separating correlation from causation in this way may change your views about how we should spend education resources.

The Basic Problem

When can a correlation between two variables be plausibly interpreted as evidence for a causal relationship? We just saw an example of how we might mistakenly think a correlation indicates causation and how it can really matter for decision making. But let's try to be a bit more systematic about why correlations aren't always evidence for causation so we can think more clearly about when we do and don't have evidence of a causal relationship.

Remember, we define causal relationships counterfactually. In Chapter 3, we introduced the notion of potential outcomes, which helps us think more clearly about such counterfactuals.

Let's suppose we're trying to estimate the effect of going to a charter school on academic performance, as measured by standardized test scores. So the *outcome* of interest is standardized test scores and the *treatment* of interest is attending the charter school.

Represent the outcome, standardized test scores, with Y . And represent the treatment, going to the charter school, with a binary variable T . If $T = 1$ for some individual, that means she attended the charter school. If $T = 0$ for some individual, that means she attended a public school. We sometimes say that a unit with $T = 1$ is *treated* and a unit with $T = 0$ is *untreated*, although it's often arbitrary which groups are labeled treated

versus untreated (e.g., we could similarly talk about the effect of attending traditional public schools versus charter schools).

In a metaphysical sense, for each individual there is some standardized test score that she would have gotten had she gone to the charter school and some standardized test score that she would have gotten had she not gone to the charter school. However, we only ever get to observe one of these. Nonetheless, having notation for each will help us think clearly about counterfactuals:

$$Y_{1i} = \text{outcome for unit } i \text{ if } T = 1$$

$$Y_{0i} = \text{outcome for unit } i \text{ if } T = 0$$

Using this notation, the effect of going to the charter school on person i 's test scores is:

$$\text{Effect of Charter School on } i\text{'s Test Scores} = Y_{1i} - Y_{0i}.$$

We say that causality is about counterfactual worlds because we can only observe, at most, one of the two quantities— Y_{1i} and Y_{0i} —for any individual at a particular point in time. This means that we can't directly measure the effect of going to the charter school on an individual.

Matters are better when we think about the average effect of going to a charter school across a bunch of individuals in some population of interest. For whatever population we are interested in, let's define notation for the average test score *if everyone went to the charter school* and the average test score *if everyone went to the public school* as follows:

$$\bar{Y}_1 = \text{average outcome if all units had } T = 1$$

$\bar{Y}_0$ = average outcome if all units had $T = 0$.

With this notation, we can now write:

$$\text{Average Effect of Charter Schools on Test Scores (ATE)} = \bar{Y}_1 - \bar{Y}_0.$$

For convenience, we'll often refer to the average treatment effect as the ATE. Of course, we can't directly observe this average effect any more than we can observe the effect of charter schools on an individual. We never see all units both treated *and* untreated. Indeed, at any given time, each unit is only one or the other. But we can try to estimate this average effect.

A first thing we might do to try to estimate the average treatment effect is simply compare the average test scores among students who go to the charter school (treated) to the average test scores among students who go to the public schools (untreated).

Start by thinking of our population as divided into two groups, those who went to charter schools (T) and those who went to public schools (U). Denote the average test scores in each of these groups by

$$\bar{Y}_{1T} = \text{average outcome among units with } T = 1$$

$$\bar{Y}_{0U} = \text{average outcome among units with } T = 0.$$

We will refer to the difference in the average test scores between these two groups as the *population difference in means*. It is just:

$$\text{Population Difference in Means} = \bar{Y}_{1T} - \bar{Y}_{0U}.$$

Of course, we might not observe the whole population, we might observe just a sample.

For instance, perhaps we only observe the students from one particular charter school. So the difference in average test scores that we observe in our sample is equal to the difference in average test scores among students who go to charter and public schools in the whole population plus some noise. So, we have:

$$\text{Sample Difference in Means} = \frac{\overline{Y}_{1T} - \overline{Y}_{0U}}{\text{Population Difference in Means}} + \text{Noise},$$

which is just a measure of the correlation between standardized test scores and attending a charter school in our sample.

Of course, we want to know the average *effect* of going to the charter school, not just the correlation. To start thinking about the difference between these, it helps to introduce two more concepts—the *average treatment effect on the treated* (ATT) and the *average treatment effect on the untreated* (ATU). The ATT is the average effect of going to the charter school among those students who in fact went to the charter school. That is,

$$\text{Average Effect of Charter Schools on Test Scores Among Charter School Students (ATT)} = \overline{Y}_{1T} - \overline{Y}_{0T}.$$

And the ATU is the average effect of going to the charter school, among those students who in fact went to the public schools. That is,

$$\text{Average Effect of Charter Schools on Test Scores Among Public School Students (ATU)} = \overline{Y}_{1U} - \overline{Y}_{0U}.$$

Notice two things. First, the ATE is just a weighted average of the ATT and the ATU, where the weights depend on how many kids are in each group. Second, just like the ATE, the ATT and ATU are both fundamentally unobservable. We don't observe the test scores students who go to the charter school would have made had they gone to the public schools (Y_{0T}). And we don't observe the test scores students who go the public schools would have

made had they gone to the charter school (Y_{1U}).

Ok, now that we have all that notation, we should be able to think clearly about the difference between correlation and causation. To start doing so, let's compare the difference in means that we in fact observe to the ATT—the effect of going to charter schools among students who went to charter schools. This will help us build some intuition that we will then also be able to apply to thinking about the comparison of the difference in means to the ATU and, ultimately, the ATE.

We are going to want to get back to with our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}.$$

That is, we want to find a way to write;

$$\text{Sample Difference in Means} = ATT + \text{Bias} + \text{Noise}.$$

How do we do this?

Let's start by remembering, from above, that

$$\begin{array}{l} \text{Sample Difference in Means} = \underbrace{\bar{Y}_{1T} - \bar{Y}_{0U}}_{\text{Population Difference in Means}} + \text{Noise}. \end{array}$$

Now, we are going to cleverly rewrite the population difference in means by adding and subtracting $\bar{Y}_{0T}$ from it. We know that seems weird. But, trust us, it's going to help. And, for now, it should at least be clear that we aren't doing any harm since, by adding and subtracting the same term, we are really just adding zero. Anyway, when we do that, we get:

$$\text{Sample Difference in Means} = \underbrace{\bar{Y}_{1T} - \bar{Y}_{0T}}_{\text{ATT}} + \underbrace{\bar{Y}_{0T} - \bar{Y}_{0U}}_{\text{Bias}_{ATT}} + \text{Noise},$$

where we've subscripted Bias with ATT to indicate that this is the bias we get when using the difference in means to estimate the ATT.

Ok, our algebraic trick was actually pretty cool, right? By adding and subtracting the same term, we were able to write things in terms of our favorite equation. The sample difference in means is equal to the ATT (estimand) plus a bias term plus noise!

But what exactly does that bias term say? If we are trying to estimate the effect of going to a charter school, our comparison of test scores among students who did and did not go to the charter school is biased if we expect that those two groups of students would have made different average scores on their standardized tests even in the counterfactual world where they all went to the public schools. When this is the case, we say the two groups have *baseline differences*.

So far, we've seen how the difference in mean test scores between charter public school students might be a biased estimate of the true effect of charter school attendance on those students who attend charter schools (the ATT). We could do a similar analysis for the true effect of charter school attendance on those students who attend public schools (the ATU):

$$\text{Sample Difference in Means} = \overbrace{\bar{Y}_{1U} - \bar{Y}_{0U}}^{\text{ATU}} + \overbrace{\bar{Y}_{1T} - \bar{Y}_{1U}}^{\text{Bias}_{ATU}} + \text{Noise},$$

Here we find a similar bias, but now the baseline differences we are worried about have to do with differential outcomes between the treated and untreated groups if both groups were to receive treatment. Since the overall average treatment effect (ATE) is itself just an weighted average of the ATT and the ATU, the bias associated with using the difference in means to estimate the ATE comes from both of these kinds of baseline differences.

Okay, now think back to the difference in academic performance between students at the Preuss School (treated) and the students at the San Diego City Schools (untreated).

We were concerned that the relationship might not be causal because, say, the students at Preuss were more academically talented, on average, than the students at the San Diego City Schools. If the Preuss students are in fact more academically talented, then there are baseline differences between the two groups of students—a difference in academic performance would exist between the treated and untreated students even if all students in both groups attended the same school. Because this comparison is so clearly not apples-to-apples, we can't be certain that the difference in average performance between the two groups is evidence of an effect of the Preuss School.

The lottery was convincing evidence precisely because it randomized people into treated and untreated. Randomization guarantees that, on average, the two groups are the same with respect to potential outcomes. That is, if we ran the randomization over and over again, on average the two groups would have the same baseline outcomes. (Of course, for any one run of the randomization, there could still be non-causal differences in academic performance between the two groups, just because of sampling variation or other kinds of noise.) Hence, differences in average outcomes between those two groups are evidence of a causal effect of the school.

When talking about causality, we often use language that evokes experiments, as we have here by discussing “treatment”. We do so because experiments provide a clear way to think about inferring causality from a correlation. If there is experimental randomization into treatment, then there are no systematic baseline differences between the treated and untreated groups.

Importantly, though, in many circumstances where we are interested in causality, we don't actually get to run an experiment. Instead, some people get the treatment and others do not, for reasons that we are not in charge of. In those circumstances, we have to be very careful about interpreting a correlation between outcome and treatment status as evidence of a causal relationship. As we saw, the positive correlation between test scores

(outcome) and going to the Preuss school (treatment) in the overall population was not in fact indicative of a causal relationship. The reason is because, out there in the world, social processes gave rise to baseline differences between the treated and untreated groups.

Sources of Bias

To take the proper care in interpreting correlations we need to be able to think clearly about when there will be systematic baseline differences because it is these baseline differences that give rise to bias. Fortunately, there are basically two sources of such differences: *confounders* and *reverse causality*. Understanding these is a big step towards being able to think about when you can and cannot learn something credible about causality from a correlation.

Confounders

A *confounder* is a feature of the world that satisfies two conditions:

1. It has an effect on treatment status.
2. It has effects on the outcome over-and-above the effect it has through its effect on treatment status.

Confounders create baseline differences and, thus, bias. Suppose some feature of the world makes people more likely to receive treatment. And suppose it also makes people more likely to have a particular outcome. Then, because of the confounder, there will be a correlation between that outcome and treatment, for reasons separate from any actual effect of the treatment. Hence, if there are such confounders (and we haven't done anything to account for them, which we will discuss in coming chapters), then it is a mistake to interpret a correlation between outcomes and treatment status as evidence of a causal effect of the treatment.

To be a little more concrete, remember our concern that the correlation (before the lottery) between going to the Preuss School and academic achievement was not causal. That concern was about baseline differences that resulted from more academically gifted kids being more likely to seek out the Preuss School. Another way of expressing that same concern is that the underlying academic talent of a kid is a confounder. Academic talent has an effect on treatment status—kids who are more academically talented are more likely to be in the treatment group (i.e., go to Preuss). And academic talent has an effect on outcomes over-and-above the effect it has through its effect on treatment status—more academically gifted kids are going to do better on tests for reasons over-and-above the fact that they seek out better schools.

Consider one further example. Many studies show a strong negative correlation between a country's per capita income and whether it experiences civil war. There are reasons to think that there could be a causal relationship underlying that correlation—for instance, perhaps when the economy is doing better, people have better lives and, thus, are less likely to be willing to mobilize to fight. But before interpreting the correlation as causal, one needs to think about whether there are potential confounders. One confounder you might worry about has to do with politics. A democratic political system, by incentivizing the government to adopt better public policy, is likely to positively affect a country's per capita income. Moreover, by giving people a non-violent means to express various grievances, democracy may also directly reduce the risk of civil war. Hence, democracy is a potential confounder. And, so, we are not justified in interpreting the correlation between per capita income and civil war risk as an unbiased estimated of the causal relationship.

So the first step in assessing whether a correlation is evidence of a causal relationship is to ask yourself whether there are any confounders. The schematic picture in Figure 9.3 might help you remember to do so. The question the figure asks is, for any given treatment and outcome, are there any factors that you suspect belong in the confounders box? To fit

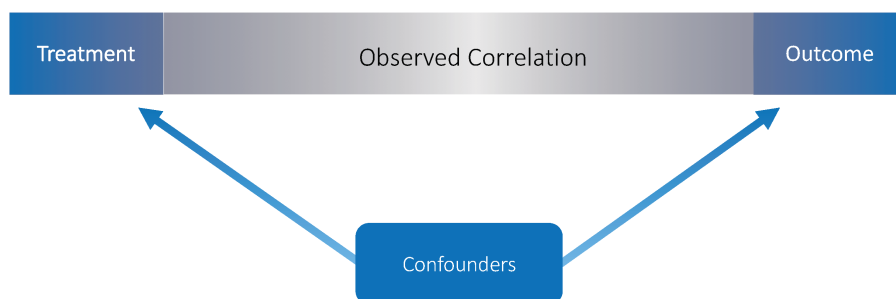


Figure 9.3: Confounders have an effect on treatment status and an independent effect on outcome.

in the box, two things must be true. The arrow from confounder to treatment has to make sense—that is, you must believe the confounder might exert an effect on the treatment. And the arrow from the confounder to outcome has to make sense—that is, you must believe the confounder might exert an effect on the outcome that doesn’t run through the treatment. If you can fill in factors that satisfy both these conditions, then you have a reasonable concern about confounders and it is a mistake to given a causal interpretation to the correlation between treatment status and outcome.

Reverse causality

The second source of bias we need to worry about is reverse causality. There is ***reverse causality*** if the outcome affects treatment status. Reverse causality creates baseline differences because, if an outcome effects whether or not a unit receives treatment, there will be systematic differences in outcomes between the treated and untreated groups that are not due to the effect of the treatment.

Consider, again, our example of the negative correlation between per capita income and civil war. We’ve already seen that there could be confounders underlying this relationship. But there might also be reverse causality. For instance, during the course of fighting a

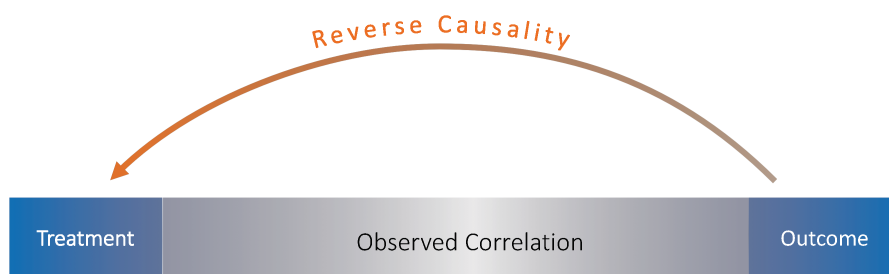


Figure 9.4: Reverse causality is when the outcome affects treatment status.

civil war, infrastructure is destroyed, production is disrupted, and people are killed. All of these effects of civil war directly reduce per capita income. Hence, the negative correlation between per capita income and civil war might reflect the effect of war on income, rather than the effect of income on war. The potential for such reverse causality is yet another reason that a causal interpretation of this correlation is not justified.

The schematic picture in Figure 9.4 is a way of helping to remind yourself to check for reverse causality before interpreting a correlation as causal. The question the figure asks is, for any given treatment and outcome, are there potential sources of reverse causality? That is, can we think of reasons that a causal arrow might run from the outcome to the treatment. If so, then you have a legitimate concern about reverse causality and it is a mistake to give a causal interpretation to the correlation between treatment and outcome.

In general, if someone shows you a correlation between an outcome of interest and a treatment of interest, without some additional information and investigation, you might have no way of knowing whether that correlation arose because the treatment affects the outcome, the outcome affects the treatment, confounders affect both the treatment and the outcome, or perhaps some combination of all of these possibilities.

Figure 9.5 provides an overall schematic for thinking about the two sources of bias. Now that we've got the conceptual material summarized in Figure 9.5 firmly in hand, let's get

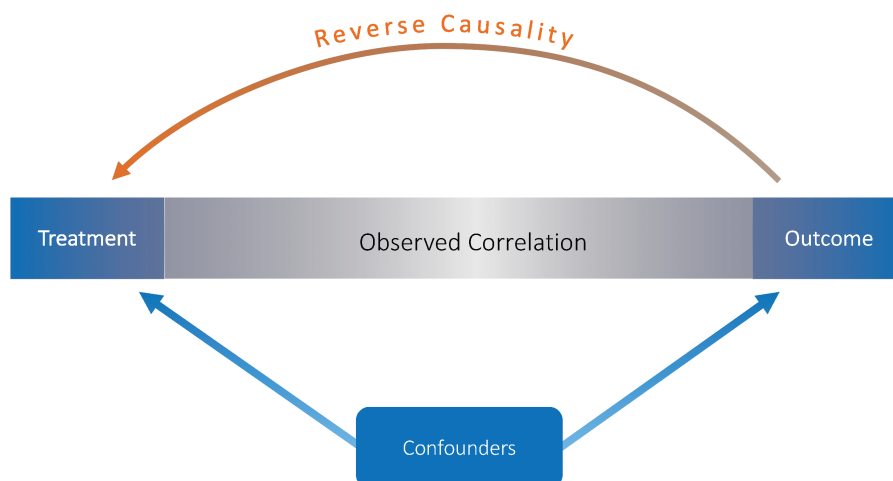


Figure 9.5: Confounders and reverse causality—two key sources of bias for estimating causal relationships.

some practice thinking clearly about correlation versus causation, confounders, and reverse causality by talking through a couple examples in some detail.

The 10,000 hour rule, revisited

You probably don't know the name Dan McGlaughlin, but he got quite a bit of press back in 2010. In April of that year, McGlaughlin quit his job as a photographer to pursue the dream of playing professional golf. He planned to practice golf for at least 30 hours every week for over six years, until he had put in 10,000 hours of deliberate practice. He believed that by the end of those 10,000 hours he would be an expert golfer, ready to qualify for the PGA Tour. It hasn't quite worked out as McGlaughlin planned. He didn't make the PGA Tour but did open what appears to be a pretty cool artisanal soda company.

McGlaughlin's quixotic plan should sound familiar. He took Malcolm Gladwell's "10,000 hour rule", which we discussed back in Chapter 4, to its (il)logical extreme. Talent, the argument goes, is secondary; great success is all about putting in those 10,000 hours. So

any of us can achieve virtually anything, even a career in professional sports, if we just commit ourselves to 10,000 hours of really serious practice.

As we've already discussed, Gladwell's evidence for the 10,000 hour rule—even just as a statement about a correlation—is dubious because of lack of variation. But McGlaughlin did not hang his hat entirely on this evidence. He was also inspired by research done by the psychologist K. Anders Ericsson of Florida State University.

Ericsson argues that the key to super high performance at anything is deliberate practice. Once a certain level of expertise has been achieved at some task, he claims, people's performance tends to plateau even if they keep gaining experience or general practice. The only way to continue improving at that point is through deliberate practice—working on exercises specifically targeted to particular aspects of performance. The more deliberate practice, the better performance. What distinguishes true masters of an activity from good, but not great, experts, is the amount of time devoted to deliberate practice.

The 10,000 hour idea derives from a seminal study of expert musicians by Ericsson and collaborators. Unlike Gladwell, Ericsson does have variation. He studied violin students at an elite music school in Berlin. All of the students in the study were expert violinists. But they could nonetheless be distinguished in terms of quality. Ericsson asked the faculty to identify three groups—the very best violinists who were likely to go on to careers as soloists or in major orchestras, good violinists who were less likely to have successful performance careers, and the weakest group of violinists, who would likely become teachers. (We are trying not to take offense.)

The violinists were interviewed about their history of practice—age at which they started, hours per week, the type of activities they engaged in during practice, level of concentration, and so on. They were also asked to keep diaries recording their practice habits. Armed with this information, the researchers were able to compare the practice behaviors of the different groups of violinists. The finding: the best violinists had practiced

at least 10,000 hours by the time they were 18, while the less accomplished violinists in the second group had only practiced about 7,500 hours, and the future teachers in the third group had practiced only about 5,000 hours. Moreover, the best violinists are distinguished by spending a greater share of their practice time deliberately. Similar studies report analogous findings for chess players, athletes, and others. There is a positive correlation between deliberate practice and high performance.

Based on this evidence, it looks like both the 10,000 hour rule and the focus on deliberate practice might not be so far fetched. The highest performing experts, in an array of fields, don't seem to be distinguished by measurable physical characteristics. The key appears to be that the best performers are those who practice the longest number of hours and in the most focused way. So maybe practicing deliberately for 10,000 hours really can make you world class. Maybe Dan McGlaughlin had a pretty good shot at becoming the next Tiger Woods.

But before you stop reading to go become a professional golfer, let's think a bit more. Ericsson didn't make Gladwell's mistake. He had variation and, so, established a correlation between deliberate practice and achievement. But that doesn't mean the correlation is evidence of causation. To reach that conclusion, we need to think about confounders and reverse causality.

Here's one possible concern. Suppose innate natural talent really is important. Imagine two kids, both of whom love to play the violin. One kid is more innately talented than the other. They both practice hard, putting in hundreds of hours a year. The hard work pays off and both progress rapidly. But over time the talent differential starts to manifest itself. The more talented kid masters difficult pieces of music more quickly and with greater precision. She receives more accolades and performance opportunities than the less talented kid.

Time progresses and the two kids become teenagers. New opportunities and distractions—

dating, sports, etc.—arise. Each teen has to decide how much time and energy to continue to devote to violin practice. The more talented of the two teens finds that every time she devotes a day to violin, she masters new skills and repertoire. This progress and achievement is inspiring. It creates a positive feedback loop whereby practice leads to success which inspires further practice and greater focus. So she continues to devote herself to deliberate violin practice, achieving those magic 10,000 hours by the time she is 18.

The less talented of the two teens also progresses each time he devotes a day to the violin, but he does so more slowly and with less proficiency. A piece that takes the more talented teen a week to master might take him a month. Even then, he plays it with less technical accuracy and musicality. His achievements are slower to come and met with fewer accolades. This lack of progress is frustrating. Met with less positive feedback, he finds practicing less rewarding. As a result, while he still loves and continues to work hard at violin, but as new opportunities come up, he is more likely to take a few hours or a day off of violin practice to pursue them. And even while practicing, perhaps he is less focused because he has other things to think about. By the time he is 18, he's put in only three-quarters as much practice time as his more talented friend, and less of it deliberate.

Two young people like those we've just described could easily have ended up music school classmates in Ericsson's study. The more talented of the two would have been identified by the faculty as one of the best violinists, while the less talented would have been identified as good but not great. Comparing them, Ericsson would have found, as he did, that the stronger of the two violinists put in 10,000 hours of deliberate practice, while the weaker put in only about 7,500 hours of less deliberate practice.

From this comparison, Ericsson wants to conclude that the correlation between deliberate practice and achievement is entirely causal. But as we've seen, this causal interpretation of the correlation is not warranted. As highlighted in Figure 9.6, innate talent could well be a confounder—affecting the amount of deliberate practice (treatment) and having a direct

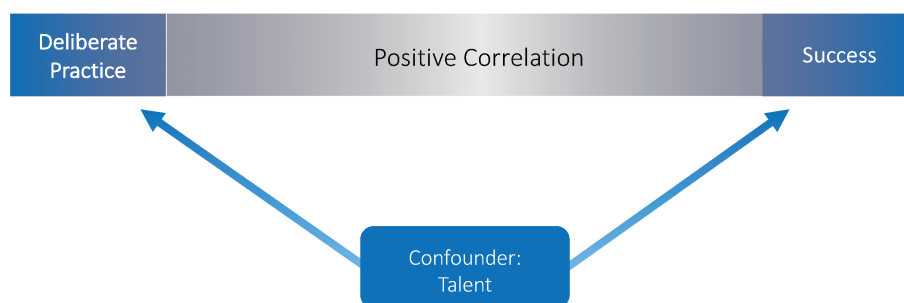


Figure 9.6: Is the correlation between deliberate practice and performance an unbiased estimate of the causal relationship?

effect on performance (outcome) over and above its effect through practice. Differences in talent cause baseline differences in achievement—if you had forced the more and less accomplished students to do the same number of hours of deliberate practice, they still wouldn't be equally good violinists.

The extent to which the correlation reflects a causal relationship is important here. To see why, think back to Dan McGlaughlin. Before quitting his job, McGlaughlin hadn't shown any of the signs of a world class golfer. That fact might help explain why he hadn't previously put in thousands of hours of deliberate practice. If at least part of the correlation between practice and success is due to both being affected by innate talent, then McGlaughlin's efforts were unlikely to yield the desired outcome. Profoundly talented people practice a lot and have great success. That doesn't mean that a person who lacks profound talent but forces himself to practice a lot will meet with that same success. And so, perhaps, he should have anticipated that his journey would end in an artisanal soda company, rather than a PGA Tour card. (For what it is worth, Ethan thinks running an artisanal soda company sounds *way* more fun than playing professional golf. Anthony, who is something of a scratch golfer, disagrees.)

Diet soda

At the time of this writing, there is a near consensus among nutrition experts that diet soda is bad for you. Studies by experts in respected scientific journals have linked diet soda consumption to a range of health problems including obesity, diabetes, and heart attacks.

Curiously, despite all of the purportedly hard evidence on the dangers of diet soda, scientists don't yet have a compelling explanation (aside from the adverse dental consequences of diet soda—acid is bad for your teeth). They have been racking their brains to explain why beverages with virtually no calories are somehow making people overweight.

Several theories have been put forward. One explanation is that diet soda has *chemicals*, which might be bad for us. Of course, everything we consume has chemicals, so this isn't much of an explanation, and it just kicks the can of soda one more step down the road. Another explanation is that diet soda confuses your body and makes it somehow want more calories. After consuming diet soda, the story might go, your brain expects you to receive some calories from this sweet beverage, and when it doesn't, it urges that you to raid your pantry for cookies and chips. In this way, your brain is like a child who is told they're about to get some candy only to have it revoked at the last minute. A third explanation is that diet soda (and presumably anything sweet) desensitizes your taste buds, meaning that you need to eat more and more sugary foods to get your fix.

None of these explanations sounds overwhelmingly compelling to us. Indeed, we could easily imagine the effect going in the opposite direction. Diet soda might allow someone with a sweet tooth to enjoy a refreshing treat without consuming extra calories. And diet soda might even trick your brain into thinking you've ingested calories and therefore speed up metabolism, which could be good for health and weight management. We're not experts, but it seems at least as plausible that diet soda is good for health, especially as a substitute for sugary beverages. So why have experts so strongly agreed that diet soda is bad for your

health?

We have scoured the studies, and the extent of the evidence seems to be the following. There is a negative correlation between drinking diet soda and health outcomes. People who drink diet soda are more likely to be obese, have diabetes, and suffer from a range of other health problems than are people who do not drink any kind of sweet beverages.

Before agreeing with the nutritionists that this correlation reflects a genuine causal effect of diet soda on health, we should think about whether there are confounders or reverse causality.

What if, for example, snacking makes people more likely to both drink soda (because the soda goes well with the snacks) and also more likely to be obese? Then snacking would be a confounder. Or perhaps it's reverse causality—what if obesity or diabetes makes people more likely to drink diet soda? Presumably, if you like soda and become diabetic, you'll switch to diet soda. Similarly, we ourselves could imagine switching from diet soda to sugary beverages if only we were healthier. Clearly, confounders and reverse causality are serious concerns, and we should not give the correlation between diet soda and health outcomes a causal interpretation.

How Different are Confounders and Reverse Causality?

While we are thinking about them, it is worth pausing to reflect on how confounders and reverse causality relate to one another. Often a problem that appears to be about reverse causality can also be thought of in terms of confounders, where the relevant confounder is simply the anticipated outcome.

To see what we mean, think back again to our example of the negative correlation between per capita income and civil war risk. We've seen that there are both confounders and reverse causality that invalidate a causal interpretation of this correlation. Consider

one more problem. Suppose that, for a variety of reasons (e.g., lack of democracy, ethnic divisions, nearby civil wars) people believe some country is at high risk for a civil war. This risk of civil war might deter investment in the country, lead to capital flight, cause a brain drain, and so on. In this way, anticipation of a future civil war can cause the country to have low per capita income. You could think of this as a case of reverse causality: civil war risk causes low income. But it may be more clarifying to think of it as a case of confounding, where the confounders are whatever factors lead people to believe the country is at high risk of civil war. Those factors both cause low income by deterring investment and causing brain drain. And, presumably, they lead people to believe the country is at high risk of civil war precisely because they exert an independent effect on civil war occurring.

Or, consider another example—campaign spending.

Campaign spending

Political candidates spend huge amounts of their time raising money for their campaigns. Members of Congress, for example, often spend several hours per day in a call center phoning wealthy constituents and asking them to help fund their next reelection effort. (It turns out that being in Congress isn't a particularly glamorous job.)

Of course, politicians do this because they believe that campaign dollars are essential for their electoral prospects. And a cottage industry of campaign consultants constantly reminds them how much they should be spending on television ads, digital ads, direct mail, and personal voter outreach. Electoral campaigns are clearly a big business predicated on the notion that candidates can improve their chances by raising and spending more money.

Given the scale of campaign spending, political scientists have devoted a lot of time trying to estimate the returns to these efforts. Can spending on advertising really influence election results? And are those effects big enough to justify the millions of dollars donated to

finance campaigns and the thousands upon thousands of hours spent raising those dollars?

One of the earliest and most influential studies of campaign spending was conducted by Gary Jacobson in 1978. Jacobson concludes that campaign spending seems to significantly help challengers' electoral prospects, but has little benefit for incumbents. Indeed, campaign spending by incumbents might even be counterproductive, hurting their electoral fortunes!

What is Jacobson's evidence for this claim that campaign spending helps challengers but not incumbents? Challenger spending is strongly positively correlated with challengers vote share. But incumbent spending is negatively correlated with incumbent vote share.

One explanation for these correlations, Jacobson speculates, is that incumbents typically raise and spend more money than challengers. Maybe some initial amount of spending at the levels that we typically see for challengers helps a candidate to obtain name recognition and persuade voters. But perhaps too much spending from an already well-known incumbent annoys and turns off potential supporters. On this account, incumbents are making systematic mistakes, both in spending their time raising money and in spending that money once they've raised it.

Of course, the comparisons underlying these correlations may not be apples to apples. We need to think about confounders and reverse causality.

One big concern along these lines has to do with electoral strength. Which kinds of challengers tend to be able to raise and spend lots of money? Presumably, popular challengers with a real shot at victory. It is those electorally strong challengers who donors are likely to be willing to invest in. But, of course, strong challengers are those who were expecting to do well in the election even before they raised the money—perhaps they are charismatic, well known, or particularly talented. So it would be a mistake to interpret the positive correlation between challenger spending and challenger electoral performance as purely causal. It, at least in part, reflects baseline differences in electoral strength between challengers

who can and can't raise a lot of money. And, in particular, because electoral strength has a positive effect on both fundraising and votes, we should expect this correlation to be an overestimate of the true effect.

The thing we want you to notice in this example is that you can think of the problem of electoral strength as one of reverse causality or as one of confounding. Thought of as reverse causality, you might describe it as follows: “when a challenger is going to do well, she can raise and spend more on her campaign.” Thought of as a confounder, you might describe it as follows: “when a challenger has characteristics that make her competitive, this affects both her ability to raise and spend money and how well she does in the election.” Both sentences describe the same concern, just framed slightly differently.

A similar argument holds for incumbents. In general, although they spend and raise a lot of money, most incumbents in U.S. elections are electorally pretty safe. The ones who really need to exert a lot of effort raising and spending money are those who are perceived as electorally vulnerable. So we might expect exactly the opposite relationship for incumbents as for challengers. Incumbents spend a lot of money not when they are strong, but when they are weak. So the observed lack of (or even negative) correlation between campaign spending and electoral performance for incumbents might be an underestimate of the true effect. And, again, you can view this problem in terms of reverse causality—“weak incumbents spend more money”—or in terms of confounders—“characteristics that weaken incumbents, making the race competitive, cause them to spend more money and lead to weaker than average electoral outcomes for incumbents.”

Subsequent studies using randomized experiments and other clever approaches to try to tease out the causal relationship generally suggest that campaign spending does have positive effects for both challengers *and* incumbents, although the substantive size of those estimated effects is typically small. A campaign might have to spend hundreds of dollars to swing a single vote, which means that meaningfully influencing the outcome of a large

election through campaigning donations is typically unaffordable. For example, consider a gubernatorial or senatorial race in a large U.S. state. Even in a race thought to be very close, the outcome will likely be decided by hundreds of thousands of votes. This means that if donors wanted to influence the outcome of the election, they would have to spend tens of millions of dollars and hope that their spending does not trigger an offsetting response from supporters of the opponent. Because of this, even the very largest donors have likely swung very few elections.

As you can see, there isn't a ton at stake as to whether we think about such cases as reverse causality or as confounders. What really matters is that we interrogate correlations for possible baseline differences, whether from confounder or reverse causality, and if there are, that we show proper caution before interpreting a correlation as implying causation.

Signing the Bias

When there are unaccounted for confounders or reverse causality, we cannot interpret the correlation between treatment and outcome as an unbiased estimate of the true causal relationship. But sometimes we can make some progress by asking whether the observed correlation over- or under-estimates the causal effect. Indeed, we just did this above, when we argued that the positive correlation between challenger spending and challenger electoral outcomes was likely an over-estimate of the true causal effect.

Let's think back to our favorite equation, this time written in terms of causal inference:

$$\text{Observed Correlation (Estimate)} = \text{True Causal Effect (Estimand)} + \text{Bias} + \text{Noise}.$$

Suppose the observed correlation between administering some medical treatment and survival rates following a stroke is positive. But also suppose there are confounders that you

have not accounted for, so there is bias in your estimate of the true causal effect. If you have reason to believe that the bias is positive, then the observed correlation is an *overestimate* of the true causal effect of the treatment. This means that you can't be confident, based on the observed positive correlation, that the treatment does anything at all. Even if the true causal effect is zero, you would observe a positive correlation on average due entirely to confounders creating positive bias.

But now suppose you have reason to believe that the bias is negative instead of positive. In this case, the observed correlation is an *underestimate* of the true causal effect of the treatment. So, if you are confident that the observed correlation is positive, you should be even more confident that the true causal effect is positive. And this can be useful to know. For instance, suppose administering the treatment would be a good idea (given its various costs) even if the true effect was equal to the observed correlation. Then the fact that the observed correlation is an underestimate of the true effect suggests that you should administer the treatment.

Note, of course, you could still end up being wrong because of noise. Even if you are under-*estimating* the true causal effect, that doesn't mean that any one estimate is in fact lower than the true effect. It just means that your estimates will be lower than the true effect on average.

Because this kind of thinking can sometimes be valuable, it is useful to spend a little time thinking conceptually about when confounders imply that the observed correlation overestimates versus underestimates the true effect.

To see how this works, return to our discussion of the positive correlation between attending a charter school and standardized test scores. There, we said, one possible confounder is that students who go to the trouble of applying to a charter school may on average be more academically gifted or motivated than the general student population. And the fact that those students are more academically gifted or motivated may have a direct effect

on their test scores.

If this story is right, does this confounder tend to make the correlation between charter school attendance and standardized scores an over- or under-estimate of the true effect? Let's think about it. Being academically gifted has a positive effect on the likelihood a student goes to a charter school. And it also has a positive effect on test scores. That means, part of the observed positive relationship between going to a charter school and test scores is the result of differential academic giftedness or motivation. Hence, this confounder is pushing the observed correlation to be an *overestimate* of the true effect. That is, the bias in our favorite equation is positive.

It is straightforward that the same would be true if we had a confounder that negatively affected both attending a charter school and test scores. Indeed, this is simply the same case, but with relabeling. If we think of the confounder as “lack of academic motivation” instead of “academic motivation”, then that confounder is negatively correlated with treatment and outcome but, still, obviously, leads the observed correlation to be an overestimated of the true effect. Thus, as illustrated in Figure 9.7, if you have a confounder that has the same sign effect on both treatment and outcome (whether negative or positive), then failing to account for this confounder will create positive bias. In such circumstances, the observed correlation will tend to be larger than the true effect.

Now let's think about a confounder that has differently signed effects on treatment and outcome. For instance, suppose students from particularly difficult neighborhoods are more motivated to apply to charter schools, but also are expected to do worse academically (all else equal) because of challenges in their living environment. This, again, is a confounder—it exerts an effect on both treatment (whether or not a student attends a charter school) and on the outcome (performance on standardized tests). But unlike the case of academic ability (which was positively associated with treatment and control), this confounder will create negative bias. Hence, the observed correlation between charter school attendance

and standardized test scores is an underestimate, rather than an overestimate, of the true effect.

Why is this the case? On our new story, living in a difficult neighborhood has a positive effect on the likelihood a student goes to a charter school. And it has a negative effect on test scores. That means, the observed relationship between going to a charter school and test scores reflects that the kids at charter schools over-represent difficult neighborhoods, relative to the full population. This fact tends to drag test scores down for reasons that having nothing to do with the effect of the charter school. If charter schools and public schools had the same proportion of kids from good and bad neighborhoods, the positive association between charter school attendance and test scores would be even more positive. Hence, this confounder is pushing the observed correlation to be an *underestimate* of the true effect.

It is again straightforward that the same would be true if we had a confounder that negatively affected the likelihood of attending a charter school and positively affected the outcome. Thus, as illustrated in Figure 9.7, if you have a confounder that has one sign effect on treatment and the opposite sign effect on outcome, this confounder creates negative bias. Hence, the observed correlation will be an underestimate of the true effect.

Signing the bias is even easier in the case of reverse causality. The outcome is, obviously, positively related to itself. So, if the outcome also has a positive effect on the treatment, the bias is positive. This means the observed correlation is an overestimate of the true causal effect. And if the outcome has a negative effect on the treatment, the bias is negative, so the observed correlation is an underestimate.

In addition to simply signing the bias, if we had a lot more information, we might be able to say something about the magnitude of the bias. Under some assumptions, the bias induced by a confounder is simply the effect of the confounder on the outcome multiplied by a measure of the correlation between the confounder and the treatment (measured by

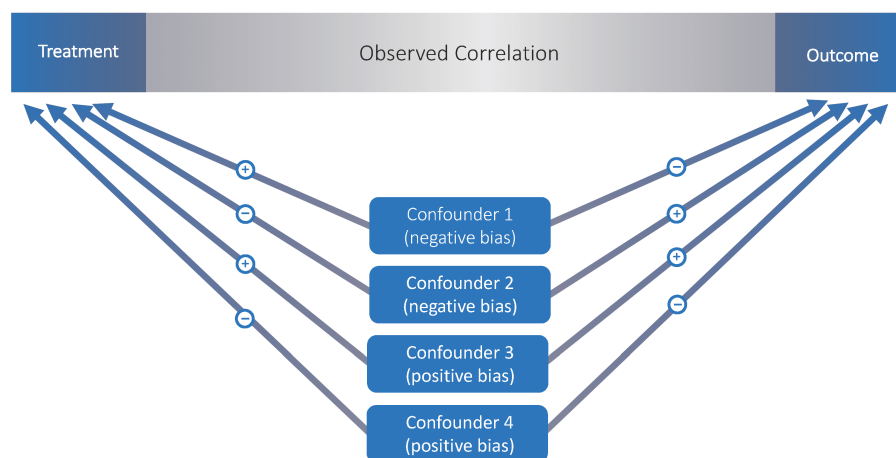


Figure 9.7: Signing the bias from a confounder.

the coefficient you would get from regressing the confounder on the treatment).

As we'll see in the next chapter, if we have data on this confounder, we can try to remove this bias by controlling. But if we don't have that data, one could still make some guesses about the extent to which the confounder affects the outcome and is correlated with the treatment in order to gauge the extent of the bias.

This discussion illustrates that we can still learn something about the quantity of interest from biased estimates. It's not as if we have to throw away all our analyses just because there might be confounders, and if we have good guesses about the direction and magnitude of the biases, then we might still be able to learn a lot about the quantity of interest. But often, it's difficult to know how much an observed correlation is the result of bias, which is why naive correlations are not our preferred approach for learning about causal relationships.

Another approach is to work in reverse. Instead of inferring how big an effect must be by making guesses about bias, one can assume that the true effect of interest is zero and ask how big the biases would have to be in order to explain an observed correlation. If the extent of that bias is implausibly large, then we would conclude that the effect of interest

is not zero. This kind of analysis is often referred to as a *sensitivity analysis*. We won't discuss the details in this book, but as a general rule of thumb, it's good to think about sources of bias, their likely signs, their likely magnitudes, and what that implies for your quantity of interest.

Let's use our new found skill of signing the bias to think through a more extended example.

Contraception and HIV

One of the greatest public health scourges of our time is the spread of HIV and AIDS in Africa. Researchers have worked hard to determine why these horrifying diseases are spreading so quickly and to try to stem the tide. One disturbing hypothesis that has received attention from scholars and public health officials alike is that the use of hormonal contraception by women may increase the risk of HIV transmission by inducing changes in the immune system or body tissue that increase virulence.

In a 2012 study in *The Lancet Infectious Diseases*, researchers presented evidence supporting this hypothesis. The researchers analyzed data on over three and half thousand couples from several countries in east and southern Africa. Each couple had one partner who was infected with HIV and one partner who was not. The researchers could observe a variety of self reported behaviors—e.g., condom use, other sexual partners—and also whether the woman received injectable hormonal contraception from the clinic that was conducting the study. The researchers also learned whether the non-infected partner contracted HIV over the course of a year or two. Finally, the researchers used genetic screening and statistical techniques to try to determine, in those couples where the partner did contract HIV, whether it was transmitted partner to partner or whether it was transmitted from some third source (e.g., another sexual partner).

There are two big findings. First, HIV-negative women who used hormonal contraception were twice as likely to acquire HIV from their infected male partners as were HIV-negative women who did not use hormonal contraception. Second, HIV-infected women who used hormonal contraception were twice as likely to transmit HIV to their HIV-negative male partners as were HIV-infected women who did not use hormonal contraception. These results held true controlling for self-reported condom use. (We'll talk more about what "controlling" means in the next chapter.) From these findings, the authors, the New York Times, National Public Radio and many other sources reported that hormonal contraception likely increases the risk of HIV transmission in Africa.

This study was a major improvement over existing studies on this critically important issue. But it was a long way from comparing apples to apples. What might be going wrong?

In a study like this, the biggest worry is the possibility of confounders—women who take hormonal contraception are different from women who don't in lots of unmeasured ways, some of which may also be relevant for HIV transmission risk. If this is the case, then the observed correlation between hormonal contraception use and HIV transmission may be a biased estimate of the true causal relationship.

The researchers who authored the Lancet study were not able to randomly assign some women to take hormonal contraception and other women not to. Women received hormonal contraception if they wanted it.

One concern is that women who intend to be more sexually active might also be more likely to use hormonal contraception. Sexual activity with an infected partner is, in and of itself, a risk factor for HIV transmission. So, independent of anything else, more sexually active women are at greater risk from HIV transmission. If the women who are taking hormonal contraception are systematically engaging in more sexual activity, they will have higher transmission rates, even if the contraceptives themselves are playing no direct biological role.

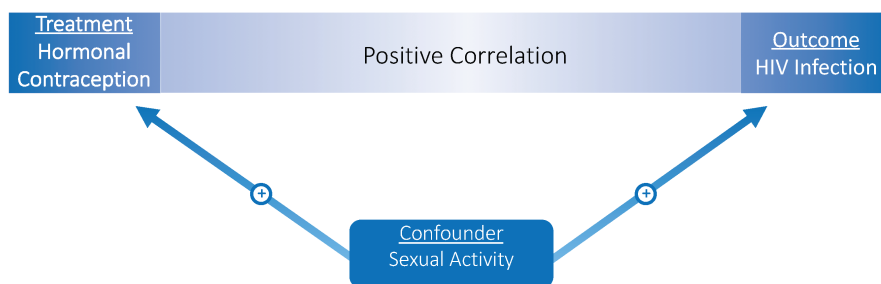


Figure 9.8: Underlying level of sexual activity tends to make the correlation between oral contraception use and HIV transmission an overestimate of the true causal relationship.

Which direction would this confounder bias the findings? As highlighted in Figure 9.8, the thought is that sexual activity is positively associated with using hormonal contraception and also positively associated with HIV transmission. So the bias is positive. As such, this confounder tends to make the observed correlation an over-estimate of the true causal relationship between hormonal contraception and HIV transmission.

The Lancet authors are aware of these types of concerns and make some attempts to address them. In particular women were asked about past sexual behavior and condom use, which the authors then control for. But self reported behavior is notoriously unreliable, especially for sensitive topics like sexual activity and condom use. Indeed, unlike in many other studies, in this study self-reported condom use was similar between the two groups of women—those using hormonal contraceptives and those not using hormonal contraceptives—which further raises concerns about whether the authors were able to effectively account for these potential confounders.

Mechanisms vs Confounders

It is easy to get a little confused about what is and what is not a confounder. One particularly common error is to mistake the mechanisms by which a treatment affects an outcome

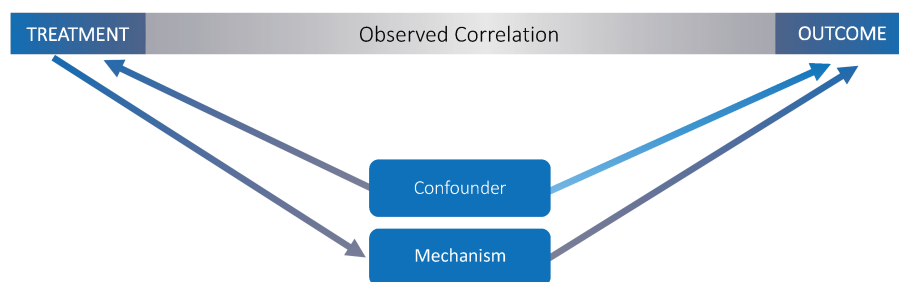


Figure 9.9: The distinction between confounders and mechanisms.

for confounders. A *mechanism* is some feature of the world that the treatment affects which then, in turn, affects the outcome. So a mechanism, rather than being a confounder, is part of the way that the treatment has its causal effect on the outcome.

For instance, one way that a charter school might cause students to get better test scores than they would if they went to their local public school is by providing more advanced placement classes that better prepare students for tests. When looking at a correlation that says charter school students perform better than public school students on standardized tests, it is tempting to say, “yeah, but that is confounded by the fact that those charter school students had access to more advanced placement classes.” But this isn’t right.

Remember, a confounder is not simply a feature of the world that is correlated with treatment and outcome (which, on this story, AP classes are). It is a feature of the world that *affects* both treatment and outcome. But, on our story, access to AP classes doesn’t affect whether a student goes to a charter school (treatment). Rather, it is affected by the student going to a charter school and then, in turn, affects the students performance on standardized tests. Thus, access to AP classes is not a confounder, it is one of the mechanisms by which charter schools improve test scores. Figure 9.9 illustrates the distinction (note the direction of the arrows).

As we say, it is easy to get confused about these issues. So, let’s talk through a couple

examples.

Suppose that a medical study of middle-aged men finds that those who take statins are less likely to die of heart attacks. You note that those men who take statins are on average wealthier and have lower cholesterol. Which of these is a confounder and which might be a mechanism? Think about it for a moment before we tell you the answer.

Let's start with wealth. Remember, when assessing whether some feature of the world is a potential confounder, you need to ask whether it could affect both treatment and outcome. So we ask two questions:

1. Could a person's wealth affect whether he takes statins? Surely the answer is yes. Wealthier men are, presumably, better able to afford medication and also probably more likely to see a doctor who would prescribe that medication to them.
2. Could a person's wealth affect her risk of dying from heart disease? Again, the answer is yes. Wealthier people might be better able to afford heart-healthy lifestyles (e.g., joining a gym) and are more likely to get swift access to health care in the event of a heart attack.

Thus, we should worry that wealth is a confounder here.

What about lower cholesterol? Medical evidence suggests that higher cholesterol might affect the likelihood of having a heart attack (although it's hard to tease out the causal effect). But does cholesterol affect whether or not a person takes statins? Here, we might need a little more information. In particular, when exactly were the cholesterol levels measured.

If cholesterol was measured before the person started taking statins, then it is a good candidate for a confounder. After all, people typically choose to take statins when they have high cholesterol. (Using your skills from the previous section on "signing the bias", does

this confounder make you think the study under- or over-estimates the efficacy of statins?)

But if cholesterol levels were measured after the person started taking statins, then it is a mechanism. We suspect that one of the ways that statins might reduce the risk of heart disease is by lowering cholesterol. If this is true, and we randomly assigned some people to take statins and others not to, we would expect the ones who took the statins to have lower cholesterol (and lower risk of heart disease). This difference in cholesterol levels isn't a problem for inferring the efficacy of statins; rather, it is a mechanism by which that efficacy is achieved.

Here's another example. Suppose, we are interested in whether a good economy helps reduce the risk of civil war. We find that there is indeed a negative correlation between per capita income and the frequency with which a country experiences civil war. But we also note that democracy is positively correlated with per capita income and negatively correlated with civil war risk. Should we think of democracy as a confounder or a mechanism in this case?

This is a bit of a tricky one. You can certainly see how democracy might be a confounder. Having a democratic form of government might improve the quality of governance. And good governance might cause a country's economy to grow. Moreover, being a democracy might give people non-violent ways to resolve political disputes, thereby directly reducing the risk of civil war. On this story, democracy is a confounder, since it has a direct causal effect on both treatment (per capita income) and outcome (civil war).

But you can also see how democracy might be a mechanism. Perhaps as countries become richer, citizens become more informed, better educated, more able to take actions for their own benefit, and so on. In this way, having a higher per capita income might directly increase the probability that a country becomes a democracy. And then, for the reasons already stated, democracy might decrease the risk of civil war. On this story, rather than being a confounder, democracy is part of the mechanism by which higher income

reduces civil war risk.

As this example highlights, the distinction between a confounder and a mechanism is important, but not always cut-and-dry. For now, it is important to see the distinction at a conceptual level, even if in many real-world scenarios you are not always sure whether some factor is the one or the other. We will return to this theme in the next chapter, when we talk about the benefits and limitations of *controlling*.

Bias and Noise

We'd like to pause to remind you not to forget the lessons from Part II about assessing whether a relationship exists just because we've now turned our attention to thinking about causal questions. In this spirit, let's think about how you should assess the evidence when someone shows you a correlation and interprets it as an estimate of some causal relationship.

The first question you might want to ask is whether they have actually established that a correlation exists. Recall from Chapter 4 that people often think they have measured a correlation when they haven't because they didn't collect data with variation in one of the key variables. So, for instance, you need to make sure that they didn't just look at instances where the outcome of interest occurred or where the purported treatment was always present. If they made this kind of mistake, you can't even know from the data presented whether the variables are correlated, let alone related causally.

Once you're sure they've established that a correlation exists in the data, you might then ask yourself whether you think the estimated correlation reflects a genuine relationship in the world. For example, suppose someone shows you that peanut butter consumption is correlated with appendicitis in a sample of 100 people—within that sample, people who eat more peanut butter were more likely to get appendicitis. You might ask yourself a series of questions. Is the correlation statistically distinguishable from no correlation? Why do they

only have data on 100 people? Did they collect the data with the goal of measuring this particular correlation? Would they have told you about this finding if they had found no correlation? If you're worried about p-hacking or p-screening, then you might be skeptical that there's actually a correlation between peanut butter and appendicitis in the broader population, and you'd want to collect an independent sample of data to see if the correlation persists in that new sample. If it doesn't, you should worry that the true estimand (the correlation in the population) is zero and that they found a positive correlation in their 100-person sample because of noise.

If you're convinced that the correlation is likely to reflect something real, then you'd want to assess whether this correlation is convincing evidence of a causal relationship. You'd want to ask whether they're comparing apples to apples—are there confounders or reverse causation that bias the estimated correlation away from the true causal relationship? For example, if someone shows you that ice cream consumption is correlated with sunburns across days of the year, you'd probably believe that they've identified a genuine correlation. If they collected a new sample of days, they'd probably continue to find a strong correlation between ice cream and sunburns. But that doesn't mean the correlation constitutes evidence that ice cream causes sunburns. It might. Maybe eating ice cream inspires people to go outside. But a far more likely explanation is that sunshine increase both ice cream consumption (for reasons unrelated to sunburns) and sunburns (for reasons unrelated to ice cream).

To help us put all of this together, let's return to the special case of our favorite equation when we are doing causal inference:

$$\text{Observed Correlation (Estimate)} = \text{True Causal Effect (Estimand)} + \text{Bias} + \text{Noise}.$$

There are two kinds of ways an estimated correlation can deviate from the causal effect

of interest. First, there could be noise. Noise here refers to idiosyncratic factors that affect our estimate. This could come from sampling variation in cases where you care about a population but you only have data on a sample. Or it could come from other idiosyncratic variation in your variables of interest that are independent of any kind of causal connection between your variables (e.g, measurement error). We might think that noise would, on average, be zero, so we can just ignore it. But that doesn't mean the noise in your particular case is zero. And furthermore, in the presence of p-hacking and p-screening, the average noise won't be zero. This was the focus of Chapter 7. Second, there could also be bias—i.e., confounders or reverse causation, which were the focus of this chapter.

When confronted with a correlation that is presented as evidence of causation, it helps to consider all three factors—a causal effect, bias, and noise—and try to think through the role each plays in explaining the correlation. Of course, it is often the case that an estimate reflects some combination of all of three.

In some cases, it's tricky to separate bias and noise or even to think about them in a conceptually clear way. Let's see some examples of this. Tyler Vigen's book *Spurious Correlations* identifies pairs of trends over time that happen to correspond with one another, even though there's no good reason to think that those two trends are causally or logically connected in any way. The term *spurious correlation* is certainly apt, although we personally avoid it because it doesn't typically clarify whether the person using the term thinks the correlation arose because of bias or noise.

Figure 9.10 illustrates one of Vigen's examples. It shows the correlation over time between suicides by hanging and government spending on science in the U.S. Although it's not presented in a conventional way, this figure shows a positive correlation. If you think of each year as a unit of observation, it's clear that years with more hangings than usual also have higher-than-average spending on science. In fact, the correlation coefficient is .9979, essentially the strongest correlation one can find.

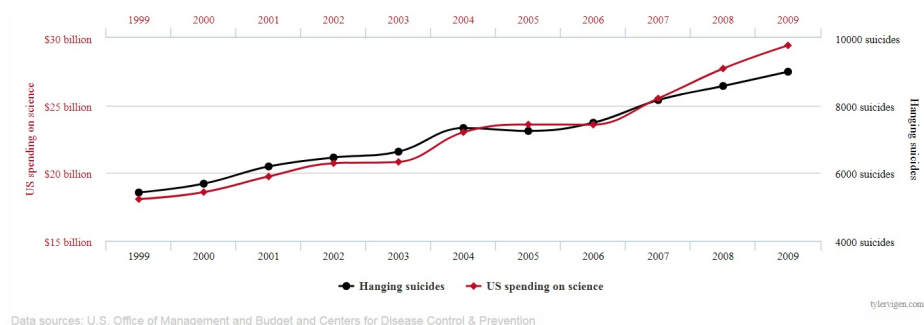


Figure 9.10: The strong correlation over time between suicides by hanging and government spending on science.

What's going on here? Is this correlation attributable to a true causal effect of science spending on suicides, to bias, or to noise? It's theoretically possible, but very unlikely, that science spending has a large, positive effect on suicide by hanging (or vice versa). Noise certainly seems like a plausible explanation. If you look at enough variables, you're bound to find two of them that happen to correspond by chance, and we know that this is exactly what Vigen did. He checked for correlations over time for many variables and selectively reported the correlations that were significant.

But maybe it's also bias. What's an example of a confounder here? How could there be a variable that affects both hangings and also science spending? One potential confounder is population. Over this period (1999-2009), the US population grew steadily from about 279 million to 307 million. And population growth could plausibly increase both suicides and science spending.

To explore whether bias or noise is the more important explanation, it might help to think about whether you expect this correlation to also hold for years before 1999 and after 2009. If you suspect that this correlation would likely hold more generally outside this sample of data, then it can't just be noise. Alternatively, if you think that this correlation is just a fluke, unlikely to hold outside of the short period for which Vigen collected data,

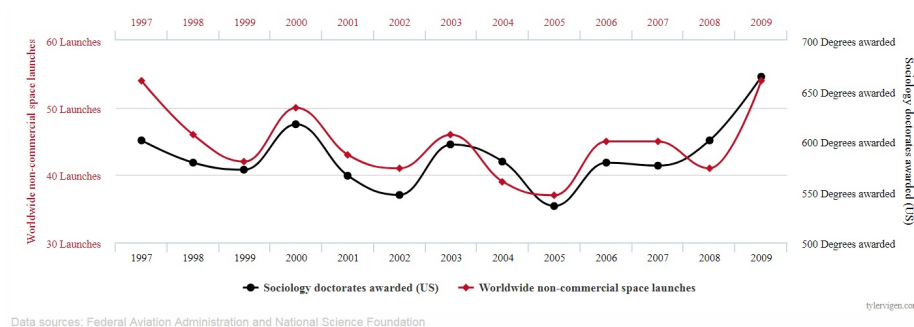


Figure 9.11: The quirky correlation over time between sociology doctorates and space launches.

then it's just noise, due to neither a causal relationship nor bias.

Let's take a look at another couple examples. Figure 9.11 shows the correlation over time between sociology doctorates awarded in the U.S. and worldwide non-commercial space launches. Again, there's a very strong correlation over time. Furthermore, it's not so easy to simply attribute the result to population growth (or something else changing over time) because the correlation is not simply driven by the two variables generally increasing over time. On average, space launches and sociology doctorates aren't increasing or decreasing much over time, but the years with more space launches also tend to be years with more sociology doctorates.

Here, we're pretty comfortable chalking this one up to noise or (bad) luck. There's idiosyncratic variation from year to year in space launches and sociology doctorates, and they happened to line up during this period. But we suspect that if we looked at the next 13 years of data, the correlation would be close to zero.

Finally, Figure 9.12 shows the correlation over time between the number of movies in which Nicolas Cage starred and swimming pool drownings. This one feels like another straightforward case of noise. There's surely no causal connection, and there's also probably no compelling confounder. And, as with sociology and space launches, we're willing to bet

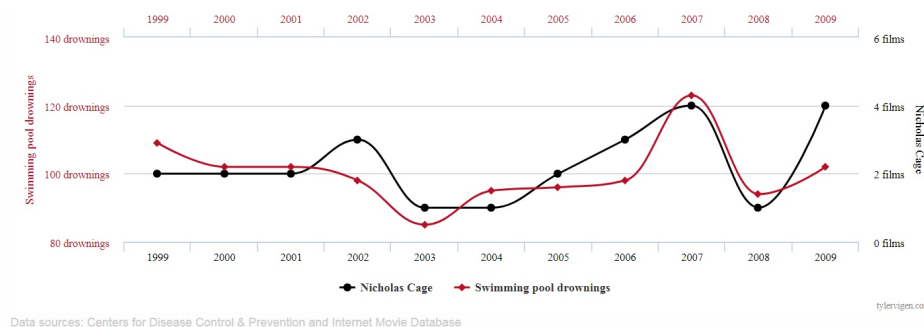


Figure 9.12: The metaphysically challenging correlation between Nicolas Cage movies and swimming pool drownings.

that this correlation won't continue to hold in future years.

However, the Nicolas Cage-drowning correlation poses a different conceptual conundrum. Suppose that this analysis included all of the years during which Nicolas Cage was acting and all of the years during which people had swimming pools (this is obviously not the case, but just imagine with us). If Nicolas Cage was no longer making movies and people no longer had swimming pools, we couldn't assess the correlation between these two variables in some future period. So how could we think about whether this correlation was the result of noise? Furthermore, what would it even mean to say that this correlation was the result of noise if we had all the data there was to have on Nicolas Cage movies and swimming pool drownings? If you have observed the entire population (here, of Nicolas Cage movies), there is no sampling variation.

One way to resolve this puzzle is to have an expansive definition of *noise*. If we think of noise as simply sampling variability or measurement error, then obviously we can't attribute a correlation to noise when we have all the relevant data and there's no measurement error. But we could expand our definition of *noise* to also include idiosyncratic variation in these two variables that is completely unrelated but just happened to be correlated.

Another way to resolve this puzzle is to make a metaphysical leap. Sure, there's an

observed correlation between Nicolas Cage movies and swimming pool drownings in this world, but that's just a small sample of a broader population of alternative, counterfactual worlds. Those worlds are just like our own, but all of the idiosyncratic, unrelated factors happen to play out differently. In those counterfactual worlds, would we have any reason to expect that Nicolas Cage movies would be correlated with swimming pool drownings? If the answer is *no*, we might say that the correlation we observed is just noise or luck, even though we have all the data there is to have about Nicolas Cage and swimming pool drownings. This is related to our discussion in Chapter 6 about how to think about statistical inference even when we have data for the whole population.

If you don't like the metaphysical leap, you don't have to make it with us. Either way, we're comfortable saying that observed correlations arise from a combination of genuine causal relationships, bias, and noise (or idiosyncratic variation), and whenever someone presents a correlation as evidence of a causal relationship, it's worth thinking about all three factors.

Key Terms

- **Causal Effect:** The change in some feature of the world that would result from a change to some other feature of the world.
- **Average Treatment Effect (ATE):** The difference in average outcome comparing two counter-factual scenarios—one where everyone in some population is treated to one where everyone in that population is untreated.
- **Average Treatment Effect on the Treated (ATT):** The difference in average outcome comparing the scenario where everyone in the population of people who in fact receive treatment do receive treatment to the counter-factual scenario where everyone in the population of people who in fact receive treatment instead is untreated.

- **Average Treatment Effect on the Untreated (ATU):** The difference in average outcome comparing the counter-factual scenario where everyone in the population of people who were in fact untreated do receive treatment to the scenario where everyone in the population of people who were in fact untreated are untreated.
- **Difference in Means:** The difference in average outcome comparing the group of people who in fact received treatment to the group of people who in fact were untreated.
- **Expected Value:** The average value of a variable if you were to sample it over and over again (an infinite number of times) from your population of interest.
- **Bias:** The expected value of the difference between your estimate and the true effect of interest.
- **Baseline Differences:** Differences in the average outcome between two groups (e.g., the treated and untreated groups), even when those two groups have the same treatment status.
- **Confounder:** A feature of the world that (i) has an effect on treatment status and (ii) has an effect on the outcome over-and-above the effect it has through its effect on treatment status.
- **Reverse Causality:** When the outcome affects treatment status.
- **Overestimate:** When the bias is positive, so that the estimate is larger than the true effect in expectation.
- **Underestimate:** When the bias is negative, so that the estimate is smaller than the true effect in expectation.

- **Mechanism:** A feature of the world that the treatment affects which then, in turn, affects the outcome.

Exercises

1. Ethan, who has spent more time than perhaps he should have doing academic administration, was once at a meeting where he was briefed on the ways in which data analytics can improve universities' operations. The example the presenter was most excited about was from a data analytics team in a major research university's development (which is jargon for fundraising) department. The data analytics team had discovered the following correlation by analyzing years of data. Alumni who donate to the university six years in a row are way more likely to be lifelong givers than are alumni who only donate five years in a row.

The presenter was excited because, in their view, this finding from the analytics team suggested a clear strategy to improve fundraising and alumni engagement. In particular, based on this analysis, they had decided to make a major push to encourage alumni who had already given for five years in a row to give a sixth. The idea being that the evidence of a correlation between giving for six-years and giving in the future suggested that giving in that sixth year had a big causal effect on future giving, so resources spent encouraging five-year givers to become six-year givers were being put to the best possible use.

Provide two arguments, based on the clear thinking skills you acquired in this chapter, to explain why this might not be good advice.

2. Shortly after the Harvard psychologist Daniel Gilbert's very nice book, *Stumbling on Happiness*, was released, he was on TV, where he informed Steven Colbert that, "marriage is one of the best investments you can make in happiness." That advice

implicitly rests on a causal claim: marriage causes happiness.

Lots of recent research documents a positive correlation between marriage and happiness. But is the relationship causal?

- (a) Provide an argument for why the correlation between marriage and happiness might be the result of reverse causation (happiness causing marriage, rather than the other way around).
- (b) Identify two confounders that you think might make a causal interpretation of this correlation problematic. For each, explain why you believe the confounder might affect both treatment (being married) and outcome (happiness).
- (c) Sign the bias for each of the confounders you identified. Having done so, explain whether each tends to make the observed correlation between marriage and happiness an over- or under-estimate of the true causal effect.
- (d) A study by Anke Zimmermann and Richard Easterlin follows people from up to four years prior to their first marriage through several years after getting married. The basic finding is illustrated in Figure 9.13 (taken from Zimmermann and Easterlin). The horizontal axis of the figure is time before and after a person's first marriage. Time 0 is when the person gets married. The line shows the average life satisfaction of a person at each stage of a relationship.
 - i. Compare the life satisfaction of people who have been married for a while to that of people who are not married but are living with their partner. Do you find this evidence supportive of or contrary to Gilbert's advice?
 - ii. Identify a confounder that this comparison suggests may have existed in the original correlation.
- (e) A study by Jonathan Gardner and Andrew Oswald also follows individuals over time, but asks a different question. It considers what happens to people's hap-

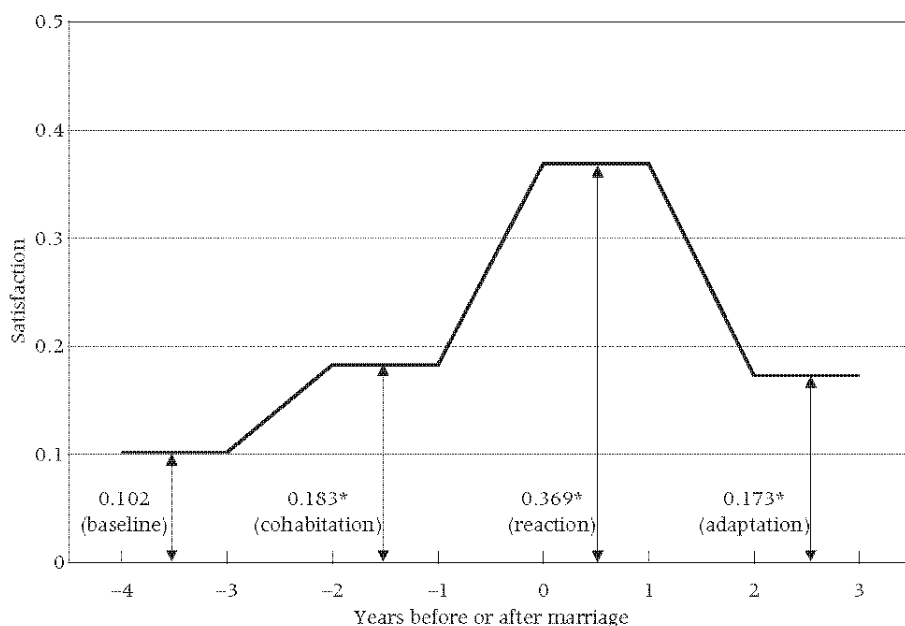


Figure 9.13: Happiness over time, before and after marriage. (source: Zimmermann and Easterlin)

piness when marriages end. The study looks at two ways a marriage might end: divorce or death of a spouse. The results are summarized in the Figure 9.14 (taken from Gardner and Oswald).

The horizontal axis shows years relative to an important event (divorce or widowhood) at time 0. The vertical axis shows life satisfaction. The green line represents people who are married and whose marriage does not end. The red line shows people whose spouse dies. The blue line shows people whose marriage ends in divorce.

- i. Compare the life satisfaction of married people who stay married (the green line) to married people who get divorced (the blue line). Does these to groups relative level of happiness prior to the divorced group getting divorced make you more or less confident in Gilbert's causal interpretation? Why?

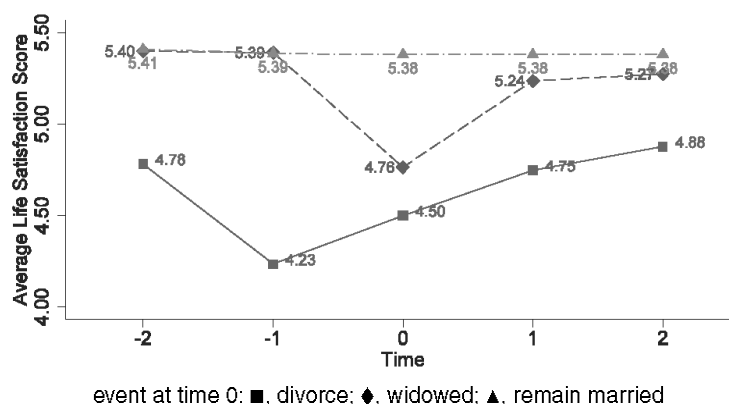


Figure 9.14: Comparing life satisfaction over time of people whose marriages do and don't dissolve. (source: Gardner and Oswald)

- ii. Now consider the widows (red line). Look at their level of happiness before the death of their spouse and starting two years after the death of their spouse. Does this make you more or less confident in Gilbert's causal interpretation? What does this comparison make you think might be going on in Gilbert's original correlation?

3. Download the data set at the following link:

`home.uchicago.edu/~bdm/HouseElectionsSpending2018.csv.`

There is a description of the variables in this data set available at

`home.uchicago.edu/~bdm/spending_readme.txt.`

- (a) Run a linear regression that finds the relationship between incumbent vote share and incumbent spending. (Note: this may require you to recode some of the variables in the data set or generate your own variables that better suit your goal. For example, if you are using Stata, you may need the following commands

in order to complete this task: *generate*, *replace*, and *regress*.)

- i. Is the correlation positive or negative?
 - ii. According to this data, do incumbents who spend more do better or worse?
 - iii. Interpret the magnitude and direction of the correlation between incumbent spending and incumbent vote share.
- (b) Do the same as above for challengers.
- (c) Let's think about whether the regressions you've run constitute compelling evidence on the effect of campaign spending on vote shares.
- i. Identify three confounders you are worried about.
 - ii. Do you have any variables in this data set that measure those confounders?
If no, identify a variable that might plausibly measure a confounder that is in the data set.
 - iii. Using linear regression, assess whether incumbent spending and challenger spending (the treatments) are in fact correlated with one of the potential confounders measured in the data set.

Notes

The study of the Preuss School is

McClure, Larry, Betsy Strick, Rachel Jacob-Almeida, and Christopher Reichher. 2005. The Preuss School at UCSD. Research report of The Center for Research on Educational Equity, Assessment and Teaching Excellence <create.ucsd.edu/_files/publications/PreussReportDecember2005.png>.

The study on the Knowledge is Power Program is

Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. “Who Benefits from KIPP?” *Journal of Policy Analysis and Management* 31(4):837–860.

The study of practice and skill among violinists is

K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. “The Role of Deliberate Practice in the Acquisition of Expert Performance.” *Psychological Science* 100(3):363–406.

The study of hormonal contraception and HIV is

Renee Heffron, Deborah Donnell, Helen Rees, and Connie Celum. 2012. “Use of Hormonal Contraceptives and Risk of HIV-1 Transmission: A Prospective Cohort Study.” *The Lancet Infectious Diseases* 12(1):19–26.

The study examining the correlation between electoral success and campaign spending for incumbents and challengers is

Gary C. Jacobson. 1978. “The Effects of Campaign Spending in Congressional Elections.” *American Political Science Review* 72(2):469–491.

The study of happiness before and after marriage is

Anke C. Zimmermann and Richard A. Easterlin. 2006. “Happily Ever After? Cohabitation, Marriage, Divorce, and Happiness in Germany.” *Population and Development Review* 32(3):511–528.

The study of happiness before and after the ending of a marriage is

Jonathan Gardner and Andrew J. Oswald. 2006. “Do Divorcing Couples Become Happier by Breaking Up?” *Statistics in Society* 169(2):319–336.

Chapter 10

Controlling for Confounders

What You'll Learn

- If we can observe a confounder, we can control for it and mitigate the bias arising from it.
- The most common way to control for a confounder is by including it in a regression, although there are other approaches as well.
- We'll provide graphical and substantive intuition for how this works.
- Controlling is not magic. It doesn't remove bias arising from unobserved confounders or reverse causation.

In Chapter 9, we saw that confounders are a big problem when we’re trying to learn about causal relationships from correlations. Here, we are going to talk about the first line of defense against confounders, controlling. You likely have heard people talk about *controlling* before, but what does it really mean? The easiest way to begin to understand is through some examples.

A not terribly surprising fact about the United States Congress is that Republicans are more likely to vote in a conservative manner than Democrats. One way of measuring this, quantitatively, is through the scores given to each Congressional representative by the right-leaning interest group the American Conservative Union (ACU). Each year, the ACU chooses 25 important bills and gives each congressperson a score between 0 and 100 based on how they voted on those bills. Since the ACU leans right, a higher score indicates a more conservative voting record.

We can back up the claim that being a Republican is correlated with a conservative voting record by checking whether Republicans have higher ACU scores than Democrats on average. Table 10.1, based on data from the House of Representatives in 1997, shows that they do. Democratic congressional representatives have an average ACU score of 19, while Republicans have an average ACU score of 83. On average, Republicans vote 64 ACU points more conservatively than do Democrats.

Average ACU Score	
Republicans	83
Democrats	19
Difference	64

Table 10.1: Comparing the voting records of Republicans and Democrats in the U.S. Congress

These data indicate that Republican and Democratic congressional representatives vote

quite differently. Moreover, lots of evidence shows that this type of polarization inside congress has been growing over time. What might explain this increasing polarization?

One idea advanced by many prominent political scientists is that party pressure causes the divergence in legislative voting behavior. Parties have lots of tools at their disposal to pressure rank-and-file members to vote the party line. Perhaps most important among these tools is help with fundraising for reelection campaigns. If campaign finance has become more important over time, then the parties may be increasingly able to discipline their members to vote the party line, causing polarization.

But, before we interpret the correlation between party membership and voting record as evidence for the effect of party discipline, we should consider possible confounders. A confounder, in this case, is some other feature of the world that both affects the party membership of a congressional representative and his or her voting record.

Ideology is one obvious candidate for a confounder. The Republican party has a conservative reputation. The Democratic party has a liberal reputation. Hence, conservatives may be more likely to run as a Republican and liberals may be more likely to run as a Democrat. Moreover, a politician's personal ideological leanings may well influence how he or she votes on legislation once in Congress. If people sort into the parties according to ideology in this way, there is reason to think that Republican representatives would vote more conservatively and Democratic representatives would vote more liberally, even if the parties exercised no discipline. So personal ideology is plausibly a confounder. In light of this, it would be a mistake to interpret the correlation between party membership and voting record as an unbiased estimate of the causal impact of party discipline on the voting behavior of representatives.

In order to address this potential confounder, we would like to control for it. In its simplest form, controlling for ideology simply means looking at the correlation between party membership and voting record, holding personal ideology constant. To do so, we first

need a measure of personal ideology. Fortunately, we have one.

In 1996, a non-partisan organization called Project Vote Smart administered a survey, the National Political Awareness Test (NPAT), to congressional candidates. The survey asked candidates their views on a wide array of issues. From their answers, Project Vote Smart then generated a liberal to conservative ranking. 76% of candidates responded to the survey, so we have a measure of the political ideology of a large number of congressional representatives.

To control for personal ideology in our analysis of the relationship between party membership and voting record, we simply compare the voting records of Democrats and Republicans with similar NPAT scores. If the NPAT is doing a good job of measuring personal ideology, these comparisons will tell us about the difference in voting records of Republicans and Democrats, holding personal ideology constant (or, “controlling for personal ideology”).

Table 10.2 below sorts congressional representatives into 5 bins, based on their NPAT scores. The left-most bin has representatives with the most liberal ideology according to their NPAT answers. The bins become progressively more conservative as we move to the right.

Looking at the data broken down in this way, a couple things immediately jump out. First, and most importantly, in no column is the difference between Republican and Democratic voting records anywhere close to the 64 point difference we found before controlling for personal ideology. This suggests that personal ideology was an important confounder in that correlation—a large portion of the difference in voting record between Democrats and Republicans was due to the members of those two parties having different underlying personal preferences about policy, rather than party pressure. The reason, of course, is as we said earlier. More conservative people tend to become Republicans and more liberal people tend to become Democrats. This fact is reflected in the observation that the number of people in each cell is increasing in conservatism for Republicans and decreasing in

		liberal ← NPAT Percentile → conservative				
		1-20	21-40	41-60	61-80	81-100
Republicans	Avg ACU Score	n/a	44	68	86	94
	# of People	0	4	45	69	69
Democrats	Avg ACU Score	10	18	41	96	84
	# of People	70	66	24	1	1
Avg Republican ACU Score –		n/a	26	27	-10	10
Avg Democratic ACU Score						

Table 10.2: NPAT Scores Controlling for Party

conservatism for Democrats (i.e., almost all the Republicans are in the NPAT 41st-100th percentile and almost all the Democrats are in the NPAT 1st-60th percentile).

Second, within a party, as you move across the ideological bins, average ACU scores are, for the most part, going up. There is one exception—Democrats in the 61-80th percentile vote more conservatively than Democrats in the 81st-100th percentile—but it is worth noting that this exception involves the comparison of exactly 2 people, since there are so few ideologically conservative Democrats.

Third, the difference between average Republican and average Democratic voting records differ across the columns. That is, the correlation between party and voting record differs depending on ideology. This is fine. But often we want a single, overall measure of the correlation between partisanship and voting record controlling for personal ideology, rather than an ideology-by-ideology measure. To get that single number, we will need to take

some sort of a weighted average of the differences from the various columns. But how do we decide what weights to give to each column?

As a starting intuition for how to decide on weights, note that there is clearly one column that is more informative than the others about the different voting behavior of Democrats and Republicans with similar personal ideologies—the column for the 41st-60th NPAT percentile. In each of the other columns, there are either very few Republicans or very few Democrats. But in the 41st-60th NPAT percentile column there are a large number of representatives from both parties. This isn't surprising—the place to look for ideological overlap across the parties is in the ideological center. So we probably want our weighted average to put a lot of weight on that column.

More generally, it is useful to think back to Chapter 5, where we learned about how regression fits a line to data to minimize the sum of squared errors. One principled way to choose weights for the five columns is to use weights that minimize our predictive error in this way. And, indeed, one can implement precisely that weighting scheme through the following regression:

$$\text{ACU Rating} = \alpha + \beta_1 \cdot \text{Republican} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon.$$

In this regression, the unit of analysis is an individual representative. The variable ACU Rating is an individual representative's ACU score. The variable Republican is what we call a *dummy variable*: it takes the value 1 if the representative is a member of the Republican party and the value 0 if the representative is a member of the Democratic party. The various NPAT variables are also dummy variables, taking a value of 1 if the representative is in the relevant percentile range and a value of 0 otherwise. (Recall, since everyone is in one of the five NPAT categories, one of them must be omitted. Here, we have omitted the 1st-20th percentile.)

The coefficient β_1 in this regression will give us the weighted average we have been talking about. That is, β_1 is the correlation between ACU Score and being a Republican controlling for personal ideology (as measured by NPAT percentile). Of course, we will also get estimates for the coefficients on the four included NPAT categories and the constant term. These also have interpretations. However, we are running the regression because we are interested in the correlation between ACU Score and Republicanism controlling for ideology, so we focus on β_1 .

Actually running this regression on our data yields an estimate of β_1 equal to 24. Not surprisingly, this is very close to the difference between average Republican ACU score and average Democratic ACU score in the column corresponding to the 41st-60th percentile, which, as we said above, is where almost all the information is. The regression, of course, puts a little weight on the other columns, dragging the estimate down from 27 to 24. But that column is basically telling us the answer.

Of course, having controlled for ideology, we still probably don't have a terribly credible estimate of the causal effect of party discipline on the voting records on congressional representatives. This is because there could be many other confounders beyond personal ideology. That is, within an NPAT bin, there may be lots of other factors that lead some people to become Democrats and others to become Republicans that also have an independent effect on their voting behavior in congress. For instance, even holding fixed personal ideology, Democrats may tend to represent districts with more left-wing voters and Republicans may tend to represent districts with more right-wing voters. If politicians choose how to vote on bills with an eye toward how their voters will react, then these differences in constituencies are yet another confounder. We are sure you can think of others.

As the list of confounders proliferates, making a table that breaks down the data into all the different possible cells becomes more difficult and unwieldy. But, as long as you can measure the potential confounders, you can control for them in a regression. Doing so will

always get you an estimate of β_1 reflecting the weighted average of the various cells in that (imagined) big table that minimizes the sum of squared errors. Given this, regression will be our most important tool for controlling for confounders. As such, it is useful to have a better understanding of exactly how controlling with regression works.

How Does Regression Control?

We've seen that controlling for a variable (X) can change the coefficient describing the relationship between some other variable of interest (T) and an outcome variable (Y). In particular, controlling for X will change the estimated relationship between T and Y , if X is correlated with T and has an independent effect on Y . Here's one way to think, graphically, about what the regression is doing when we control for a variable.

Suppose we want to know the effect of height on income, in which case both our outcome and treatment variables of interest are continuous. Figure 10.1 shows some data on income and height from the National Longitudinal Survey conducted by the Bureau of Labor Statistics. A representative sample of Americans born between 1980 and 1984 were asked about their heights and their incomes in 2014, when they were between the ages of 34 and 38.

To allow for easier visualization, we grouped respondents by height and gender, so every dot in Figure 10.1 corresponds to a group of 15 or more individuals of the same gender and height (measured in inches). The figure plots the average income of each group, measured in thousands of dollars above 20,000, and the average height, measured in feet above 5. (You'll see in a moment why we scaled our variables in such an unusual way.) The blue dots correspond to groups of men and the red dots correspond to groups of women.

Visually, we see a strong, positive correlation between height and income. What would we get if we ran a regression of income on height with this data, ignoring gender? As we've

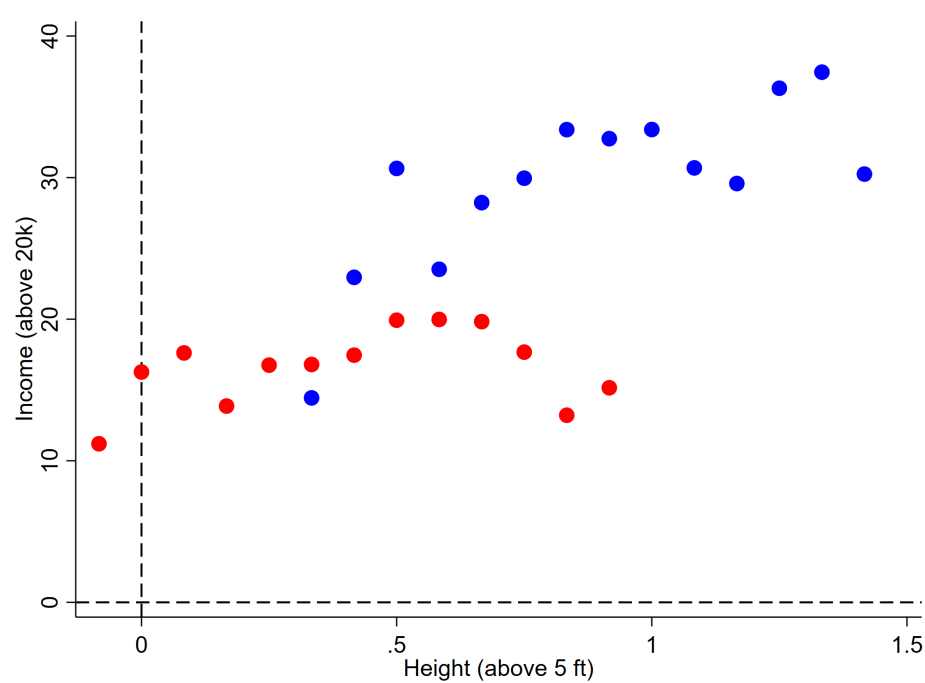


Figure 10.1: Income and height among 34-38 year old Americans in 2014 (data from NLSY97).

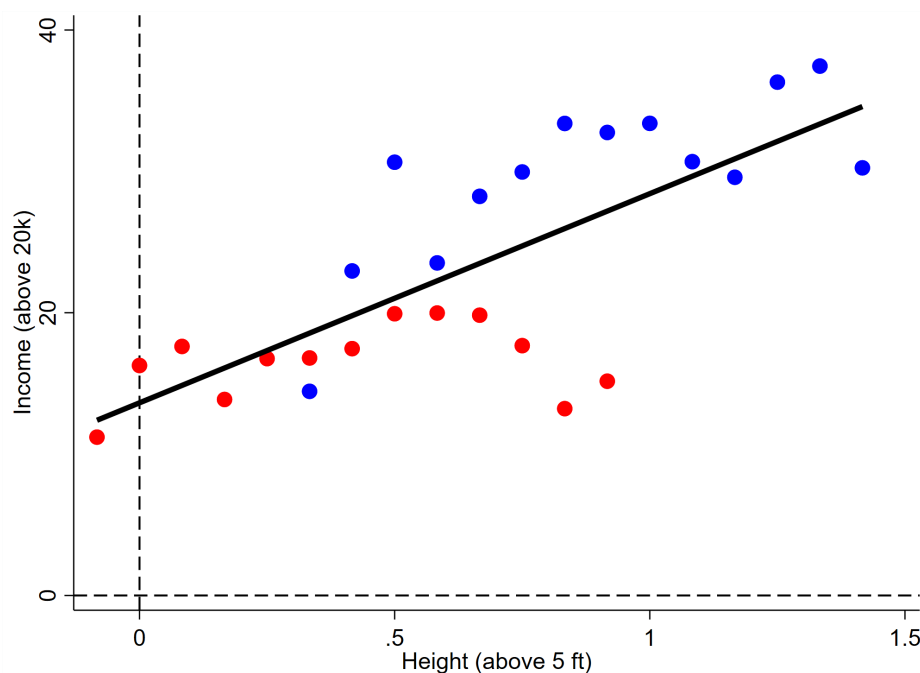


Figure 10.2: Regressing income on height.

seen previously, this would simply involve finding the line that best fits the data. Figure 10.2 plots that line. Indeed, the best-fitting line has a strong, positive slope, indicating that, on average, taller people earn higher incomes.

As we already know, the regression finds the line that best fits the data by identifying the values of α and β in the following equation that minimize the sum of squared errors:

$$\text{Income} = \alpha + \beta \cdot \text{Height} + \varepsilon.$$

These two values are illustrated by Figure 10.3. The height of the line when $\text{Height} = 0$ (i.e, when a person is 5 feet tall) is α , and the slope of the line is β . For this particular data set, we estimate a slope of about 14.8. On average, people who are one foot taller earn an extra 14,800 dollars of income per year!

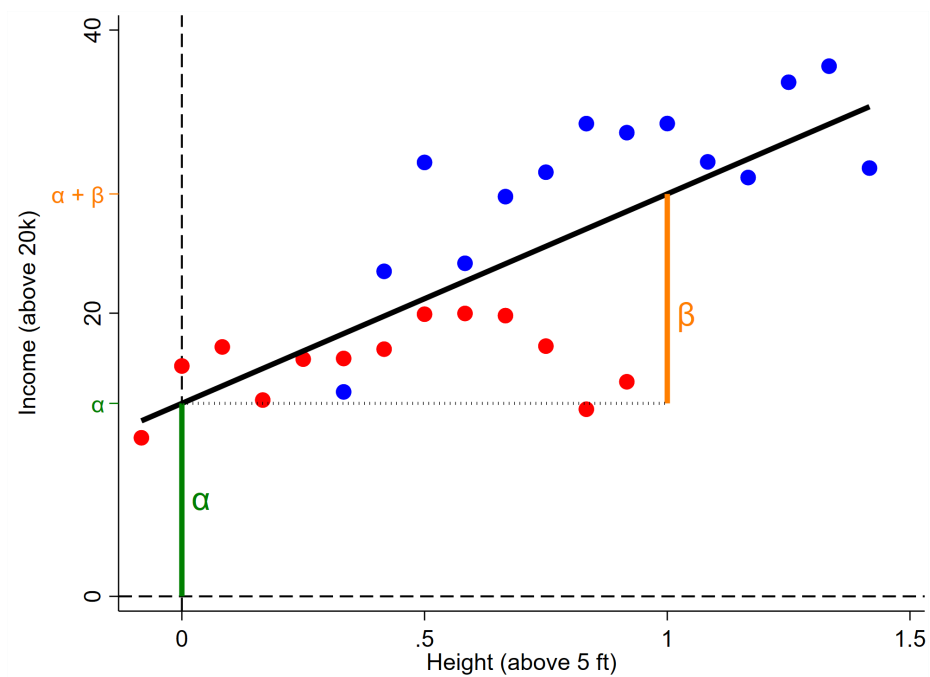


Figure 10.3: Graphical interpretation of regression coefficients.

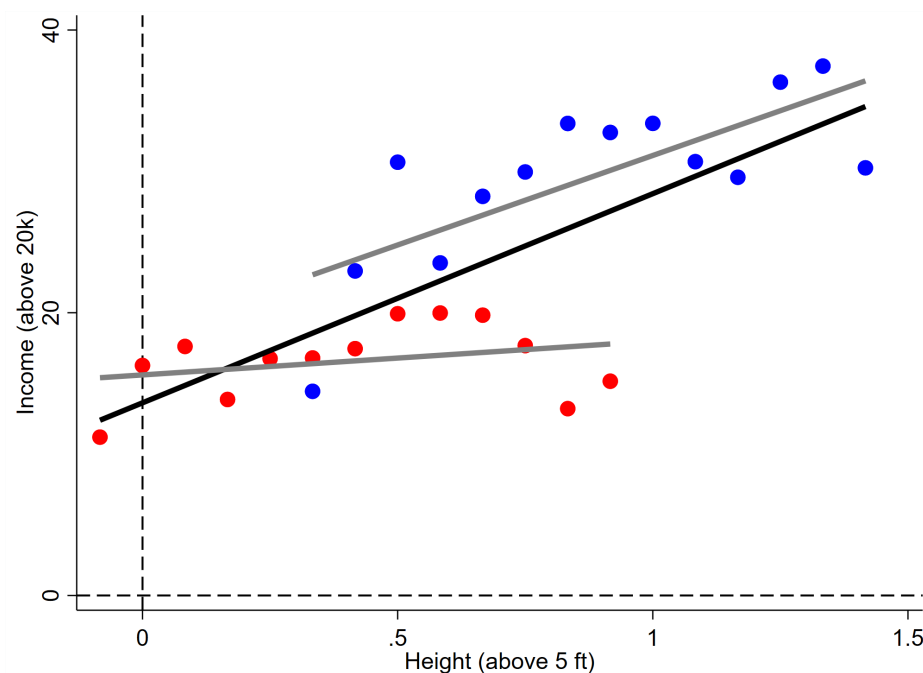


Figure 10.4: Running separate regressions for men and women.

Of course, before we draw a causal interpretation from this regression coefficient, we should think about confounders. Gender is one possibility. Men are, on average, taller than women. And we suspect that men, on average, earn higher incomes than women for reasons unrelated to height. Indeed, we can see in the picture that women do seem to have lower heights and lower incomes, on average. So gender is a confounder that we might want to control for in this regression.

One way we could start addressing this concern would be to run separate regressions for men and women:

$$\text{Income} = \alpha_W + \beta_W \cdot \text{Height} + \varepsilon.$$

$$\text{Income} = \alpha_M + \beta_M \cdot \text{Height} + \varepsilon.$$

If we did that, we would fit two lines, as shown in Figure 10.4.

The separate regression lines for men and women are shown in gray, while the previous regression line that pooled everyone together is still shown in black. Interestingly, the correlation between height and income is smaller within each gender than it is across the population as a whole. That is, both β_W and β_M are smaller than β from our earlier regression. Also notice that the slope is greater for men than it is for women, $\beta_M > \beta_W$.

This procedure of splitting the data and running separate regressions tells us the correlation between income and height separately for men and women. Thinking back to our legislative politics example, this is analogous to the cells at the bottom of the table that told us the difference in average ACU score between Republicans and Democrats for each bin of NPAT scores.

While the separate correlations are good to know, just as in the congressional politics example, we might want to have one summary estimate of the correlation between income and height, controlling for gender. Of course, that number will be a weighted average of the two slopes in gray above (just as in the legislative politics example where the single number was a weighted average of the different bottom cells of the table). But we need to know how to assign the weights.

The most straightforward way to do this is to run a regression of income on both height and gender. The equation would look like this:

$$\text{Income} = \alpha + \beta \cdot \text{Height} + \gamma \cdot \text{Male} + \varepsilon.$$

Graphically, how will this regression separately estimate α , β , and γ ? Instead of finding one line that best fits the data, we can think of finding two lines that best fit the data—one for men and one for women. But unlike when we ran separate regressions, we now constrain those two lines to have the same slope (β). Figure 10.5 shows how those two lines look if we do that.

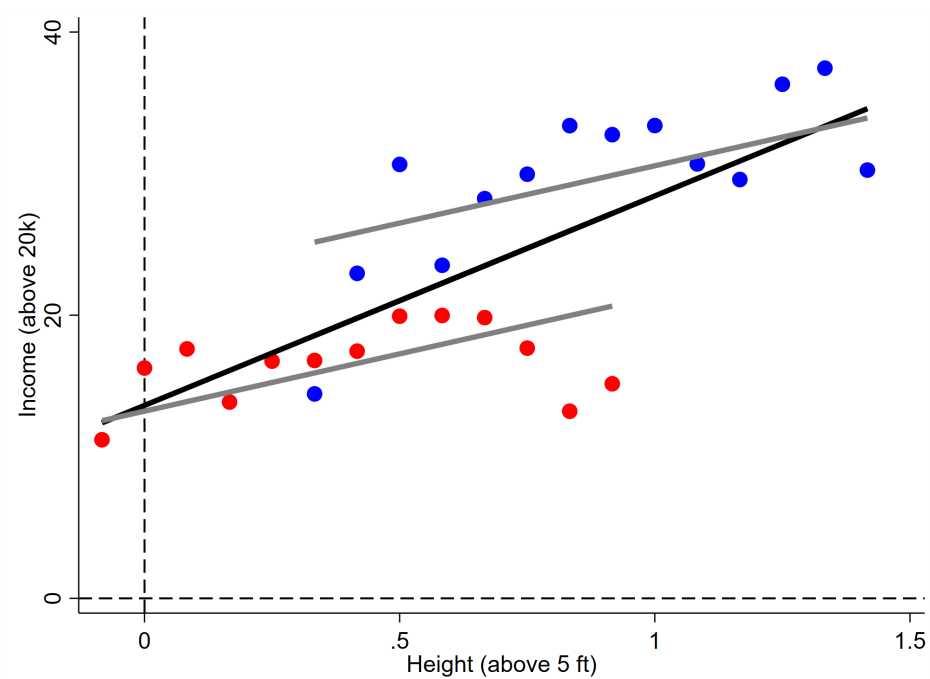


Figure 10.5: Controlling for gender by including it in the regression of income on height.

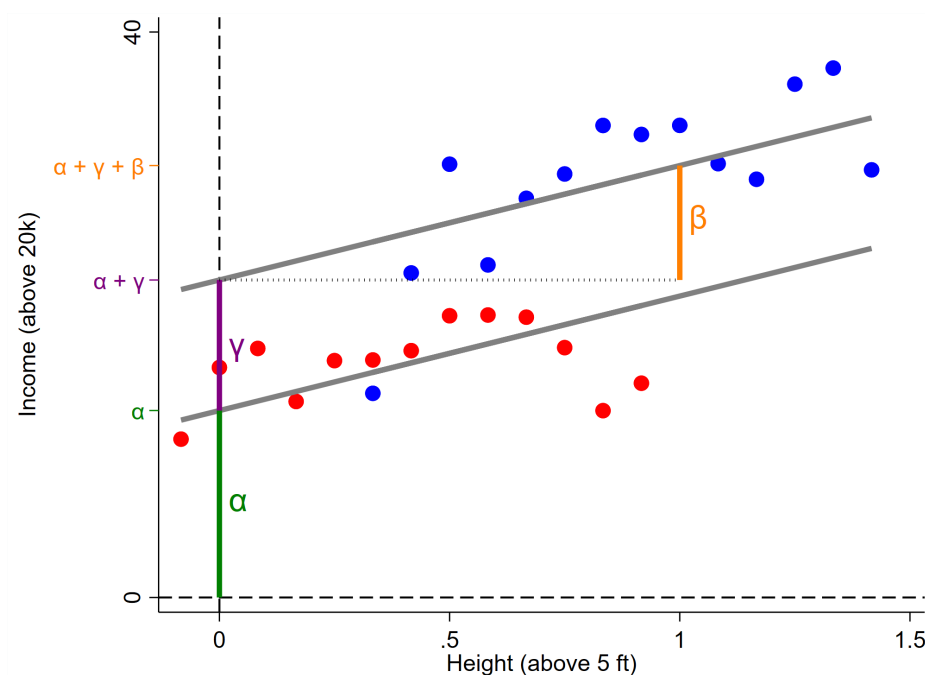


Figure 10.6: Graphical interpretation of regression coefficients when controlling for gender in the regression of income on height.

Notice that the slope of these two lines is identical, by construction. And the slope is somewhere in between the slope for those we got running the two separate regressions. That is, it is a weighted average of the two. Having estimated these two parallel lines that best fit the data, Figure 10.6 shows how we've estimated the parameters in the equation.

The y-intercept of the line for women (Male = 0) is α . The distance between the two lines is γ . And the slope of the two lines is β . Put differently, α is the average income for women who are 5 feet tall; γ is the average difference in income between men and women of the same height; and β is the average relationship between height and income, controlling for gender.

Not surprisingly, controlling for gender has significant implications for the estimated relationship between height and income. Instead of 14.8, our new estimate for the slope

is about 8.1. The change is due to the fact that gender is a confounder—it affects both height and income. Using our tools for signing the bias from Chapter 9, we know that if the confounder is positively correlated with both treatment and outcome, as is the case here, it creates positive bias. Since 14.8 was an overestimate of the true effect of height on income, when we control for gender, we get a smaller estimate.

It is also worth noting that controlling for gender not only affects our estimate of the relationship between income and height, it also affects the precision of that estimate. If we don't control for gender, there is more variability in income because we are looking across the whole sample. This noise inherently makes our estimate imprecise. By controlling for gender, we remove the part of the variability in income that is due to gender. Then, when we go to estimate the relationship between height and income, we do so with respect to the less noisy distribution of income for men and women. This fact, that controlling for a confounder improves precision, is quite general. (Of course, if you keep trying control variables until you get statistical significance, that's p-hacking, and it's a bad idea.)

Since we've talked about the analogy between what we've just done and our legislative politics example, let's revisit that example in a regression framework. Notice, in this case, the treatment (Republican or Democrat) is binary, but the potential confounder (ideology) is measured continuously by the NPAT score.

Again, start with a scatter plot, this time of the American Conservative Union rating on the vertical axis and the NPAT conservative score on the horizontal axis. In Figure 10.7, the blue dots correspond to Democrats and the red dots correspond to Republicans.

Since the treatment is binary, we can start with a simple comparison of the average ACU Rating for Republicans and for Democrats. Consider the following regression equation:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \varepsilon.$$

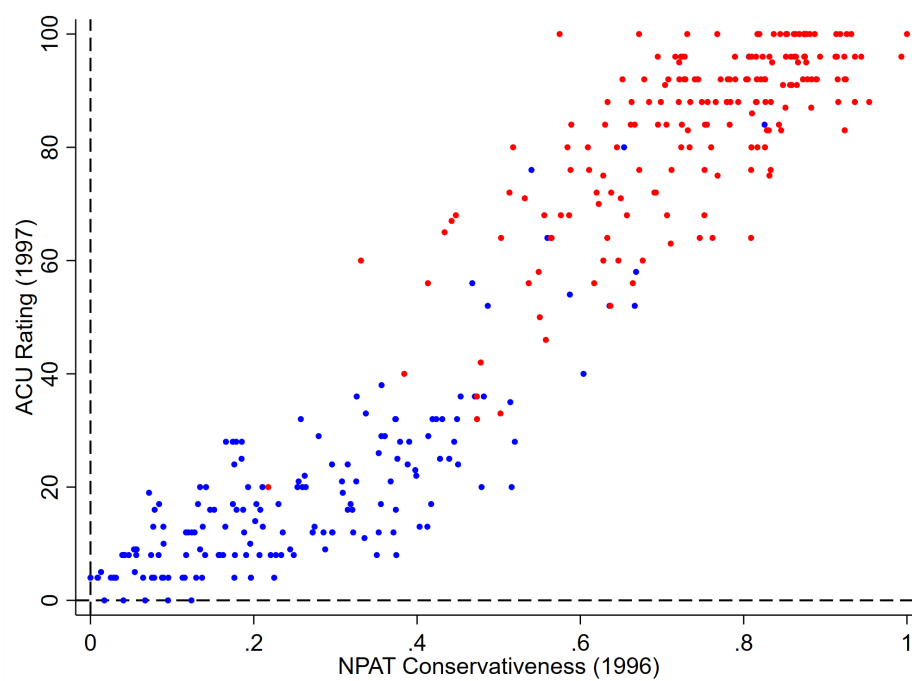


Figure 10.7: ACU score and NPAT conservative score.

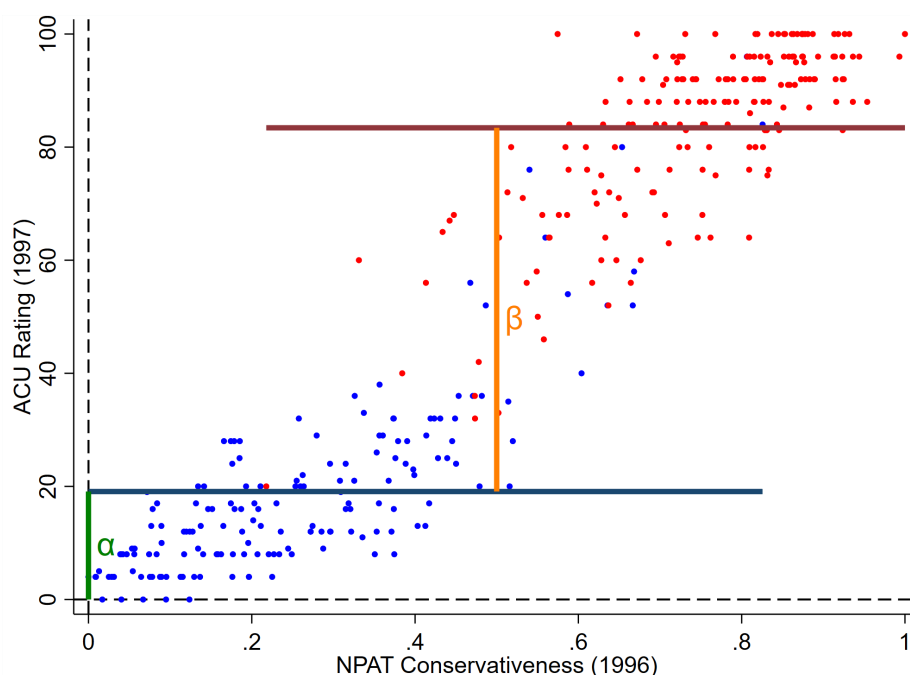


Figure 10.8: Graphical interpretation of coefficients in regression of ACU Score on party.

To minimize the sum of squared errors, the coefficient α equals the average ACU Rating for a Democrat (Republican = 0) and the coefficient β is the difference between the average ACU rating for a Republican and for a Democrat. Thus, as we’ve already seen, $\alpha = 20$ and $\beta = 84 - 20 = 64$. This is illustrated Figure 10.8, where the horizontal lines correspond to the average ACU Ratings for Democrats and Republicans.

Our concern, of course, is that personal political ideology is a confounder in this regression, so that β does not estimate the true effect of party on legislative voting behavior. So we would like to control for personal ideology. We will do so using the NPAT Conservativeness score measured on the horizontal axis.

In our previous example, we were concerned about a confounder that was measured as a binary variable—gender. Our first step in building intuition about controlling was to consider running the original regression (income and height) separately for each value of

the confounder. Then we saw that the final regression coefficient on height controlling for gender was a weighted average of the slopes of these two separately estimated regression lines.

Here, because our confounder is continuous, we cannot do a separate regression for each value of the confounder. But we can do something similar. Run a regression of ACU Rating on NPAT Conservativeness separately for Democrats and for Republicans (the subscripts P on the regression coefficients refer to the idea that this is the regression for party P):

$$\text{ACU Rating} = \alpha_P + \gamma_P \cdot \text{NPAT Conservativeness} + \varepsilon_P.$$

That gives us two regression lines, one for Republicans (in red) and one for Democrats (in blue), as in Figure 10.9.

For each value of NPAT Conservativeness, the expected ACU Rating of a Republican with that NPAT Conservativeness score is:

$$\alpha_R + \gamma_R \cdot \text{NPAT Conservativeness}.$$

And for each value of NPAT Conservativeness, the expected ACU Rating of a Democrat with that NPAT Conservativeness score is:

$$\alpha_D + \gamma_D \cdot \text{NPAT Conservativeness}.$$

Notice, this means that these regressions will give us an expected ACU Rating for a Republican with an NPAT Conservativeness of, say, .1 or of a Democrat with an NPAT Conservativeness of, say, .9 even though there are no such people in the data. Something to think about. Anyway, this means that at any given value of NPAT Conservativeness, the gap between the two lines is the difference in expected ACU Rating between Republicans

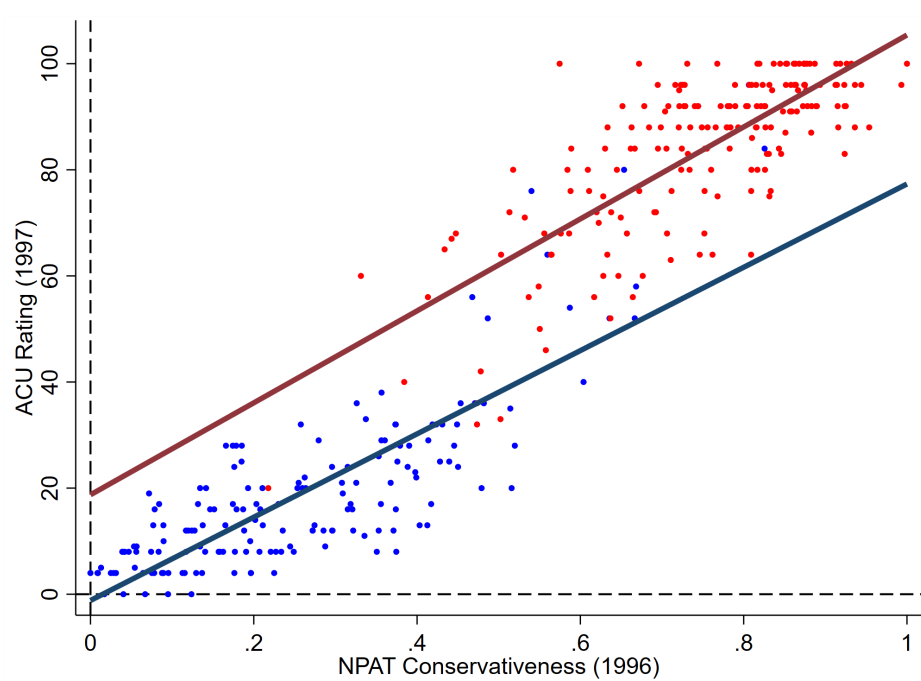


Figure 10.9: Controlling for NPAT score (ideology) in relationship between ACU score and party by running two separate regressions, one for each party.

and Democrats with that NPAT score. Hence, this regression allows us to get a continuous analogue of our earlier binary comparison. It tells us, for each value of NPAT Conservativeness, what the expected difference in mean ACU Rating is between Republicans and Democrats.

But we aren't done. As before, the goal is to get a single measure of the relationship between ACU Rating and party membership, controlling for NPAT Conservativeness. Currently, we have a separate measure of that relationship for each value of NPAT Conservativeness. As before, the final step in controlling, then, is to use regression to create the weighted average of these differences that minimizes the sum of square error. We do so with the following regression:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \gamma \cdot \text{NPAT Conservativeness} + \varepsilon.$$

Figure 10.10 illustrates this regression. The parameter α tells us the average ACU Rating of a Democrat with NPAT Conservativeness of 0. The parameter γ tells us the slope of the relationship between ACU Rating and NPAT Conservativeness. Importantly, unlike our previous two regressions, where γ_R and γ_D were different, this regression imposes that the slope of the relationship between ACU Rating and NPAT Conservativeness be the same for both parties. Hence, this slope γ is a weighted average of γ_R and γ_D . Finally, the coefficient β is the gap between the two lines. This gap is constant across NPAT Conservativeness scores because we forced γ to be the same for both parties, making the lines parallel. Hence, β is the average difference in ACU Rating between Republicans and Democrats controlling for NPAT Conservativeness.

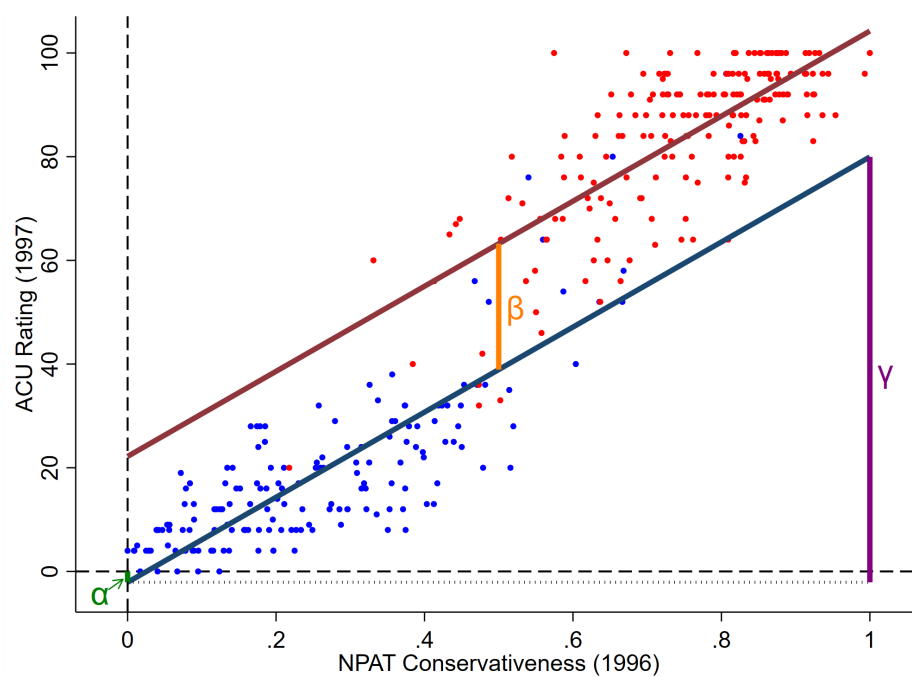


Figure 10.10: Controlling for NPAT score (ideology) in a regression of ACU score on party.

Controlling and Causation

Remember from Chapter 9 that if we want to interpret a correlation as an unbiased estimate of a causal effect, we must believe that there are no baseline differences between the treated and untreated units. In other words, our comparison has to plausibly be apples-to-apples. If we regress Y on T and X (or other possible confounders), we're still making a similar assertion if we give the coefficient on T a causal interpretation. We are saying that, other than the set of variables we controlled for (X), we believe there are no confounders in the relationship between Y and T and no reverse causality.

This might be a more defensible claim with controlling than without. But in most cases, it's still pretty strong. In our experience, it is hard to find situations (other than randomized experiments, which we will discuss in Chapter 11) where it feels plausible that there are really no omitted confounders. Typically, even if the analyst controls for lots of things, you can think of other potential confounders that are either unobservable or unmeasured in your data and, thus, can't be controlled for. For instance, ask yourself whether you can think of any potential confounders beyond gender in the relationship between income and height? The answer is, of course, yes, including a long list of economic, biological, cultural, health, and other characteristics. For instance, wealthy parents might provide their children with better nutrition, which might make them taller, and wealthy parents might also help their children in other ways that allow them to earn higher incomes. It is hard to imagine that you would be able to measure and control for all possible confounders.

Reverse causation is another reason we don't generally think controlling for confounders can uncover causal relationships on its own. In Chapter 9, we talked about how both confounders and reverse causation can prevent us from making an apples-to-apples comparison. The idea of controlling is to try to account for confounders as best as possible, but if there is reverse causation, meaning the outcome affects the treatment, there's no amount

of controlling that can make that problem go away.

Let us give you an example. There is lots of concern that exposure to social media is bad for people. And, indeed, many studies show a negative correlation between social media usage and various measures of a person's subjective well-being and mental health.

Of course, that correlation may not reflect a causal relationship. For instance, there might be reverse causality—perhaps people who are sad, lonely, or distressed spend more time on social media than people who are happier or more socially connected. Or there might be confounders—perhaps socio-economic status, education, or geography affect both social media usage and subjective well-being.

A first thing you might think to do is to get at an estimate of the causal relationship is to control for some of these confounders. We are interested in examining is how well that controlling strategy does at recovering the true causal effect.

Happily there is a study that can provide some insight into that question. A group of scholars interested in the effects of social media ran an experiment. They first identified a large group of Facebook users willing to participate in their study. (The participants didn't know what the study was about.) From each of these people they elicited measures of subjective well-being, Facebook usage, and how much they'd have to be paid to turn off Facebook for a month. Then the experimenters randomly selected some of these people and in fact paid them to turn off Facebook for a month (which they were able to monitor). The others didn't turn off Facebook, but continued to be part of the study as the control group. The researchers then measured subjective well-being again at the end to see whether turning off Facebook had changed the subjective sense of well-being for those in the treatment group, relative to those in the control group who did not turn off Facebook.

The nice thing about this study, for our purposes, is that the experiment, by randomly assigning Facebook usage, gives an unbiased estimate of the effect of Facebook usage. But,

the researchers also just asked about Facebook usage and subjective well-being at the start of the study. So by comparing levels of well-being across people with different levels of Facebook usage at the beginning of the study they can also replicate the simple correlation reported in earlier studies. Moreover, the researchers also know a bunch of details about the individuals in their study and so can control for some potential confounders—e.g., income, age, sex, education, race, political affiliation.

If we were able to control for all confounders, then the estimate of the relationship between Facebook usage and subjective well-being from controlling and from the experiment would be the same in expectation. So, by comparing the simple correlation, the correlation controlling for these potential confounders, and the experimental estimate, we can start to get a sense of how well controlling does, in this setting, at recovering the true causal effect.

Figure 10.11 shows the simple correlation, the correlation controlling for potential confounders, and the estimate from the experiment (each surrounded by a 95% confidence interval). All are measured in units of average Facebook use per day. As you can see, the simple correlation gives the biggest estimate of Facebook's negative relationship with subjective well-being. Controlling for potential confounders reduces that estimate a bit. But the experimental estimate is about one-third the size of the estimates from the simple correlation and about one-half the size of the estimate controlling for potential confounders. This suggests that the strategy of controlling, here, still leaves us with a substantial over-estimate of the true effect.

Reading a Regression Table

When you run a regression on your computer, or see regression results discussed in a report, they are often presented in the form of a table. So it is worthwhile to be able to understand and interpret the different parts of a regression table.

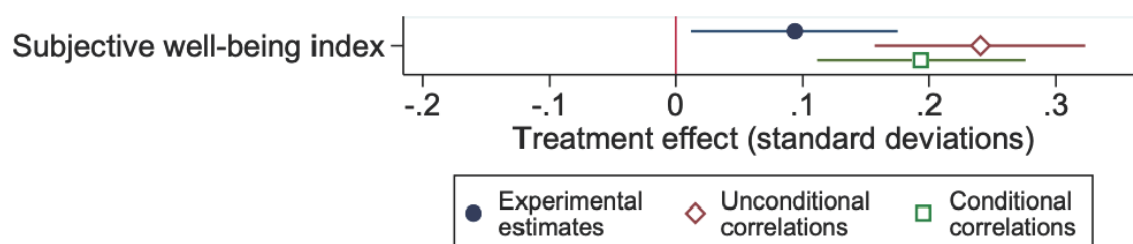


Figure 10.11: Various estimates of relationship between Facebook usage and subjective well-being. (From Allcott, Braghieri, Eichmeyer, and Gentzkow (2109).

To get a sense of how this works, let's go back to our analysis of the relationship between congressional voting records and party. In that setting, we've talked about three regressions we might consider running.

First, you might want to know the relationship between voting and party, without controlling for anything. That is described by the following regression:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \varepsilon.$$

Next, because ideology is a likely confounder, you might also want to control for NPAT score. We've talked about two different ways of doing so. You might bin congresspeople and include dummy variables representing each bin in the regression:

$$\text{ACU Rating} = \alpha + \beta_1 \cdot \text{Republican} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon.$$

This is what we did to find the right weighted average in our discussion surrounding Table 10.2.

Our approach surrounding Figure 10.10, however, probably makes more sense. There we just controlled directly for the continuous NPAT variable as follows:

$$\text{ACU Rating} = \alpha + \beta \cdot \text{Republican} + \gamma \cdot \text{NPAT Conservativeness} + \varepsilon.$$

Variables	ACU Rating	ACU Rating	ACU Rating
Republican	64.32** (1.71)	23.74** (2.25)	24.28** (1.98)
NPAT ₂₁₋₄₀		8.01** (1.76)	
NPAT ₄₁₋₆₀		32.74** (2.29)	
NPAT ₆₁₋₈₀		52.27** (2.83)	
NPAT ₈₁₋₁₀₀		59.77** (2.83)	
NPAT Conservatism			82.05** (3.44)
Constant	19.09** (1.25)	10.29** (1.24)	-2.10 (1.18)
Observations	349	349	349
R-squared	0.80	0.92	0.93

Table 10.3: Relationship between ACU rating and party. Standard errors in parentheses.
 ** $p < 0.01$, * $p < 0.05$

Table 10.3 presents the results from each of these three regressions.

Let's talk about how to read this table. The first column simply contains labels. The second column shows the results from our first regression: ACU Rating on Republican, controlling for nothing. The third column shows the results from our second regression: ACU Rating on Republican, controlling for NPAT category. The fourth column shows the results from our third Regression: ACU Rating on Republican, controlling for the continuous NPAT Conservativeness score.

Along the first row, we see the name of our dependent variable. For these regressions, this is always ACU Rating. For each of our regressions, the row labeled Republican shows three pieces of information. The top number is our estimate of the coefficient on Republican

in our regression. The bottom number in parentheses is the standard error on that estimate. And the stars indicate whether the result is statistically significantly different than zero (and at what level). Looking across this row, we see that in the first regression, the coefficient on Republican is 64.32. But once we control for NPAT score, it drops dramatically. If we control with NPAT categories, it drops to 23.74. And if we control with the continuous NPAT Conservativeness score, it is 24.28. (Not surprisingly, it doesn't matter much exactly how we control for ideology.)

Going down the rows, we then get the coefficient estimates, standard errors, and statistical significance for each of our control variables. This is why the next five rows are blank in the second column—we didn't control for anything in that regression. In the third column, the four rows associated with the NPAT categories are filled in, but the row associated with the NPAT Conservativeness score is blank. And in the fourth column, the NPAT category rows are blank, but NPAT Conservativeness is filled in. For all three regressions, the row called Constant is filled in. This is the estimate of the y-intercept (α) from that regression.

The table contains two more pieces of information. For each regression, the table tells us how many observations there were in the data. Here, the answer is 349, reflecting the number of congresspeople who filled out the NPAT survey in 1997.

And, for each regression, the table reports the R-squared. Recall, from Chapter 2 that the R-squared is the proportion of variation in one variable that can be predicted by variation in other variables. So a value of 0.93 in our final regression says that, within the sample of data we have, by knowing congresspeople's party and NPAT score, you can predict 93 percent of the variation in their ACU rating.

While that sounds pretty good, we urge you not to over-interpret the R-squared. In fact, when we run regressions, we often don't even report the R-squared. Typically, our goal is not to predict or model the variation in our dependent variable. It is to learn whether our key treatment variable matters for our outcome. For that, what we are really interested is

the coefficient estimate on that variable. Moreover, getting a big R-squared, on its own, isn't very meaningful. One easy way to successfully predict a lot of the variation in your data is to just include lots of control variables in your regression. But this doesn't mean you've understood what is going on at all. Think back to our discussion of over-fitting in Chapter 5. Just because you fit the data really well (which is all that a high R-squared means) by including lots of variables, doesn't mean you can do a good job predicting what the outcome will be when you look at observations not from your data set. And in some cases, you can have a reliable, unbiased estimate of your quantity of interest even though your R-squared is low.

The Anatomy of a Regression

Now that we have a conceptual and graphical understanding of how regression works and its uses in causal inference, let's go into a bit more detail. The key ingredients in any regression for causal inference are

- the *dependent variable* (also called the *outcome variable*),
- the *treatment variable*, and
- a set of *control variables*.

The dependent variable is the outcome you are trying to understand. The treatment variable is the feature of the world whose effect on the dependent variable you are trying to estimate. And the control variables are potential confounders that you are including in the regression to reduce bias.

In the simple case where there is only one control variable, we write the regression equation as follows:

$$Y = \alpha + \beta \cdot T + \gamma \cdot X + \varepsilon. \tag{10.1}$$

Y is the dependent variable, T is the treatment variable, X is the control variable. The regression parameters (i.e., the quantities we'd like to estimate) are the intercept or constant term α , the effect of the treatment β , and the “effect” of the control variable γ . There is also an error term, ε .

There's nothing in the regression equation that distinguishes the treatment variable from the control variable. This distinction is conceptual and is driven by the question you are trying to answer. If you want to know the effect of party on voting, controlling for ideology, the party variable is your treatment and NPAT is your control. But if you had wanted to know the effect of ideology on voting, controlling for party, this would be reversed.

This is also why the word effect is in scare quotes above when referring to the effect of the control variable. Often, we don't actually care about the value of γ . What's important is that β is an effect of interest, and we are going to try to estimate it in an unbiased way.

One way to read the equation above is to take it literally. We can pretend that we know the data generating process. For each individual i , their outcome equals a common intercept plus $\beta \cdot T_i$ plus $\gamma \cdot X_i$ plus random noise. Another way to read the equation is to acknowledge that we don't know the data generating process, but nonetheless, we'd like to estimate β —the average linear relationship between Y and T , controlling for X .

Whatever the data generating process, OLS regression always gives us the *best linear approximation to the conditional expectation function* (BLACEF). If there are no baseline differences across values of T after controlling for X , then the BLACEF corresponds to the average effect of T on Y . In this case, knowing β is very valuable.

As you've seen before, when we “run a regression,” we're estimating the values of α , β , and γ that give us the best predictions of Y . For any given α , β , and γ , our prediction of Y_i for an individual i is then

$$\alpha + \beta \cdot T_i + \gamma \cdot X_i.$$

The error associated with each prediction can be written as

$$\text{Error}_i = Y_i - (\alpha + \beta \cdot T_i + \gamma \cdot X_i).$$

As we learned in Chapter 5, OLS regression finds the values of α , β , and γ that minimize the sum of the square of these errors across our entire sample. Our computers can do this really quickly.

Suppose we know that once we control for X , there are no other omitted confounders. So the regression of Y on T and X gives an unbiased estimate of the effect of T on Y . One question we might ask is how biased our results would have been if we failed to control for X .

It turns out, we can answer that question. Call Equation 10.1 above the *long regression* because it includes X . Suppose we ran the following *short regression* instead:

$$Y = \alpha^S + \beta^S \cdot T + \xi. \tag{10.2}$$

There's no guarantee that β^S will be the same as β . In fact, they won't be the same if X is a confounder.

We can quantify the bias associated with failing to include X in the regression. Consider a third regression equation:

$$X = \tau + \pi \cdot T + \mu.$$

This is a regression of the control variable on the treatment variable. The coefficient π is a measure of the correlation between T and X —it's the slope relating changes in T to changes in X . Of course, this need not have a causal interpretation.

It turns out that the bias associated with excluding X from the regression is

$$\beta^S - \beta = \pi\gamma.$$

We sometimes call this *omitted variable bias*. In words, this means that the short regression will be biased if the control variable is correlated with the treatment variable (π) and the control variable influences the outcome variable (γ).

If we can't observe X , we can't include it in the regression, but the omitted variable bias formula gives us a way to think about the direction and extent of the bias. Indeed, the omitted variable bias formula formalizes our ideas from Chapter 9 about how to sign the bias. If there's an unobserved confounder that we suspect is positively related to both T and Y , the omitted variable bias formula tells us that $\beta^S - \beta > 0$, so we are overestimating the effect of T . If the confounder is negatively related to both T and Y , again, the bias is positive and we are getting on overestimate. By contrast, if the confounder is positively related to T but negatively related to Y , or vice versa, the bias is negative and we underestimate the effect of T .

Controlling for Confounders versus Mechanisms

Thinking clearly about controlling gets even dicier when there is some variable that affects the outcome both affects and is affected by the treatment. What can we do in this situation? That variable is a confounder: it affects both the treatment and the outcome. So it seems that we should control for it. But as we discussed in Chapter 9, that variable is also a mechanism: it is affected by the treatment and affects the outcome. So it seems that we should not control for it, since it is part of the pathway by which the treatment effects the outcome. What are we to do?

To make this conundrum more concrete, let's return to our example of trying to learn about the effect of per capita GDP on civil war. On the one hand, democracies might implement better policies that improve GDP and might also provide better opportunities for non-violent expression of political grievances, which might directly affect civil war risk. From this perspective, whether a country is a democracy or not is a confounder and, thus, should be controlled for. On the other hand, perhaps as a country becomes richer, its citizens become more informed and start demanding greater democracy, which then reduces the likelihood they turn to civil war. From this perspective, democracy is one of the mechanisms by which GDP affects the likelihood of civil war and, thus, should not be controlled for.

There really isn't a solution in such situations. You are damned if you do and damned if you don't. Which means that you aren't in a position to learn much about the causal relationship you are interested in. To do that, you'd need a more creative approach, which will be the subject of the next two chapters.

There is No Magic

People would really like to believe that they can learn about causal relationships by just controlling for all the confounders. And they'd like you to believe it too. But, as we've just discussed, in many important settings you will only believe there are no omitted confounders if you aren't thinking clearly. And so, sometimes, people will use mathematical jargon, cool sounding statistical methods, complicated computer programs, and other technical wizardry to try to get you to not think clearly. It is important to not be fooled. No matter what fancy techniques an analyst uses, if the fundamental strategy is to control for the confounders, and there are plausible confounders that are either unobservable or unmeasured in the data, they can't possibly have controlled for them. Computers aren't magical. They can control for observable confounders. But they cannot make unobservable confounders observable.

To see what we mean, consider the case of a perfectly fine and useful statistical technique called matching. Here's the idea. Suppose you have a continuous variable, X , that you'd like to control for. You could match each treated unit to whatever untreated unit has the most similar value of X . Then, you might compute a difference of means (or run a regression) on that matched dataset in order to estimate the effect of T on Y . This is called "nearest neighbor matching".

So, in our legislative politics example, you would start by matching each Democratic congressperson to the Republican congressperson with the most similar NPAT score. Then, in this matched sample, you would compute the difference in average ACU score between matched pairs of Republicans and Democrats, which would give you the relationship between ACU score and political party, controlling for NPAT score.

If you had multiple variables you wanted to control for, you'd have to define some summary measure of how similar any two observations are across those variables. There are lots of strategies for doing this. Some of them get pretty fancy in terms of computation, which can make it hard to keep thinking clearly. But you have to try to keep your wits about you.

Matching has some advantages over regression as a technique for controlling. One nice advantage is that it allows for much more flexibility in the way in which the control variable might influence the outcome of interest. For instance, whereas regression assumes the relationship is linear, matching makes no such assumption. Matching also has disadvantages relative to regression. One downside of matching is that it's often less precise than regression because you're using less information. Another downside is that matching estimates can be biased because the best match for a treated observation will have, in expectation, a higher value of X if, for example, X is positively correlated with T . There are statistical solutions to that problem as well, which can again start looking pretty technical and fancy.

As we say, matching, like regression, is a good statistical technique for controlling. We

have no objection to it. We get concerned because analysts sometimes like to present some very technical matching algorithm and then say things like “matching creates an experiment-like comparison of units that differ in the treatment but are otherwise the same.” Such claims are an attempt to blind you with science. Matching is just a tool for controlling. It creates no more of an experiment-like comparison than does a regression that includes control variables. That is to say, it controls for the few variables that were observed and matched on—nothing more. Your computer, no matter how fancy the statistical algorithm, can’t make the unobservables observable. Because that would be magic. And there is no magic.

Key Terms

- **Controlling:** Using a statistical technique to find the correlation between two variables, holding the value of other variables constant.
- **Dummy Variable:** A variable that indicates whether a given unit has some particular characteristic, taking a value of 1 if the unit has that characteristic and 0 if the unit does not.
- **Dependent or Outcome Variable:** The variable in your data corresponding to the feature of the world that you are trying to understand or explain with your regression.
- **Treatment Variable:** The variable in your data corresponding to the feature of the world whose effect on the dependent variable you are trying to estimate.
- **Control Variable:** A variable in your data corresponding to a potential confounder that you include in a regression in an attempt to reduce bias in your estimate of a causal effect.

- **Omitted Variables Bias:** The bias resulting from failing to control for some confounder when attempting to estimate a causal effect.

Exercises

1. This problem uses the data at

`home.uchicago.edu/~bdm/HouseElectionsSpending2018.csv`

and described at

`home.uchicago.edu/~bdm/spending_readme.txt`.

- (a) Run a regression of incumbent vote share (your dependent variable) on both incumbent spending and challenger spending.
 - i. Note that if challenger spending is positively correlated with higher challenger vote shares, it must be negatively correlated with incumbent vote share. In light of this, how should we interpret the estimated coefficients associated with your independent variables?
 - ii. Are the results you obtained different from those you obtained when you ran separate regressions of incumbent vote share on incumbent spending and incumbent vote share on challenger spending in Chapter 9? Why or why not?
- (b) Let's add some controls to your regression in an attempt to obtain more reliable estimates of the effect of campaign spending. As you may know, 2018 was a good year for Democrats in House elections.
 - i. Is the overall good performance of Democrats in 2018 a potential confounder

- in your regression?
- ii. Create a new variable indicating whether the incumbent is a Republican—call it `RepublicanIncumbent`. It should take a value of 1 if the incumbent is a Republican and a value of 0 if a Democrat is the incumbent.
 - iii. Rerun your regression, but include that variable as a control.
 - iv. Interpret the estimated coefficient associated with your new `RepublicanIncumbent` variable.
 - v. Did including this control variable meaningfully change your estimated coefficients of interest (i.e., the coefficients on incumbent and challenger spending)? Why or why not, do you think?
- (c) Now add in a control for the vote share the incumbent's party received in that district in the 2016 presidential election.
- i. What kind of concern might including this control variable address?
 - ii. Interpret the estimated coefficient associated with this control variable?
 - iii. Does including this control variable meaningfully change your estimated coefficients of interest (i.e., the coefficients on incumbent and challenger spending)? Why or why not, do you think?
2. Produce a regression table that shows the results of each of the regressions from Question 1, along with the number of observations and the R -squared.

Notes

The study on Facebook and subjective-well being is

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. Forthcoming. "The Welfare Effects of Social Media." *American Economic Review*.

Chapter 11

Randomized Experiments

What You'll Learn

- Randomization yields unbiased estimates of causal effects.
- All the tools of statistical inference and hypothesis testing work in experimental settings for adjudicating between genuine effects and noise.
- Even with a randomized experiment, numerous complications can arise and must be planned around.
- Sometimes people don't comply with their experimental assignment. When this happens, it is important to keep making comparisons based on randomized assignment.

We love regression, so we had a pretty good time with Chapter 10. But we can see how the message that you can basically never get unbiased estimates of causal relationships just by controlling for confounders might be something of a downer. We are going to try to make it up to you in the next couple of chapters. We'll do so by showing you that there are better ways to learn about causality. Those ways are called *research designs*. Using research designs to learn about causality often involves quite a bit of cleverness and creativity. This makes research design among the most fun topics you must master in order to think clearly in a data driven world.

In this chapter, we consider the research design called a randomized experiment. Randomized experiments are great because, if you can randomize treatment, you can in fact create a world in which there are no omitted confounders. So you can eliminate bias from your estimates. An analogy to randomized experiments also helps to explain why the approaches we consider in the next two chapters are called research designs. When you run a randomized experiment, you literally get to *design* the way in which treatment is assigned.

In Chapters 12 and 13, we turn to other research designs that are a little less “designy”. In particular, those research designs are ways of trying to learn about causality from data that you observe “in the wild”—i.e., where the world, rather than an experimenter, decided how treatment would be assigned. But before we go there, let's spend some time learning about experiments and how they work.

Breastfeeding

At the time of this writing, it is a virtual article of faith in the developed world that babies should be breastfed. Consider, for example, this official statement from the World Health Organization: “Adults who were breastfed as babies often have lower blood pressure and lower cholesterol, as well as lower rates of overweight, obesity and type-2 diabetes.

There is evidence that people who were breastfed perform better in intelligence tests.” Similarly, in 2011, the Surgeon General of the United States issued a call to action to support breastfeeding, which she said is “one of the most highly effective preventive measures a mother can take to protect the health of her infant.” The accompanying report claimed links between breastfeeding and a host of childhood scourges, including ear infection, eczema, diarrhea, respiratory disease, asthma, obesity, type 2 diabetes, leukemia, and sudden infant death syndrome.

Indeed, an enormous scientific literature documents the positive association between breastfeeding and good health outcomes for children. Given this evidence, it seems like any caring mother must breastfeed her child.

But before you jump to conclusions, consider this fact. In developing countries, breastfeeding seems to be correlated with worse, not better, health outcomes for children. In countries as diverse as Ghana, Kenya, Egypt, Brazil, Peru, Bolivia, and Thailand breastfeeding has been found to be correlated with malnutrition and decreased height and weight.

What is going on? Is it possible that breastfeeding is good for kids in the industrialized world and bad for kids in the developing world? Maybe, but let’s make sure we are thinking as clearly as possible. You have already learned that correlation does not necessarily imply causation. And in this case, the comparison of mothers who do and do not breastfeed is most likely not an apples-to-apples one.

First, think about the developing world, where breastfeeding is negatively correlated with children’s physical wellbeing. One possibility is that breastfeeding causes these adverse outcomes. It’s also possible that some confounding factor, like poverty, causes both breastfeeding and malnourishment. Breastfeeding costs a lot in terms of the mother’s time, but it doesn’t cost much money. Formula, by contrast, costs a lot of money but less time. So we might expect that economically distressed families are more likely to breastfeed their children. And those same economically distressed families may be more prone to health

problems for reasons entirely unrelated to breastfeeding. Reverse causality is also a concern. Perhaps adverse health in a child directly leads a mother to be more likely to breastfeed.

Indeed, confounding factors and reverse causality appear highly relevant in the developing world. A 1997 study published in the *International Journal of Epidemiology* tracked the health outcomes of 238 toddlers (ages 12-15 months) in a village in Peru. The study's data included information on child size, breastfeeding, complementary food intake, and diarrhea. The study found a negative correlation between breastfeeding and height—children who were breastfed were smaller on average, suggesting they were in poorer health. This relationship was strongest among those children who were getting the least complementary food and were the most sickly. It turns out that, because breastfeeding is widely believed to have health benefits, mothers whose children were sickly or lacked access to complementary food weaned their children later. Consequently, children who were already sick and malnourished were more likely to be breastfed. Thus, the study concludes, it is not the case that breastfeeding causes children not to grow in the developing world. Rather, children who are not growing because they are sick and malnourished are more likely to be breastfed.

Now think about the developed world, where parents are inundated with the message that breastfeeding is good for their children. Remember, breastfeeding is said to reduce the risk of heart disease, asthma, obesity, leukemia, SIDS, ear infections, and a host of other ailments. Unfortunately, the evidence underlying this conventional wisdom once again doesn't withstand much scrutiny. Surely you can think of lots of reasons why comparing breastfed children to children who are not breastfed in the developed world also might not be an apples-to-apples comparison. For instance, once official organizations issue statements on the efficacy of breastfeeding, we would expect wealthy, educated mothers to be particularly likely to hear this news and follow the advice. But their children were likely to have better health outcomes anyway.

Breastfeeding or not is such a high stakes decision, and so many different factors influence

this decision, that it seems impossible to ever find an apples-to-apples comparison out there in the world. However, perhaps we could generate our own apples-to-apples comparison through a randomized experiment. A team of researchers in Belarus tried to do exactly this.

The team's strategy was to run a randomized experiment. Clearly, for both ethical and practical reasons, they couldn't force mothers to breastfeed or not to, just for the benefit of their study. But they could make it more likely that a randomly selected group of mothers would choose to breastfeed through randomly assigned encouragement. To achieve this, in some randomly selected hospitals the researchers implemented a program to encourage and facilitate breastfeeding. In other randomly selected hospitals, they did not implement this program. For all the hospitals, they recorded how children were fed and tracked a variety of the children's health (and other) outcomes over time. And, indeed, mothers in the hospitals that had the breastfeeding program were much more likely to breastfeed their newborns.

Given the claims from the World Health Organization, the Surgeon General, the American Academy of Pediatrics, and the parenting industry more generally, they found surprisingly scant evidence for the large benefits breastfeeding is supposed to provide. Babies from the hospitals that received the program were slightly less likely to have eczema and gastrointestinal infections, but the researchers obtained null results for many more outcomes. In a follow up study, conducted when the children were between six and seven years old, the investigators explored whether the children who were more likely to be breastfed performed better on any observable physical, psychological, or cognitive outcomes. They found no evidence that breastfeeding provided any benefits in terms of the risk of eczema, allergies, asthma, obesity, emotional problems, conduct problems, hyperactivity, or peer problems. Indeed, if anything, the evidence went the other way, showing some limited evidence of a negative association between breastfeeding and these outcomes. The one piece of evidence they found in support of breastfeeding was that children from the hospitals that received

the breastfeeding program performed slightly better on IQ tests. Of course, as you learned in Chapter 7, if you look at that many outcomes, you're pretty likely to find at least one statistically significant finding just because of noise.

Overall, our view is that the experimental evidence is not nearly strong enough to encourage every mother around the world to breastfeed. Although there are very strong correlations between breastfeeding and health outcomes in various settings, and some reasonable arguments about the biological mechanisms through which breastfeeding might work, the best available evidence suggests that the average effect of breastfeeding is likely small. (It is worth noting that the evidence is stronger in certain sub-populations—e.g., premature babies.) This would have been much more difficult to discover without the power of randomized experimentation.

Randomization and Causal Inference

Why are experiments such a powerful tool? Suppose we want to know the effect of some treatment, say Body Vibes, on some outcome of interest, say skin health. In general, it is difficult to estimate the effect of body vibes because of all the issues discussed in previous chapters. We want to know how different a person's skin would be in the world in which they use Body Vibes versus the world in which they do not use Body Vibes. Unfortunately, for any given person, we only get to observe one of those potential outcomes. For example, if a person uses Body Vibes, we can observe their skin health in that situation, but we don't know what their skin would be like if they hadn't used Body Vibes.

As we discussed in Chapter 9, if we just compared the average skin health of people who do and don't use Body Vibes, that's not an apples-to-apples comparison. That is, there are a variety of confounders that imply that this difference in means is not an unbiased estimate of the average treatment effect. For instance, perhaps those who use Body Vibes just care

more about their skin, and they also use more moisturizer and sunscreen. Or maybe the bias goes the other way. Perhaps the people who use Body Vibes have bad skin, they've tried everything else, and they're getting desperate. Either way, because of such confounders, we can't get an unbiased estimate of the effect of Body Vibes just by comparing the average skin health of people who do and do not use them.

One way, perhaps the best way, to get rid of this bias and be sure that we're making an apples-to-apples comparison is to randomize the treatment. Our comparison of those using and not using Body Vibes is biased because these groups, on average, would likely have different skin health even if none of them used Body Vibes. However, if we randomly assign people to use or not use Body Vibes, then those two groups would, in expectation, be the same in terms of their pre-existing skin health and all other prior characteristics. That is, there would be no confounders. Why is this the case?

If a treatment of interest is determined by the flip of a coin, a random number generator on your computer, or another random process, then the only thing that distinguishes people in the treated group and the untreated group is pure chance. There is no reason why people in the treated group should be systematically taller, smarter, richer, more motivated, or have better skin than those in the untreated group before we deliver the treatment.

In terms of our potential outcomes notation, suppose we randomly assign some group (the treated group T) to receive Body Vibes and another group (the untreated group U) not to receive them. We observe the average skin health of the treated group with Body Vibes, $\bar{Y}_{1T}$. And we observe the average skin health of the untreated group without Body Vibes, $\bar{Y}_{0T}$. Thus, if we compare the average skin health in the two groups we get the difference in means:

$$\bar{Y}_{1T} - \bar{Y}_{0T}.$$

But, because of randomization there are no systematic differences between either of these

groups and the whole population. Thus, in expectation the average skin health of the treated groups while wearing Body Vibes is equal to the average skin health of the whole population if they were to all were Body Vibes

$$\mathbb{E}[\bar{Y}_{1T}] = \bar{Y}_1.$$

And similarly for the untreated group

$$\mathbb{E}[\bar{Y}_{0U}] = \bar{Y}_0.$$

Hence, the difference in means is an unbiased estimate of the average treatment effect:

$$\mathbb{E}[\bar{Y}_{1T} - \bar{Y}_{0U}] = \bar{Y}_1 - \bar{Y}_0.$$

We, of course, don't have to put this in terms of expectations. We can instead talk about the observed difference in means being an unbiased estimate of the average treatment effect—which is to say that the observed difference in mean outcome between the treated and untreated groups is just the ATE plus noise. Thus, the following is equivalent to the statement about expectations above:

$$\underbrace{\bar{Y}_{1T} - \bar{Y}_{0U}}_{\text{Observed Difference in Means}} = \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{ATE}} + \text{Noise}.$$

As in the examples from previous chapters, noise can come from sampling variability. Perhaps there's a broader population about which we care, and we happened, by chance to get an unusual sample of subjects in our experiment. But now, when we're doing an experiment, noise also comes from the random assignment of the treatment to subjects. Even for the same sample of subjects, different randomizations could have produced different

estimates, and this also contributes to the noise.

Because of noise, for any small scale experiment, there will be some differences in average potential outcomes between the treated and untreated groups, just by chance. But those differences won't be systematic—if we were to run many iterations of the experiment, we would not expect to find the same pattern of differences repeated over and over. This is what we mean by the phrase “in expectation” in the preceding discussion.

Think back to our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

Randomization guarantees that the bias is zero. So the only reason that the estimate we get from comparing the mean outcome in the treated and untreated groups in a properly randomized experiment differs from the true causal effect (i.e., the estimand) is because of noise.

As we increase the number of people in any given experiment, we expect the two groups to become more and more similar. That is, as the sample size gets big, the noise becomes small.

Randomization gives you comparability of the treated and untreated group in expectation—generating unbiased estimates. A large sample size gives you a very small amount of noise—generating precise estimates. So randomization plus a large sample size gives you comparability in your actual, realized sample—generating estimates that are very likely to be close to the true estimand.

If you think about it for a minute, you'll realize that randomization is essentially the only way to guarantee unbiased estimates of causal relationships. Suppose you tried to conduct an experiment, but instead of randomly assigning subjects to the treated and untreated groups, you tried to carefully divide the groups so that they were as similar as

possible to each other. Since you can't possibly observe and quantify all of your subjects' relevant characteristics, you'd have to make some judgment calls. Maybe you'll do this really well, and maybe you won't. What if your own subconscious biases lead you to put slightly different people in the treated and untreated groups—perhaps because you're subconsciously hoping the experiment will show a big effect? You'll have no way of knowing whether you actually did a good job. Therefore, why take the risk? Why not actually flip coins and assign the treatment randomly? If this point seems obvious now, it wasn't obvious to a lot of smart people in the past. It's only since the work of R.A. Fisher in the 1920s that scientific researchers understand the value of randomization.

There is one way in which this thought about trying to make the treated and untreated groups as similar as possible on observable characteristics does make some sense. As we know from our favorite equation, there are two ways our estimates might differ from the true causal effect: bias and noise. Randomization eliminates bias. But there could still be lots of noise, especially if the sample size is small or the experimental subjects are quite different from one another on characteristics that matter for the outcome. That is, in any given iteration of an experiment, the treated and untreated groups could end up looking very different *in reality*, even though they are the same *in expectation*.

One thing you can do to reduce this problem, is to start by grouping people based on their observable similarities. Then you can randomly assign individuals to be treated or untreated within those groupings. For example, you might be concerned that Body Vibes differentially affect men and women. It would introduce a lot of noise into your experiment if, by chance, you ended up with a treated group made up of mostly men and an untreated group made up of mostly women, or vice-versa. This won't happen in expectation (i.e., if you did the experiment an infinite number of times, the average proportion of men and women will be the same in the treated and untreated groups). But it could happen in any given iteration of your randomization. To eliminate this source of noise, you could start by dividing your

experimental population by biological sex. Then you randomly assign half the male group to be treated and half to be untreated, and likewise for the female group. You'd still have randomized treatment assignment, so you'd still get unbiased estimates. But you'd also have reduced the noise by making it more likely that your treated and untreated groups were similar in terms of biological sex, not just in expectation, but in reality. Experimentalists call dividing up the sample like this prior to randomization *blocking* or *stratification*. We'll talk more about it later in the chapter.

Estimation and Inference in Experiments

In Chapter 6, we discussed statistical inferences about relationships. All of those lessons apply in the case of an experimental estimate as well. In the simplest scenario, we can analyze the results of an experiment by calculating a difference in means, i.e., comparing the average outcome in the treated and untreated groups. In fact, as we saw in Chapter 5, if we regress the outcome on a binary measure of treatment status, the coefficient associated with the treatment variable is just the difference in means. And since regression coefficients and differences in means are just quantitative relationships, we can apply all of the statistical tools of Chapter 6 to experiments as well.

Suppose we conduct a randomized experiment and estimate the average treatment effect by comparing the average outcome for subjects with the treatment to the average outcome for subjects without the treatment. This estimate is unbiased. But it might be imprecise (i.e., there may be lots of noise).

We'd, of course, like to know how close our estimates are to the true causal effect (i.e., the estimand). We can estimate the standard error associated with our experimental estimate just like we estimated the standard error of poll results and regression coefficients in Chapter 6. The standard error gives us a sense of how far, on average, our estimate would

be from the truth as a result of noise, if we repeated our experiment an infinite number of times, each time using the same procedure to generate an estimate of the treatment effect. Similar to our discussion of poll results, the true standard error actually depends on quantities that are unobservable, but there are various approximations that practitioners use for estimating the standard error.

You don't need to memorize formulas for standard errors; you can always look them up or just let your computer calculate them for you. Nonetheless, it's useful to think about how various features of experiments influence the amount of noise. Suppose we conduct an experiment with N subjects, of whom m receive the treatment and $N-m$ do not. All else equal, the greater is N , the less noise and, thus, the smaller the standard error. This should be intuitive. When the sample size is larger, the treated and untreated groups will be more similar to each other with respect to other characteristics, reducing noise, and making our estimates closer to the true causal effect.

What about m ? Suppose we have 500 people in our study. How many of them would we like to put in the treated group versus the untreated group? Obviously, we can't put all of them in either group because then we wouldn't be able to make a comparison (remember that correlation requires variation). Extending that logic, we can see that we don't want either too few or too many subjects in either group. If either the treated group or the untreated group is very small, then our estimates will be highly noisy because the average outcome for whichever group is small will be quite sensitive to idiosyncratic features of just a few subjects. Typically, then, you'll get the most precise estimates when the size of the treated and untreated groups are roughly equal.

With that being said, there are often cases where an optimal experimental design might have different numbers of subjects in each condition. Suppose you have 100,000 potential subjects, you only have enough resources to put 100 people in your treated group, but it's costless to put more subjects in your untreated group. You might as well randomly assign

100 people to the treated group and everybody else to the untreated group. Your estimates won't be nearly as precise as if you had 50,000 people in each group, but they'll be much more precise than an experiment with 100 people in each group.

The last factor that influences the noisiness of experimental estimates is how variable the outcome is in both the treated and untreated groups. If we study outcomes that vary very little within each treatment condition, our estimates will be more precise than if we study outcomes with greater variance. This is because, if the outcome varies very little based on non-treatment characteristics, then there is very little scope for noise—we'll get similar outcomes for each group across iterations of the experiment. This is why, for example, doctors and government regulators often have precise estimates of the effect of some heart medication on blood pressure (a relatively low variance outcome) but imprecise estimates of the effect of the same drug on heart attacks (a high variance outcome). Of course, sometimes, we have no control over this. The outcome of interest is what it is. And sometimes the most interesting or important outcomes (e.g., heart attacks) are high variance. But other times it might be possible to identify outcomes or methods for measuring those outcomes that reduce noise (e.g., blood pressure, which is known to play a role in heart attacks). When this is feasible, it is worthwhile.

We can also apply the tools of hypothesis testing that we learned in Chapter 6 to experimental results to assess statistical significance. For instance, dividing the estimate by the standard error generates a t-statistic, which can be used to estimate a p-value. And because we often do hypothesis testing with experimental results, we need to keep thinking clearly about the risks of over-comparing, under-reporting, mean reversion, and cosmic habituation! To reduce these risks, analysts should pre-specify the questions of interest the experiment is designed to address, the empirical tests they plan to conduct (so they can't just go fishing for a statistically significant finding), and report the results regardless of what they find. We should also interpret experiments with these issues in mind. If

analysts are not transparent about the steps they took to avoid over-comparing and under-reporting, we should be skeptical of their findings. And, the more surprising the findings, the more skeptical we should be. Remember that the ESP result arose in an experimental study! Or, more seriously, think again about the breastfeeding experiment with which we began. That study had many virtues. But one potential problem is that, because the study designers collected information on so many outcomes, when we see no evidence of an effect of breastfeeding on eczema, allergies, asthma, obesity, emotional problems, conduct problems, hyperactivity, or peer problems, but we do see an effect on IQ, we are worried that the apparent effect on IQ arose just by chance.

Problems That Can Arise in Experiments (and Other Research Designs)

Things rarely work as beautifully in practice as they do in our idealized examples. In theory, you can design a randomized experiment and estimate the average treatment effect simply by comparing means. In practice, however, problems arise that make analysis and interpretation less straightforward. Let's discuss some of those problems and the ways in which careful analysts can deal with them.

Noncompliance and instrumental variables

One common problem in experiments is that subjects fail to comply with their assigned treatment. For instance, it is pretty common in medical studies for some subjects to simply stop taking their medication. It also happened in the breastfeeding experiment. Recall, because it is unethical to force a mother to breastfeed or not, the researchers randomly assigned mothers into groups where they received more or less encouragement to breastfeed. Encouragement designs like this allow researchers to experimentally study lots of topics that

would otherwise be off-limits for logistical or ethical reasons. But they inevitably involve the additional complications that arise from noncompliance, since surely some mothers who were encouraged nonetheless did not breastfeed and some mothers who were not encouraged did breastfeed.

Suppose we designed a randomized experiment to estimate the effect of Body Vibes on skin health. We randomly assigned some individuals to the treatment condition—we gave them Body Vibes and, for the sake of science, tried to convince them to wear them. We also randomly assigned some individuals to an untreated condition where they were given no Body Vibes and told to go about their normal lives. Then, despite our best efforts, some of the subjects in the treated group forget or simply refuse to wear their Body Vibes. And amazingly, a few of the more gullible members of the untreated group hear about Body Vibes elsewhere, spend their hard-earned money on the product, and wear Body Vibes. Shoot! What do we do?

One idea would be to simply compare people who did and didn't wear Body Vibes, ignoring whether each subject was initially assigned to the treated or untreated group. But this won't work. It brings us right back to the problem we were trying to solve through our experiment. The people who voluntarily wear or do not wear Body Vibes are likely different from one another, so a comparison of those to groups is not an apples-to-apples comparison.

Another idea would be to drop the subjects that did not comply with their treatment assignments. In other words, we could remove from our analysis the people who were assigned the treated group but didn't wear Body Vibes and the people who were assigned to the untreated group but did wear them. After doing that, we might just proceed as normal, comparing the mean skin health among the remaining members of the treated and untreated groups.

Unfortunately, this is still not an adequate solution. To see why, think about the people

who were in the treated group but refused to wear Body Vibes. They might be special in important ways—e.g., have better skin health, be less gullible. Presumably, there were also people just like them in the untreated group. But, because we didn't ask those people to wear Body Vibes in the first place, we can't figure out who they are. So we can't similarly remove them from the untreated group. As such, if we throw this group of people out of the treated group, the comparison of the treated and untreated groups is no longer an apples-to-apples comparison. The kinds of people who wouldn't wear Body Vibes even if given them would be present in the untreated group but not the treatment group.

So what *can* we do in light of noncompliance? Well, one thing we can always do is estimate the effect of being assigned to the treatment group (as distinct from the effect of the treatment itself). We do this by comparing the outcomes for the people assigned to the treated and untreated groups, regardless of whether they actually took up the treatment. This comparison won't give us an unbiased estimate of the effect of wearing Body Vibes. But it will give us an unbiased estimate of the effect of being given Body Vibes and encouraged to wear them. We sometimes call this the *intent-to-treat effect*.

There are situations where a policymaker or decision maker actually cares more about intent-to-treat effects than actual treatment effects. Suppose a charitable organization is trying to decide whether it should provide free Body Vibes to high school kids with bad skin. They know that not everyone provided with Body Vibes will wear them. Furthermore, all they can do from a policy perspective is provide the Body Vibes; they can't force anyone to use them. They conduct an experiment to estimate the benefits of free Body Vibes. What quantity should go on the benefits side of their cost-benefit analysis to inform them about whether this is a good policy? It's not the average effect of Body Vibes for an individual who uses them. It's the average effect of being provided Body Vibes, regardless of whether an individual uses them or not, since this is what the charitable organization can actually do. So the intent-to-treat effect is the relevant number. More seriously, in many settings,

all a policy maker or organization can do is provide a service, they can't force people to take it up. In any such situation, the intent-to-treat effect may in fact be the most important quantity.

In other situations, however, we are interested in the actual effect of the treatment, not just the intent-to-treat effect. Suppose, for instance, that we're trying to decide whether we should wear Body Vibes ourselves. Or, more seriously, suppose someone is deciding whether to try an experimental medical treatment, a new study regimen, a new teaching technique, or a new productivity enhancing management strategy. In those case, we want to know more than just the intent-to-treat effect. We want to know the likely effect of taking up the treatment. So what more can we do with our experimental results, plagued as they are by issues of non-compliance?

To make some additional progress, let's think about the different ways a subject's decision to wear Body Vibes can respond to our experimental encouragement. Our sample consists of up to four different kinds of people.

1. There are **compliers** who will wear Body Vibes if they're assigned to treatment and will not wear them if they're not assigned to treatment.
2. There are **always-takers** who will wear Body Vibes regardless of whether we provide them or not.
3. There are **never-takers** who will not wear Body Vibes regardless of what we do in the experiment. (We are both never-takers when it comes to Body Vibes.)
4. And, in principle, there could be a perverse group of **defiers** who won't wear Body Vibes if they're in the treated group but will wear Body Vibes if they're in the untreated group.

Obviously, when we do an experiment, we're hoping for lots of compliers. The whole

idea of an experiment is that we want to randomly assign treatment, and the compliers are those subjects who are willing to let us do that.

Every subject in an experiment fits neatly into one (and only one) of these categories. However, we can't just look at our experimental subjects and figure out which people are compliers, always-takers, never-takers, or defiers. Why is that? Suppose we see that someone is in the untreated group and didn't wear Body Vibes. We know that they are either a complier or a never-taker. But we have no way of knowing which, because we don't know whether they would have worn Body Vibes if they were in the treated group. More generally, we can think about these categories within our Body Vibes experiment as follows:

	Treated Group	Untreated Group
Wore Body Vibes	Compliers & Always-Takers	Always-Takers & Defiers
Didn't Wear Body Vibes	Never-Takers & Defiers	Compliers & Never-Takers

Dividing people up into these groups helps us think clearly about when we are or are not making an apples-to-apples comparison. In particular, in order to insure that we don't have confounding, we want the groups we compare (say, treated and untreated groups) to have the same share of compliers, always takers, never takers, and defiers as one another.

To get a sense of how this helps us understand the problem, let's start by assuming that everyone is either a complier or a never-taker. In other words, none of those people who might buy and wear Body Vibes on their own happened to participate in our experiment. (We'll relax this in a bit.) Now let's revisit the various ways we might deal with experiment subjects who don't behave according to their experimental treatment assignment. In a world with only compliers and never-takers, our experiment looks like this:

It's easy to see why we can't just compare people who did and didn't wear Body Vibes, ignoring their treatment assignment. The group that wears Body Vibes is made up of

	Treated Group	Untreated Group
Wore Body Vibes	Compliers	N/A
Didn't Wear Body Vibes	Never-Takers	Compliers & Never-Takers

just the compliers in the treated group. The group that doesn't wear Body Vibes is a combination of the compliers in the untreated group, the never-takers in the untreated group, and the never-takers in the treated group. So the comparison of Body Vibe wearers to non-Body Vibe wearers is not apples-to-apples.

Similarly, it's easy to see why we can't just drop the people who visibly didn't comply with our experiment. We would drop the never-takers from the treated group. But we wouldn't drop anyone from the untreated group. As a result we'd be comparing the compliers from the treated group to a combination of the compliers and the never-takes from the untreated group—again, not an apples to apples to comparison.

It seems like we're still stuck in a place where all we can do is compare the treated and untreated groups, estimating the intent-to-treat effect. But, actually, we can do better. Let's see how.

A key step in doing better involves estimating the proportion of compliers in our sample. We don't know exactly who the compliers are. But, in our simplified example with only compliers and never-takers, we can estimate what proportion of the sample is compliers. We do so by calculating what proportion of the treated group in fact takes up the treatment. This is the proportion of compliers in the treated group. And, because of random assignment, the treated and untreated groups have the same proportion of compliers in expectation. Therefore, the proportion of compliers in the treated group is an unbiased estimate for the proportion of compliers in the whole sample (i.e., the treated group and untreated group combined).

We now have unbiased estimates of both the intent-to-treat effect (by comparing the average outcomes in the group assigned to be treated and the group assigned to be untreated) and the proportion of compliers in the sample. How does that help us?

We want to know the effect of Body Vibes on some outcome like skin health. If we assume that the only way that treatment assignment could have influenced skin health is through the actual use of Body Vibes, then what is the intent-to-treat effect? Well, no one in the untreated group used Body Vibes, so their skin health is just what it would have been at baseline. In the treated group, the compliers used Body Vibes and the never-takers did not. So the average skin health in the treated group is a weighted average of the compliers after using Body Vibes and the never-takers at baseline, with weights given by the proportion of compliers and never-takers in the treated group. Hence, when we compare the average skin health in the treated and untreated groups, we are getting the effect of Body Vibes, weighted by the proportion of compliers. That is, the intent-to-treat effect is the true effect of Body Vibe usage for the compliers times the proportion of compliers. Therefore, we could divide the intent-to-treat effect by our estimate of the proportion of compliers to estimate the average effect of Body Vibes for the compliers.

Let's do a little example to see how this works. Imagine that Body Vibes actually work. (Remember, a lot of this book is about counterfactual worlds.) In particular, suppose that you could measure skin health on a scale of 1–10, with 10 being perfect skin and 1 being very bad skin.

Now let's imagine we conducted an experiment on 100 people to study the effects of Body Vibes. We randomly assigned 50 people to receive the treatment and 50 people not to. The people assigned to receive the treatment got Body Vibes. The other people did not. A month later, we measured the skin health of each person on our 1-10 scale. Suppose the data looked like this:

Our estimate from the data of the intent-to-treat effect is 1.6—i.e., on average, people

	People Assigned to Be Treated	People Assigned to Be Untreated
Average Skin Health	7.8	6.2

given Body Vibes had a skin health score that was 1.6 points higher than people not given Body Vibes.

You dig a little deeper and discover that, while no one in the group assigned to be untreated went and bought Body Vibes, only 40 of the 50 people assigned to treatment wore them. From this you estimate that the proportion of compliers in your sample is 80% ($\frac{40}{50}$) and the proportion of never-takers in your sample is 20%. You can now recover the true effect of Body Vibes on the compliers.

How does this work? To make sure we are thinking clearly, let's return to our potential outcomes notation. Call the average skin health of a complier without treatment (i.e., without Body Vibes) with Y_{0c} ; call the average skin health of a complier with treatment (i.e. with Body Vibes) Y_{1c} ; and call the average skin health of a never-taker without treatment Y_{0n} . Given that the sample is 80% compliers and 20% never-takers, we have the following two equations:

$$7.8 = 80\% \cdot Y_{1c} + 20\% \cdot Y_{0n}$$

and

$$6.2 = 80\% \cdot Y_{0c} + 20\% \cdot Y_{0n}.$$

The first equation says that the average skin health for those assigned to the treated group (7.8) is a weighted average of the average skin health of compliers with treatment (with weight 80%) and of never-takers without treatment (with weight 20%). Similarly, the average skin health for those assigned to the untreated group (6.2) is a weighted average of the average skin health of compliers without treatment (with weight 80%) and of never-

takers without treatment (with weight 20%).

We can subtract the left-hand sides of these two equations from one another and the right-hand sides of these two equations from one another to get

$$1.6 = 80\% \cdot (Y_{1c} - Y_{0c}).$$

The left-hand side is the intent-to-treat effect: the difference in average outcomes between the group assigned to be treated and the group assigned to be untreated. On the right-hand side, the 80% represents the proportion of compliers in the sample. And the term in parentheses is the average effect of the treatment for compliers (usually called the *complier average treatment effect* or CATE). So, we can recover the complier average treatment effect by dividing both sides by 80%:

$$\begin{aligned} \frac{1.6}{80\%} &= \overbrace{Y_{1c} - Y_{0c}}^{\text{CATE}} \\ &= 2. \end{aligned}$$

It is important to note the distinction between the complier average treatment effect and the overall average treatment effect. It is possible that wearing Body Vibes has the same effect on skin health for all the different kinds of people. If this is true, we say there are *homogenous treatment effects*. But, of course, it need not be true—Body Vibes could differentially effect the skin health of the kind of person who would never use them (never-takers) and the kind of person who uses them if given them (compliers). In this case, we say there are *heterogeneous treatment effects*. As the algebra above shows, dividing the intent to treat effect by the share of compliers recovers the complier average treatment effect. If there are homogenous treatment effects, the complier average treatment effect is the same as the overall average treatment effect. But if there are heterogeneous treatment effects,

they are not the same and we have to keep in mind that we are only able to estimate the average treatment effect for this specific part of the population. The intuition for why is straightforward. It is only the compliers who are actually changing their behavior in response to treatment. So they are the only part of the population about whom we are actually gaining information.

It was easy to see how all this works in a simplified world where everyone was either a complier or a never-taker. But we can do the same basic thing even if we move away from this simplified world and also allow for the possibility of always-takers. For now, let's continue to assume that there are no defiers because they muddy the waters. (There are lots of situations, including this hypothetical Body Vibes experiment, where we think that there should be few to no defiers.)

In this more complicated world, we can describe our experimental sample as follows:

	Treated Group	Untreated Group
Wore Body Vibes	Compliers & Always-Takers	Always-Takers
Didn't Wear Body Vibes	Never-Takers	Compliers & Never-Takers

How do we estimate the proportion of compliers when there are compliers, never-takers, and always-takers? First, the people in the group assigned to be treated who actually wear Body Vibes are either compliers or always-takers. So the size of this group gives us an estimate of the proportion of always-takers plus compliers. Second, the people in the group assigned to be untreated who wear Body Vibes are definitely always-takers. So the size of this group gives us an estimate of the proportion of always-takers. By subtracting this second number from the first, we get an estimate of the proportion of compliers. With that in hand, we can again proceed as above.

Therefore, our general procedure for estimating the complier average treatment effect

is as follows. Estimate the effect of being assigned to the treated group on the outcome of interest. As we’ve discussed, we call this the intent-to-treat (ITT) effect. It is also sometimes referred to as the *reduced form effect*. Also estimate the effect of being assigned to the treated group on the actual take-up of the treatment. Assuming there are no defiers, this gives us an unbiased estimate of the proportion of compliers. This is also sometimes called the *first stage effect*. We then recover an estimate of the CATE by dividing the intent-to-treat effect by the proportion of compliers. This ratio is called the *Wald Estimator*, after statistician Abraham Wald who first developed it, though in a different context.

The Wald Estimator is a special case of what is called *instrumental variables* (IV) analysis. This kind of analysis is appropriate when the treatment of interest is not randomly assigned, but there is some other variable (called an instrument) that (i) affects the treatment of interest, (ii) does not affect the outcome of interest except through the treatment, and (iii) is randomly assigned (or, there is some other way to credibly estimate its effect on the treatment and the outcome).

To be more precise there are 4 key conditions that must hold for IV analysis to work:

1. *Exogeneity*: The instrument must be randomly assigned or “as if” randomly assigned, allowing us to obtain unbiased estimates of both the first stage and reduced form effects.
2. *Exclusion Restriction*: All of the reduced form effect must occur through the treatment. In other words, there is no other pathway for the instrument to influence the outcome except through its effect on the treatment. If this isn’t the case, then the reduced form effect includes both the effect of the treatment on the outcome for compliers and these other pathways. Then, even after we divide by the first-stage, the resulting estimate still includes these other pathways and, thus, does not reflect the CATE.

3. *Compliers*: There must be some compliers.
4. *No Defiers*: If there are defiers, then our estimate will give us a weighted average of the average effect for compliers and the average effect for defiers, but with the defiers getting a negative weight (since their behavior changed in the wrong direction). How big a problem the presence of defiers is depends on how many of them there are. If there are very few defiers, then the bias that comes from their presence is negligible. But if they are a big part of the population, they are a big problem for the IV analysis.

In the case of our Body Vibes experiment, the experimental assignment to be treated or untreated was an excellent instrument. It clearly satisfied exogeneity because we randomized treatment. It also seems unlikely that being assigned to the treated group had any way of affecting skin health other than through Body Vibes, so it quite plausibly satisfied the exclusion restriction. So, as long as there were compliers (i.e., people who actually used the Body Vibes because they were assigned them) and no defiers, our analysis yielded an estimate of the complier average treatment effect.

There are more flexible ways to implement IV analysis than the Wald estimator. In particular, it can be implemented using regression, which is important because that allows us to accommodate control variables, if necessary, as well as situations with multiple instruments or treatments and instruments that are not binary.

Some analysts think of IV as a method or research design unto itself. For example, an analyst might implement our design above and say that they estimated the effect of Body Vibes using instrumental variables. That's technically true but misleading. The important research design in our example is the randomized experiment. We're using instrumental variables to deal with noncompliance, acknowledging the additional assumptions (above and beyond randomization) that doing so requires. In particular, the exclusion restriction is defensible in our example because all the experiment did was hand out stupid stickers to

some people and not to others. In other contexts, however, the exclusion restriction will be harder to justify and will require a lot of thought. We will return to this later, when we discuss natural experiments.

Chance imbalance

Randomization guarantees that the treated and untreated groups are, in expectation, the same in terms of prior to treatment. But the term *in expectation* is important. Just because two groups are the same in expectation, doesn't mean they are the same in actuality. As we've discussed, in any given experiment, the treated group could differ from the untreated group in lots of ways, just due to chance. This is why there is noise term, in addition to a bias term, in our favorite equation.

Experimenters often assess the *balance* between their treated and untreated groups by comparing them in terms of measurable pre-treatment characteristics. For example, in our Body Vibes experiment, we could compare the average age, gender, weight, diet, and skin health of the subjects in the treated and untreated groups before the treatments were delivered. We could even test for statistically significant differences. The hope, of course, is that we don't find any differences. If we do, we must worry that, even though our estimate is unbiased, it might nonetheless be quite different from the true effect because of noise.

What should a careful analyst do if, despite randomization, the treated and untreated groups turn out to differ in substantively or statistically significant ways? Let's consider three potential responses.

1. **Throw out the “broken” experiment.** You had good intentions when you ran the experiment, but you got unlucky and now you can't trust your results, so you should just forget the experiment and move on. Maybe you should do another one and hope for better balance.

We think this is an inappropriate response. Remember the problem of over-comparing. If you test for balance on enough pre-treatment variables, you are virtually guaranteed to find statistically significant imbalance on some of them. Therefore, by this logic, the more pre-treatment variables you can measure, the more likely you are to have to throw out the experiment, which seems perverse. Moreover, even “broken” experiments contain information. Importantly, they are unbiased (remember, bias is about getting the answer right on average, across lots of iterations of the experiment). And, so, they could be pooled with other evidence (perhaps from other iterations of the same experiment) and incorporated into a larger analysis that will ultimately contribute to knowledge.

Our response above assumes that the analyst is confident that the treatment was indeed randomly assigned. Our recommendations would change if this weren’t sure. Suppose you (or your computer) didn’t do the randomization directly. Instead, suppose you were running a large scale experiment and the randomization was implemented by a big team or by a partner organization. In a situation like this, if you detect enough imbalance, you might start to worry that your planned randomization wasn’t faithfully implemented. In that case, throwing out the experiment (probably following some investigation into whether your suspicions are well founded) could be appropriate.

2. **Proceed as normal.** Unbiasedness is a property *in expectation* so the experimental estimate is still unbiased. You could report the imbalance for the sake of transparency while still reporting the same estimated treatment effect that you originally planned. Of course, the treated and untreated groups are sometimes different by chance. That’s exactly why we report standard errors or other measures of noise.

This strategy may seem unsatisfying. As you’ll recall from Chapter 6 and our favorite equation, even an unbiased estimate can be very far from the truth. When we find

an imbalance between the treated and untreated groups that we think is strongly related to the outcome of interest, we might worry that this chance imbalance reflects us having gotten one of those draws of our procedure that results in an estimate that is far from the truth, despite the absence of bias. Nonetheless, there is still some merit to proceeding as planned and reporting our unbiased (if probably quite wrong) estimate. This is especially the case if we are talking about the kind of experiment that will be replicated lots of times, so that the lack of balance in any one iteration will be washed out in the long run through averaging across many iterations of the experiment.

But we also might wonder whether there is some way that we can account for the imbalance and generate an estimate which is likely to be closer to the right answer right now. Which leads us to our third possible response.

3. **Use the techniques discussed in Chapter 10 to control for any unbalanced variables.** As we learned in Chapter 10, controlling for variables, by soaking up the variance in the outcome that is due to those variables, has the advantage of improving the precision of the estimate of the effect of the treatment. This is the sense in which controlling may help you get closer to the truth. But it has disadvantages as well. Because of randomization, you can be sure that estimating the treatment effect without controlling (e.g., just comparing the average outcome in the treated and untreated groups) leads to an unbiased (if potentially very far from true) estimate of the true effect. By contrast, controlling for variables after the fact can produce a biased (if more precise) estimate. This means that if you were to run your experiment lots of times and always control for whatever variables turn out to be imbalanced, the average of your estimates might not converge on the true effect. So there are trade-offs to think about between reducing noise and increasing bias.

There is no easy answer or quick fix to the problem posed by imbalance following randomization. Our view is that you should probably do some combination of options 2 and 3. Also, whenever feasible, you should try to replicate experiments multiple times. No matter what, be honest and transparent about the choices you make.

Of course, what we'd really like is to avoid these difficult decisions by avoiding imbalance in the first place. And there are ways to do this. If you can identify and measure important characteristics ahead of time, you can design your experiment to insure balance. We've already briefly mentioned how. Prior to treatment, divide your sample into groups based on those characteristics and then randomize within those groups. This approach is called a *blocked* or *stratified* experimental design. Recall, before, we suggested that you might be concerned that Body Vibes differentially affect men and women, so you want to make sure your treated and untreated groups are balanced by biological sex. You achieve this by first dividing your sample into a male group and a female group. Then you randomize treatment assignment within these groups. This guarantees that biological sex is balanced between the treated and untreated groups (reducing noise), while still assigning treatment randomly (preserving unbiasedness). We can save ourselves a lot of headaches by following a procedure like this for pre-treatment characteristics that would cause us concern if they turned out to be imbalanced after the fact.

Lack of statistical power

Sometimes, an otherwise excellent experiment yields inconclusive results because the standard error is so large that even a reasonably sized effect would not be statistically distinguishable from zero. In this case, we say that your experiment lacked the *statistical power* to detect the effect of interest. Ideally, an experimenter would think about this problem beforehand and take steps to improve the statistical power of the experiment—for instance, by increasing the sample size. Thinking about power is an essential part of good experimental

design.

That said, sometimes, because of costs or other constraints, it turns out that you've run an under powered experiment. If you've already run the experiment and obtained imprecise estimates, what can you do? Here, the debate mirrors that around chance imbalance. You can try to improve precision by controlling for some variables, but, as we've already discussed, that has downsides as well. Sometimes, you may just have to accept that you don't have a convincing answer to your question, even after running an experiment.

Attrition

Sometimes people drop out of an experiment after treatment assignment. Such attrition differs from non-compliance by never-takers. The non-compliance situation involves people who were supposed to take up the treatment but chose not to. At least we get to observe the outcome for these non-compliers. When people drop out of the experiment, we don't even get to observe their outcome.

Suppose, for example, that Body Vibes make some people feel so young and care-free that they forget to come back for their follow-up meeting where we were planning to measure their skin health. This is bad! If attrition happened at random (i.e., was uncorrelated with skin health)—an unlikely scenario— then we could still obtain an unbiased estimate of the effect of our treatment by comparing the remaining members of the treated and untreated groups. We just lose some statistical power because our sample got smaller. If attrition is nonrandom but unrelated to the treatment assignment, then we can at least estimate the average effect of the treatment for the kind of people that choose to remain in the experiment. This is a genuine effect, but we've kind of changed the question. And, of course, most of the time, if there is attrition, we're left worrying that the attrition was both nonrandom and influenced by the treatment. For instance, maybe people left the study because the Body Vibes worked so well that they stopped worrying about skin health

entirely. In that case, were we to compare the remaining members of the treated and untreated groups, we'd be getting a biased estimate of the effect. As with many problems, it's much better if you can anticipate and mitigate attrition at the design stage rather than trying to account for it after the fact.

If attrition is unavoidable, what should an analyst do? First, you can test whether the experimental treatment influenced the rate of attrition. If it did, then we know we no longer have an apples-to-apples comparison. And relatedly, you can see whether the treated versus untreated units that remained in your sample differ systematically on other covariates that might be related to the outcome.

If those tests come out poorly or if you have other reasons to think that attrition biases your estimates, what can you do? Do we just have to throw out the experiment? There is one last resort that doesn't require the analyst to make any assumptions about the nature of attrition. You can try to bound the extent of the bias arising from attrition.

To see how this works, imagine an experiment with a binary outcome (1 = healthy skin, 0 = poor unhealthy skin). Suppose that 50 percent of the subjects in both our treated and untreated groups appear to have healthy skin, suggesting no effect of Body Vibes, but 5 percent of subjects in each group never showed up to have their skin health measured. We don't know whether attrition was affected by the treatment. But we can ask how bad the bias could be if it was.

The best case scenario for the hypothesis that Body Vibes are good for skin health would be if all of the people in the treated group who didn't show up had good skin, and all of the people in the untreated group who didn't show up had bad skin. In that scenario, 52.5 percent of subjects in the treated group would have good skin health compared to 47.5 percent in the untreated group, implying a positive effect of Body Vibes on skin health of 5 percentage points. Alternatively, in the worst case scenario for this hypothesis, those numbers would be flipped, and there would be a negative effect of 5 percentage points. We

can't be sure that attrition doesn't bias our estimates, but we can say that the bias can't possibly be greater than 5 percentage points.

Interference

If the treatment status of one unit affects the outcome of another unit, this can bias the results of an experiment. To see what we mean, consider the following story we heard from our colleague, Chris Blattman, about a pilot study for an experiment he ran in Liberia.

Blattman was interested in understanding what kinds of interventions might help high risk young men in post-conflict settings. In particular, he was trying to evaluate the impact of two kinds of interventions: offering high risk young men small cash grants to start some income-generating business and offering high risk young men cognitive behavioral therapy designed to instill self discipline.

One thing you might do, if you were interested in whether either of these two approaches works, is to start an organization offering each of them. You could then compare those who received either (or both) of these interventions to those who didn't, to see if those who received them did better in some important way.

Such an approach, however, would fail to compare apples to apples. It could well be that the young men who self select into receiving grants or therapy are already different from the average young man in the sample. They might be more ambitious, more healthy, smarter, or what have you. Thus, it would be a mistake to attribute the entire difference in performance between those who received the grants or therapy and those didn't to the causal effect of the intervention.

To address these concerns, Blattman designed a randomized experiment in which he randomly assigned the different interventions to different groups of Liberian young men. Everyone was given a small fee for participating in the experiment at all. But, beyond that

fee, some got nothing (the untreated group), some got a cash grant of about \$200, some got therapy, and some got both a grant and therapy.

Blattman's plan was to compare levels of criminality and homelessness among the young men assigned to different groups. The idea was that if the young men receiving one of the treatments had better outcomes than the untreated group, this would constitute apples to apples evidence that the intervention had a positive impact. So far so good.

The problems started when the young men in the study found out that half of them would receive \$200 while the other half would not. They explained that they did not want to play this lottery. They preferred to each receive \$100, eliminating any risk of getting nothing. Of course, giving them each \$100 would ruin the experiment. After all, the purpose was to randomly give some more than others and see whether those who received more actually did better. So Blattman's team dispensed the cash grants as per their experimental protocol—randomly giving only half the participants in the study the extra \$200.

But these young men were one step ahead of the researchers. They seem to have reached an understanding that they would provide one another with a sort of insurance. As a result of this insurance agreement, the winners of the lottery each gave some of their money to the losers who had received nothing. This kind of interference biases the estimates that come out of the experiment, since, now the carefully constructed untreated group had in fact received some of the treatment and the carefully constructed treated group had given up some of the treatment.

Here we see how hard it can be to design a clean experiment. The world, and your experimental subjects, can conspire to undo your efforts.

Blattman's failed pilot is an example of what we call *interference*. When you design an experiment, you randomly assign a treatment of interest across different units of observation (e.g., individual subjects, households, petri dishes, etc.). When you do that, you're assuming

that those units of observation are independent from one another. However, if the treatment status of one unit actually affects the outcomes in another unit, that's interference, and that can bias the results of your experiment. In this experiment, the interference concern is that the treatment status of the group that got the cash grants affected the outcomes in the untreated group because the treated subjects actually shared some of the treatment with the untreated subjects.

How do careful analysts deal with interference? Sometimes it's interesting enough that the interference itself becomes the object of investigation. Do the taxes in one state influence economic development in a neighboring state? If a campaign mobilizes a group of supporters, will that subsequently mobilize a group of opponents? If a public health program vaccinates children in one school, will this help protect children in another school? Researchers can sometimes design studies with the goal of estimating these kinds of spillover effects. For example, Blattman could have randomly assigned some friend groups to have one person treated with cash and other friend groups to have nobody treated with cash. Then, he could have tested whether the individuals who weren't given money behaved differently when a friend was given money.

In general, careful analysts need to anticipate interference and design their studies in ways that mitigate these possibilities. This is exactly why researchers do things like running pilot studies. In Blattman's case, when he scaled up the experiment after the problematic pilot, he made sure that it remained secret which subjects had and had not been assigned cash grants, to reduce the risk of interference.

Natural Experiments

For many interesting and important questions, we'd like to learn about causal relationships, but an experiment would be infeasible, unethical, unrealistic, or prohibitively expensive.

But sometimes the world creates something like experimental randomization for us, even without our intervening to actually run an experiment. We've already seen one example of this kind of *natural experiment*, in our discussion of the effect of charter schools on academic outcomes in Chapter 9. Although no quantitative analyst has been able to conduct their own experiment where they randomly send some kids to charter schools and others to public schools, many charter schools themselves randomize admissions. The schools didn't randomize for scientific reasons, but rather because they were required to by law. The law presumably exists because of concerns about fairness and equal opportunity. But regardless of the motivation, these lotteries create randomization "in the wild" that allows us to more credibly estimate the effects of attending charter schools vs. regular public schools than we could by simply comparing the performance of students at the two types of schools and trying to control for all the many potential confounders.

Natural experiments almost always involve some level of non-compliance—i.e., not everyone who wins the charter school lottery ultimately attends a charter school. Thus, in such settings, we are typically either estimating an intent-to-treat effect (i.e., the reduced-form relationship between winning the lottery and academic outcomes) or we need to take an instrumental variables approach to recover the complier average treatment effect. In this example, the instrument would be winning the lottery, the treatment is attending the charter school, and the outcome is some measure of academic performance (e.g., test scores).

When taking the instrumental variables approach, we need to think seriously about the conditions we described earlier. If there is natural randomization, we can have confidence in exogeneity. That is, we can credibly estimate the effect of winning the lottery on academic performance and on attending the charter school. But we have to think very carefully about the exclusion restriction. That is, are there ways that winning the lottery might affect academic performance other than through its affect on attending a charter school?

It might well be that in the charter schools example, the exclusion restriction is reason-

able and that we really can estimate the complier average treatment effect. But let us give you another example where the exclusion restriction is a bit more fraught.

The effect of military service on future earnings is of considerable interest to economists. But, of course, people who serve in the military and do not serve in the military differ in lots of ways that matter for earnings. Hence, a comparison of the earnings of veterans and non-veterans (even controlling for a bunch of stuff) is hopelessly confounded. Such a comparison does not provide a plausibly unbiased of the causal effect.

Fortunately (for social scientists), there is a natural experiment to help. During the Vietnam War, draft-eligible men were randomly assigned draft numbers. People were only actually drafted if their randomly assigned number was sufficiently low. Hence, we have a source of random variation in military service.

Of course, there was not perfect compliance with the draft lottery. For instance, many young men volunteered to serve in the military, despite having a high draft number. (In our earlier terminology, such men are always-takers.) So, if we want to get an estimate of the causal effect of military service on earnings (rather than the reduced-form effect of lottery number on earnings, which seems less interesting), we need to take an instrumental variables approach, which has been done in many studies. The idea is to use whether or not someone was drafted as the instrument, military service as the treatment, and future earnings as the outcome.

In this context, exogeneity is quite plausible. As best we can tell, the government really did assign draft numbers randomly. (Technically, they randomly assigned birthdays, so everyone with the same birthday was in the same boat, but whether one's birthday was selected was random.) So we really can estimate the effect of draft number on military service and on future earnings.

But what about the exclusion restriction? For the exclusion restriction to hold, it needs

to be the case that lottery number has no effect on future earnings other than through its effect on military service. How might this be violated?

One possibility concerns how people responded to receiving a low draft number. Such people may have been more likely to engage in various activities that would allow them to avoid the draft. For instance, they may have been more likely to flee the country. Or they may have been more likely to pursue higher education in order to receive a student deferment, which excused them from the draft while they remained in school. Becoming an expatriate or going to college might both directly affect future earnings. As such, these are alternative paths by which lottery number might affect earnings, other than through military service. Because of such violations of the exclusion restriction, it might well be that, even with random assignment of draft numbers, the instrumental variables approach will not allow us to use the draft lottery to recover the effect of military service on future earnings.

Of course, even exogeneity is tricky. For many causal questions we care about, the world does not provide natural randomization. In the next two chapters, we will talk about approaches that might allow us to estimate causal effects using natural sources of variation even without a true natural experiment.

Key Words

- **Research design:** Approaches to obtaining unbiased estimates of a treatment effect or other estimand.
- **Random assignment:** Deciding which units are assigned to receive treatment in a random fashion (e.g., by flipping a coin or using a random number generator).
- **Non-compliance:** When an experimental subject chooses a treatment status other

than the one to which it was assigned.

- **Compliers:** Units that take-up the treatment status they are assigned.
- **Always-takers:** Units that are always treated, regardless of whether they are assigned to be treated or untreated.
- **Never-takers:** Units that are never treated, regardless of whether they are assigned to be treated or untreated.
- **Defiers:** Units that take-up the opposite of the treatment status they are assigned.
- **Intent-to-treat (ITT) or reduced-form effect:** The effect on the outcome of being assigned to the treated versus untreated groups. This need not be the average treatment effect because of non-compliance.
- **First-stage effect:** The effect of being assigned to the treated group on take-up of the treatment. This corresponds to the fraction of compliers.
- **Complier Average Treatment Effect (CATE):** The average treatment effect for the compliers.
- **Heterogeneous treatment effects:** When the treatment effect differs across units.
- **Homogenous treatment effects:** When the treatment effect is the same across units.
- **Exogeneity:** An instrument is exogenous if it is randomly assigned or “as if” randomly assigned such that we can get an unbiased estimate of both the first-stage and reduced-form effects.
- **Exclusion Restriction:** An instrument satisfies the exclusion restriction if it affects the outcome only through its affect on treatment, not through any other channel.

- **Chance imbalance:** The situation where, despite random assignment, the treated and untreated groups differ in important ways because of noise.
- **Attrition:** The situation where experimental subjects drop out of the experiment, such that you do not observe outcomes for those subjects. Attrition is different from non-compliance.
- **Interference:** The situation where the treatment status of one unit affects the outcome of another unit.

Exercise

Download "GOTV_Experiment.csv" and "README.txt".

We will be analyzing data from a randomized experiment to estimate the effects of get-out-the-vote interventions on voter turnout.

Several factors complicate the analysis of this particular experiment. First, the probability of being randomly assigned to treatment was different for urban vs. non-urban areas. Second, some people assigned to treatment did not receive the treatment. And third, we are unable to observe turnout for some of the subjects. See the README file for more details.

1. Calculate the mean value of turnout for people who did and did not receive the treatment, and interpret the implied effect of get-out-the-vote interventions on turnout. Think about the likely biases that arise from the three complications listed above. If you had to guess, would you say that you are likely over- or under- estimating the average effect with this analysis.
2. Using the lessons from Chapter 10, try to account for the fact that the probability

of treatment varied between urban and non-urban places. How did your estimate change? Why?

3. Using the lessons from this chapter, let's try to account for noncompliance. First, try to estimate the intent-to-treat effect. Then, use some strategy to estimate the complier average treatment effect.
4. Now think about the attrition problem. What are you implicitly assuming if you just drop the subjects for whom we don't observe their turnout? Let's see how our estimates change under different assumptions. Estimate the complier average treatment effect assuming that none of the subjects who attrited would have voted. What would your estimate be under the worst case scenario for the effectiveness of GOTV? What about the best case scenario?

Notes

The study showing that sickly children in Peru were weaned from breastfeeding later is

Grace S. Marquis, Jean-Pierre Habicht, Claudio Franco, and Robert E. Black. 1997. "Association of Breastfeeding and Stunting in Peruvian Toddlers: An Examples of Reverse Causality." *International Journal of Epidemiology* 26(2):349–356.

The randomized experiment on breastfeeding in Belarus is

Michael S. Kramer, Tong Guo, Robert W. Platt, Stanley Shapiro, Jean-Paul Collet, Beverley Chalmers, Ellen Hodnett, Zinaida Sevkovskaya, Irina Dzikovich, and Irina Vanilovich. 2002. "Breastfeeding and Infant Growth: Biology of Bias?" *Pediatrics* 110(2):343–347.

There are many papers on the Vietnam draft lottery. Two of them (one classic, one recent) are:

Joshua D. Angrist. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80(3): 313–36.

Joshua D. Angrist and Stacey H. Chen. 2011. "Schooling and the Vietnam-era GI Bill: Evidence from the Draft Lottery." *American Economic Journal: Applied Economics* 3(2): 96–118.

Chapter 12

Regression Discontinuity Design

What You'll Learn

- Even when experiments are infeasible, there are still some special situations that allow us to estimate causal effects in an unbiased way.
- One such circumstance is when treatment changes discontinuously at a known threshold. Here a regression-discontinuity design may be appropriate.
- Regression discontinuity designs estimate a *local* average treatment effect for units right around the threshold where treatment changes.

Another such circumstance is when treatment changes over time for some groups but not for others. Here a difference-in-differences design may be appropriate.

We've just seen some examples of how clever natural experiments can help us learn about causality, even when we can't run an actual experiment. The idea is to look for ways in which the world creates situations where we can make apples-to-apples comparisons without running an experiment. Sometimes, as with charter schools, the world does this through actual randomization. Other times, you have to be a little more clever. This is the idea behind creative research designs.

In this chapter, we'll discuss one special situation that can help us generate more credible causal estimates—when treatment changes discontinuously at a known threshold. In the next chapter we'll consider another such situation—when treatment changes over time for some groups but not for others.

In Chapter 10, we discussed trying to learn about causal relationships by controlling for confounders. We don't typically have much faith in such approaches because it is so hard to measure all of the confounders out there. And if you can't measure something, you can't control for it. However, there are rare situations where we have a lot of information about the assignment of the treatment that may make this plausible. One example is a randomized experiment. If we know that treatment was assigned randomly, we know there are no confounders. Another is settings in which treatment is assigned according to some sharp rule. For instance, suppose each person (or other unit of observation) is associated with a score of some sort. People whose score is on one side of a threshold get the treatment and people whose score is on the other side of the threshold don't. Very close to that threshold, people on either side are likely to be very similar to one another on average. So a comparison of those two groups (one of whom got treatment and the other didn't) may be very close to apples-to-apples.

Let's be a little more concrete. Suppose that we want to estimate the effect of receiving a merit scholarship to college on future earnings. In general, it is difficult to do this credibly because the kinds of students who receive merit scholarships are probably different in many

ways that matter for future earnings—smarter, harder working, more ambitious—than those who do not. And of course, it's very difficult to measure and control for all these differences.

But what if the scholarship was awarded according to a strict scoring rule. The committee generates a score from 0-1000 for every applicant based on GPA, test scores, community service, extracurricular activities. Everyone with a score of 950 or above gets the scholarship, and everyone below does not. Now, even though nothing is randomized, we might be able to learn about the effect of receiving the scholarship for those applicants who were right around the threshold of 950. How exactly does this work?

Assume that the scholarship committee and the applicants can't precisely manipulate the scores. That is, the students put in effort without knowing exactly where their scores will fall, and the committee honestly evaluates the students also without knowing exactly where the scores will fall. Then, in expectation, the people with scores of 950 are almost identical to those with scores of 949. Nothing is randomized, but there are likely many idiosyncratic factors that could have easily pushed a 949 up to a 950, or vice versa. Had the 949s taken their standardized test on a slightly less stressful day, logged one more hour of community service out of hundreds, gotten one teacher who was a slightly more generous grader in one class, they would have been 950s and won the scholarship. Similarly, had the 950s had one minor, idiosyncratic thing not go their way, they would have been 949s and lost the scholarship. So it seems reasonable to say that, on average, the 949s were essentially the same as the 950s before the scholarship decision is made. And therefore we have something like a natural experiment. The comparison of individuals right around the threshold—some of whom got the scholarship (the 950s) and some of whom did not (the 949s), for essentially random reasons—is apples-to-apples. By comparing the future earnings of these two groups, we can estimate the causal effect of winning a merit scholarship.

Here's a more general way to think about this kind of situation. We want to estimate the effect of a binary treatment on some outcome. Treatment assignment is perfectly determined

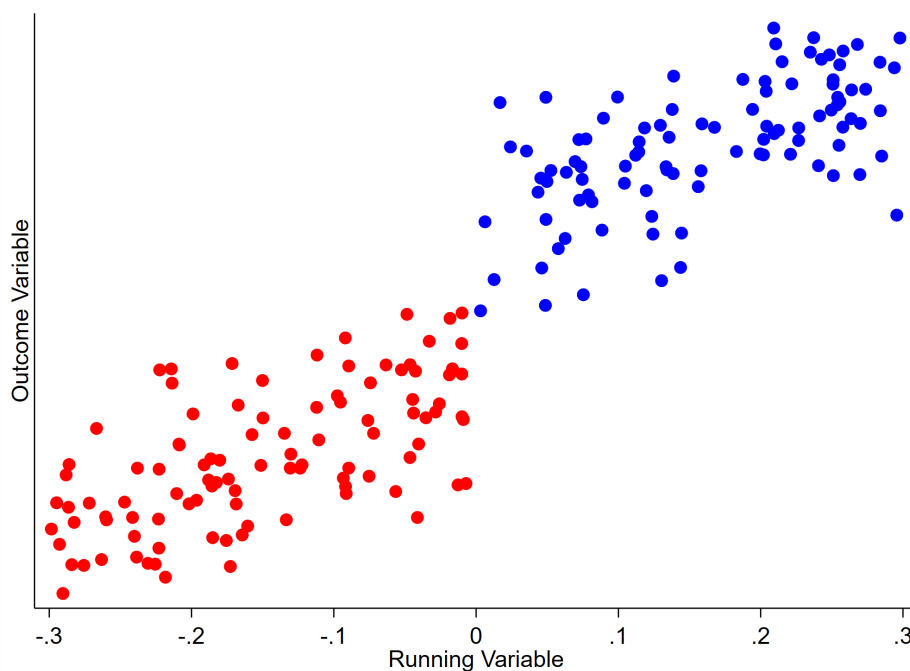


Figure 12.1: Scatter plot with treatment discontinuity when running variable equals zero.

by some third variable (like the score above) that we call the *running variable*. Specifically, if the running variable is above some threshold for a given unit, then that unit receives the treatment ($T = 1$), and if the running variable is below that threshold, that unit does not receive the treatment ($T = 0$). Such a situation might produce data that looks like Figure 12.1, with blue dots corresponding to treated units and red dots corresponding to untreated units.

How can we estimate the effect of the treatment in this kind of situation?

At first glance, it looks like there's not much we can do. The running variable is strongly correlated with the outcome of interest. In the scholarship example, this makes sense because the committee wants to select high-ability people, and, not surprisingly, the criteria they use to create the scores turn out to be highly correlated with future earnings, regardless of whether a student wins the scholarship. The committee uses a cutoff rule, so

everyone who receives the scholarship has higher values of the running value than anyone who does not. Clearly, then, if we compare those who do and do not receive treatment, we know that the inputs to the score are confounders. And, because of the cutoff rule, we can't make an apples-to-apples comparison by finding students with the same value of the running variable, some of whom did and some of whom did not receive treatment (i.e., the scholarship). Everyone with the same score has the same treatment status.

But don't give up yet. Let's think more about what we can do here. We can estimate the expected value of the outcome for a given value of the running variable. For units whose score on the running variable is above the threshold, this will tell us the expected outcome under treatment at that value of the running variable. We can estimate this quantity for every value of the running variable all the way down the threshold. Similarly, for units whose score on the running variable is below the threshold, this will tell us the expected outcome under control at that value of the running variable. We can estimate this quantity for every value of the running variable all the way up to the threshold. Therefore, right at the threshold, we have estimates of the expected outcome with and without the treatment. The difference between those two values might well be a good estimate of the effect of the treatment, at least for those units with a value of the running variable right at the threshold.

We could estimate this quantity by comparing units on either side of the threshold, all of which have values of the running variable very close to the threshold. This was the idea of behind comparing the 949s to the 950s to learn about the effect of merit scholarships. But there are actually somewhat better approaches.

One strategy is to run a regression of the outcome on the running variable and use it to predict the outcome with and without treatment right at the threshold. From these predictions we can estimate the "jump" or "discontinuity" in the outcome as the running variable crosses the threshold. That discontinuity is the causal effect of the treatment for units right at the threshold. For this reason, we call this strategy a *regression discontinuity*

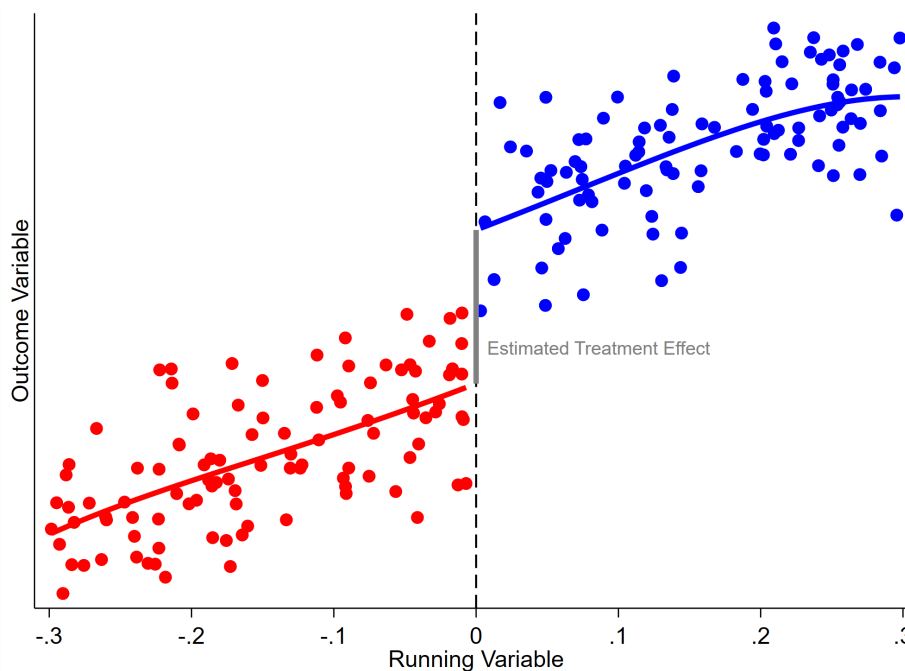


Figure 12.2: The regression discontinuity design estimates the jump in outcomes at the threshold, which is the causal effect of the treatment for units at the threshold.

(RD) design. Figure 12.2 illustrates the idea.

One thing worth emphasizing is what is referred to as the *localness* of the average treatment effect that a regression discontinuity design estimates. It is possible that the average effect of the treatment is different at different values of the running variable. This localness is illustrated in Figure 12.3.

To be more concrete, in our example, the effect of winning a scholarship on future earnings could be different for low- and high-achieving students. The regression discontinuity estimates the average treatment effect for units with values of the running variable right at the threshold. So, in our example, it estimates the effect of winning a scholarship on the future earnings of students with scores of 950, which might be different than the effect on students with scores of, say, 700. We sometimes refer to this estimand as a *local average*

ne Variable



treatment effect (LATE). Importantly, the LATE can differ from the overall average treatment effect in the population. So it is important, when using a regression discontinuity design, to think about whether the quantity estimated is really the one you are interested in.

Regression discontinuity designs are important in a variety of settings. One common application is in estimating the effects of government programs. Many policies change discontinuously at known thresholds. For example, individual-level government benefits are often means-tested, with eligibility determined by whether some continuous measure of income or poverty is on one or the other side of a threshold. County-level policies are often determined by population thresholds or by the share of residents of a certain type. Regression discontinuity designs provide a straightforward way to estimate the effects of these programs. Furthermore, these designs estimate the effects of the programs for the kinds of people or places where we care the most—the marginal unit that was just barely eligible or ineligible. So if policy makers are trying to figure out whether they should shrink or expand a particular government program, these regression discontinuity estimates should be highly informative.

Another setting amenable to regression discontinuity designs is in elections. Let's look at an example of that now.

How to Implement an RD Design

There are different ways for analysts to go about implementing their own regression discontinuity designs and there are pros and cons associated with each one.

The simplest approach, as mentioned above, is to simply compare the mean outcome for small bins on either side of the threshold. For example, we might compare the average earnings for applicants who scored between 950 and 954 to the average earnings for appli-

cants who scored between 945 and 949. For reasons you'll see in a moment, we often call this the *naive* approach.

A clear advantage of the naive approach is its simplicity. What makes it naive is the fact that it is virtually guaranteed to produce biased estimates. Why is this? The running variable is typically correlated with potential outcomes. Why would the committee use the scores to allocate scholarships if they didn't believe the scores corresponded to ability, effort, motivation, or some other factor that is likely correlated with earnings in the future?

Because the running variable is correlated with the potential outcomes of interest, there will always be some average difference between the groups just above and just below the threshold. Of course, as the size of the bins being compared shrinks (we often call this the *bandwidth*), the bias should shrink, but it will never disappear.

We can already see that one of the important decisions an RD analyst must make is to select a bandwidth. And when they make that decision, they often face a trade-off between reducing bias and improving precision. Smaller bandwidths will generally yield less biased estimates but also less precise estimates.

To try to reduce the bias that arises from the naive approach, another option is the *local linear* approach. Here, we again select a bandwidth, and for observations within that bandwidth, we run linear regressions of the outcome on the running variable separately on either side of the threshold. We estimate the predicted values of the regressions right at the threshold, and the differences in those predicted values give the predicted magnitude of the discontinuity.

With this approach, we're allowing for the possibility that there is a relationship between the running variable and the outcome, we're allowing that relationship to be different on either side of the threshold, and we're assuming that this relationship is approximately linear (at least for the small window of data that we're analyzing).

To make our lives easier and to obtain an estimate of the standard error, there is a way to implement this local linear approach with a single regression. First, rescale the running variable so the threshold is zero (i.e., subtract the value of the threshold from the running variable). Second, generate a treatment variable indicating whether an observation is above or below the threshold. Third, generate an interactive variable by multiplying the treatment variable and the running variable. And lastly, regress the outcome on the treatment, the running variable, and the interaction of the two for the observations within your bandwidth. The estimated coefficient associated with the treatment provides the estimated discontinuity.

A third common way that people implement RD designs is with polynomial regressions. An analyst might regress the outcome on the treatment, the running variable, and higher order polynomials (i.e., the running variable to the second power, third power, and so on). This approach accounts for a possible nonlinear relationship between the running variable and the outcome. A downside is that data points that are far from the threshold can have a big effect on the estimated discontinuity.

When implementing an RD design, the researcher clearly has a lot of degrees of freedom, so they have to be careful to avoid the problem of over-comparing and under-reporting. Your particular decisions should depend on your substantive knowledge and beliefs about the relationship between the running variable and the outcome and also how much bias you're willing to accept in exchange for a gain in precision, or vice versa. The best approach is to justify your choices with a combination of theory, substantive knowledge, and data analysis, and perhaps most importantly, show results for different specifications. If your estimates are robust across different bandwidths and specifications, this will lend additional credibility to your results.

Are Extremists or Moderates More Electable?

In both the 2016 and 2020 presidential elections, the Democratic party engaged in a heated debate about the electability of extremist versus moderate candidates. In particular, the liberal wing of the party was disappointed by the nominations of Hillary Clinton and Joe, both of whom they perceived as too moderate. The way to win elections, they argued, isn't to appeal to centrist voters. Rather, parties should nominate ideologically pure candidates that can turn out the base. Bernie Sanders, the argument went, was in a better position to defeat President Trump in the general election than either of his more moderate rivals. Sure, there might have been some moderates turned off by some of Sanders' more extreme policy proposals. But Sanders would have more than made up for those losses by mobilizing progressives who had lukewarm feelings about Clinton and Biden.

How can we assess whether this argument is right? On the one hand, moderate candidates might persuade more people in the middle to support their party. On the other hand, extremists might mobilize the base. So if you want to maximize the chances that your party wins the general election, who should you support in the primary election? It is, of course, impossible to say with confidence what would have happened if Sanders won the 2016 or 2020 Democratic nomination. But maybe we can say more about what happens, on average, when a party nominates a more extreme versus more moderate candidate.

To try to get a handle on this, let's turn to congressional elections where we have a lot more data than we do for presidential elections. At first glance, it looks like the advocates of ideologically pure candidates might be onto something. After all, it sure looks like Congress has a lot of ideological purists in it. If moderation is a winning strategy, why are there so many extremists in office?

For starters, we have to make sure we aren't forgetting the lessons of Chapter 4: correlation requires variation. The fact that many congresspeople are ideologically extreme does

not imply a positive correlation (to say nothing of a causal relationship) between ideological extremism and electoral success. To figure out which is true, we need to compare the electoral fortunes of extremists and moderates. Sure, one possible explanation of the large number of extremists in congress is that extremism really is correlated with winning. But another is that there are just very few moderates running.

Moreover, it may be misleading to think about extremism and moderation on a national scale. Rather, for the purpose of thinking about electoral strategy, we want to know whether a candidate is extreme or moderate relative to the preferences of her particular electorate or constituency. Sanders is surely an extreme liberal relative to the median voter in the U.S. But when he's running to represent Vermont in the Senate, perhaps he's only somewhat left of center. Indeed, maybe many congresspeople appear pretty ideologically moderate relative to their constituencies, but ideologically extreme relative to the country as a whole. This could happen if the constituencies are themselves constructed to be ideologically extreme compared to the country—some far to the left and others far to the right. But in this case, you wouldn't want to interpret the presence of lots of ideological extremists as evidence that extremism itself is an effective electoral strategy. Because the winning congressional candidates would not have been perceived as ideological extremists by the voters that elected them.

Given these concerns, what we really want to know is not the correlation between ideology and electoral success, but the causal effect of ideological extremism on electoral fortunes. To find an unbiased estimate of this, we need to compare how parties do in elections when they nominate an extremist versus a moderate, all else equal. On average, is the party better off running an extremist or a moderate candidate?

Of course, a naive comparison of the correlation between electoral outcomes and ideological extremism of candidates isn't be apples to apples. Presumably, the times, places, and situations where a party nominates a moderate are different than those where a party

nominates an extremist for all sorts of reasons that are consequential for electoral outcomes. For instance, most likely, liberal Democrats win primaries in liberal places where the Democratic Party is strong, and moderate Democrats win in conservative places where the party is weak. So if we found that extremists do better in general elections, that wouldn't tell us that parties are better off when they elect extremists. The causal interpretation of that correlation would obviously be confounded. We could try to control for differences across time and place, but we would always be worried that there are still unobservable baseline differences between places nominating extremists and moderates.

We can do a better job using a regression discontinuity design.

Major party congressional candidates are selected in primary elections. And election outcomes are determined by a sharp threshold. Suppose we analyze a large sample of primary elections that pitted one extreme candidate against one moderate candidate. The treatment we are interested in is the nomination of an ideologically extreme candidate. We want to know the effect of that treatment on the party's vote share in the general election. To set up the RD, define the running variable as the vote share of the extreme candidate in the primary. If that vote share is below one-half, the party runs the moderate in the general election; if it exceeds one-half, the party runs the extremist. We can now estimate the effect of running an extremist by implementing an RD design, comparing a party's general election outcome when it just barely nominated an extremist in the primaries versus when it just barely nominated a moderate in the primaries.

Andrew Hall did exactly this in a 2015 study. He estimated a large, negative discontinuity in a party's general election results at the threshold. That is, on average, a party that nominates an ideological extremist instead of a moderate significantly decreases its performance in the general election. Despite the predictions of the Sanders supporters, the evidence suggests that nominating extremists, on average, is a bad electoral strategy.

Hall's design is illustrated in Figure 12.4 The two lines represent separate linear re-

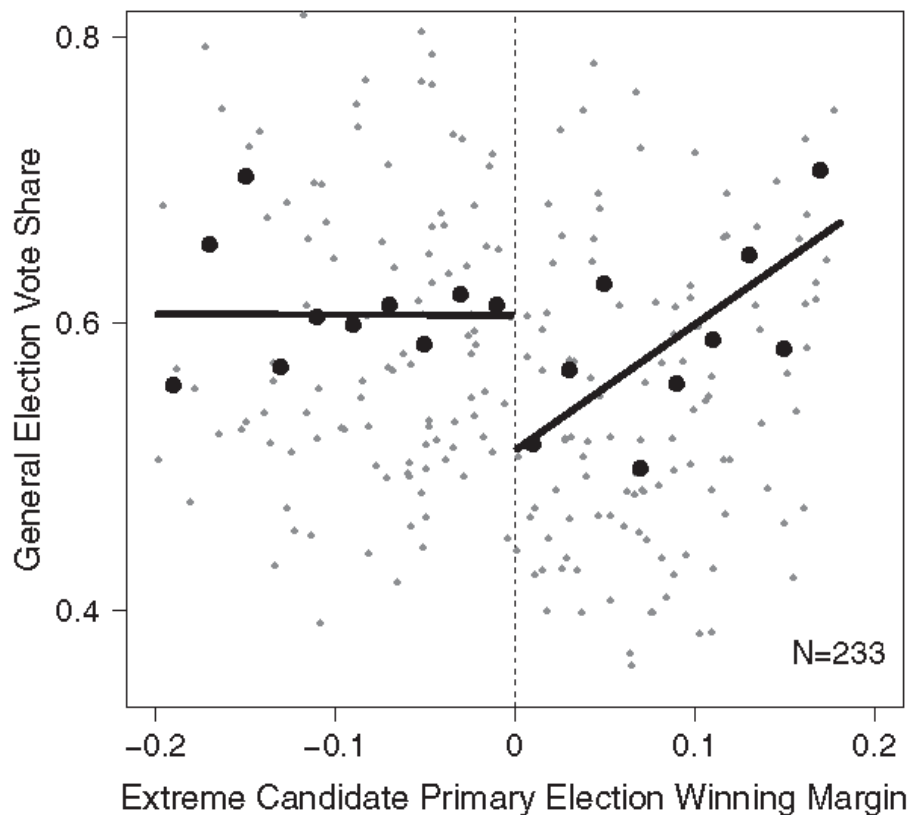


Figure 12.4: The effect of running an extremist on electoral prospects. (Source: Hall)

gressions on each side of the 50% threshold. The large, negative discontinuity right at the threshold is the estimated effect on general election vote share of nominating an extremist instead of a moderate, for a race where the primary election was evenly split between a moderate and an extremist.

What explains this result? In a follow-up study, Hall and Dan Thompson investigate further. Using a similar regression discontinuity design, they study the effect of nominating an extremist on voter turnout. Interestingly, contrary to the predictions of the Sanders supporters, there's no evidence that extremist candidates turn out the base. Or, rather, nominating an extremist does appear to turn out the base, but the wrong one. When a

party runs an extremist candidate, more people from the *other* party turn out to vote in opposition. Therefore, if we had to guess, these results suggest that if Bernie Sanders had won the Democratic primary, he might have done even worse than Clinton. He likely would have lost some of the centrist voters that preferred Clinton over Trump, *and* likely would have motivated Republican voters to turn out in greater numbers.

Continuity at the Threshold

In order for the regression discontinuity approach to provide an unbiased estimate of the causal relationship we want, it has to be the case that treatment status changes sharply at threshold *and* that nothing else that matters for outcomes does so. If baseline characteristics also change discontinuously at the threshold, then any differences in expected outcomes right around the threshold could be due to those changes in baseline characteristics, rather than treatment. That is, the comparison of treatment and control units would no longer be apples-to-apples, even right at the threshold, because those two groups would be differentiated by things other than just treatment status. But if average baseline characteristics of the units change continuously (rather than in a discrete jump) as the running variable passes through the threshold, then we can obtain an unbiased estimate of the effect of the treatment for units with a value of the running variable that is right at the threshold because the only thing that will differentiate units just on one or the other side of the threshold, on average, will be their treatment status. We call the requirement that baseline characteristics don't jump at the threshold, *continuity at the threshold* (or just *continuity* for short).

Let's see why continuity is crucial. Figure 12.5 illustrates what it looks like if the continuity condition is satisfied. The filled in dots are units we actually observe. The solid lines plotted through them are the average potential outcome functions (for the relevant value of treatment assignment). The hollow dots are units we don't observe (since we don't

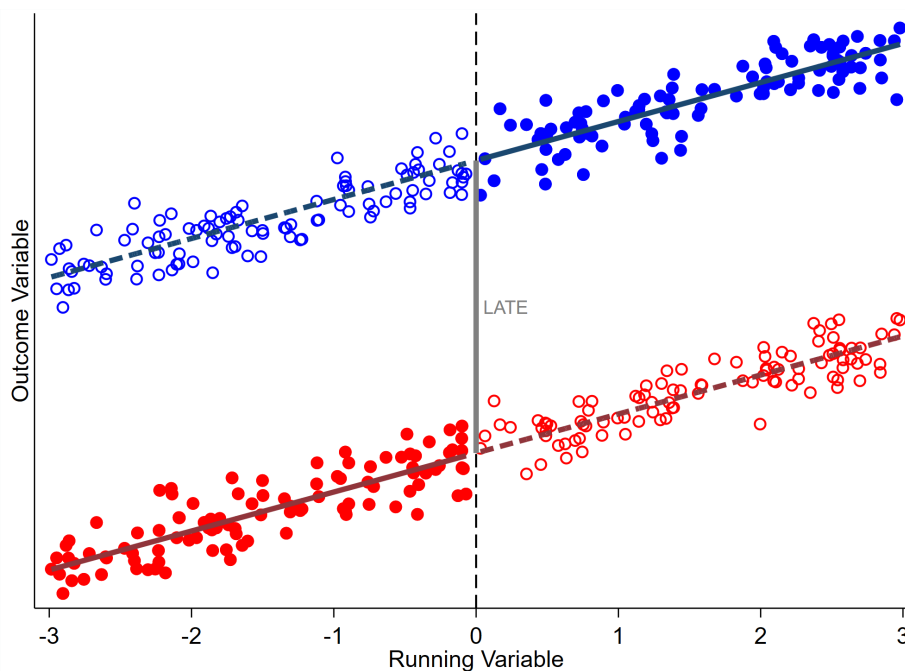


Figure 12.5: A case where average potential outcomes satisfy continuity.)

ever observe, say, the potential outcome under treatment for a unit with a value of the running value below the cutoff). The dashed lines plotted through them are the average potential outcome functions (again, for the relevant value of treatment assignment). Continuity is satisfied because these average potential outcome functions have no jump. That is, the average potential outcome under each of treatment and no treatment is continuous at the threshold. All that changes at the threshold is that units go from being untreated to treated.

Importantly, if continuity holds, then the gap between the red and blue dots at the threshold is in fact the local average treatment effect at that threshold, which is just what we want.

But what if continuity does not hold, so that the potential outcomes look like Figure 12.6?

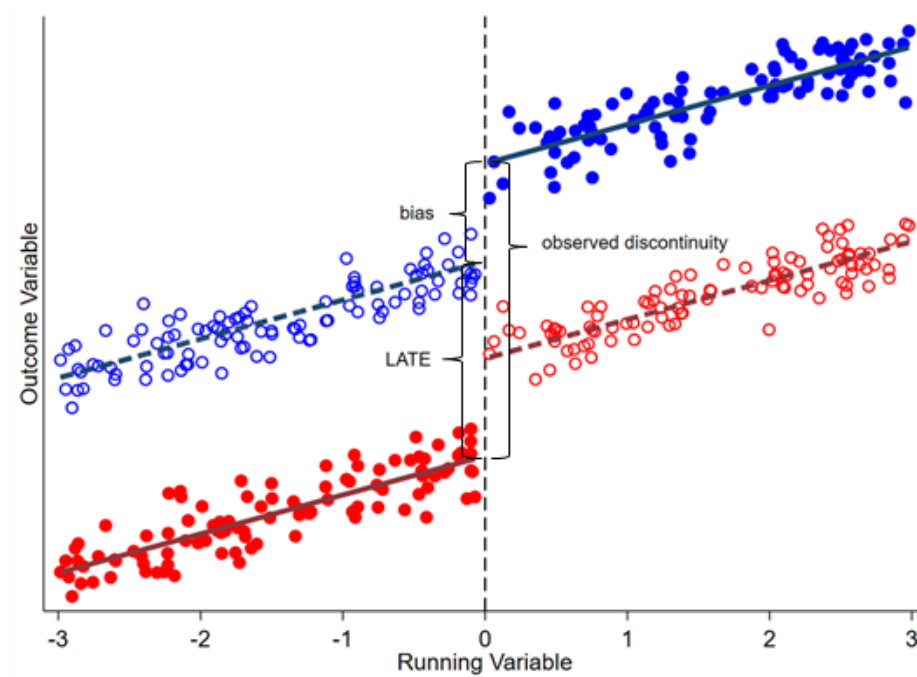


Figure 12.6: A case where average potential outcomes do not satisfy continuity at the threshold.

The true average treatment effect at the threshold is the difference between the filled in red dots and the hollow blue dots at the threshold. (You could also define it as the difference between the filled in blue dots and the hollow red dots.) But, right at the threshold, the potential outcomes jump up, even absent a change in treatment. We don't know why, but something besides treatment is changing right at the threshold. As a consequence, not all of the observed gap—i.e., the jump between between the filled in red and the filled in blue dots—is the result of the change in treatment. Some of it is the result of whatever else is changing. As such, that gap is a biased estimate of the LATE—in this case it is a big over-estimate of the true effect of the treatment change—since the gap includes both the effect of the treatment change and also the effect of whatever else is changing. Thus, without continuity at the threshold, the RD will give a biased estimate of the local average treatment effect.

When it comes to implementing an RD design, there are many different paths the analyst can take. However, viewed in the correct light, once a researcher has established that plausibility of continuity of potential outcomes at the threshold, her job ahead is clear. Using the sort of techniques we have already discussed (e.g., regression), she simply has to generate unbiased estimates of two things—the average outcome with and without treatment at the threshold.

To think about when an RD design is appropriate, we want to think about when continuity at the threshold is or is not plausible. It is worth noting that the continuity requirement is less demanding than you might have expected for credibly estimating causal relationships. For instance, it does not require that the treatment is assigned randomly (even by nature). In our scholarship case, we were able to use an RD even though, for every single student, treatment assignment was deterministic (i.e., there was no randomness at all). It also does not require that the outcome be unrelated to the running variable. Again, in our scholarship example, the running variable reflects genuine academic merit and, thus, is

positively correlated with future earnings outcomes. Finally, it does not require that units have no control over their value of the running variable or that units have no knowledge of the threshold. In our scholarship example, students could do all sorts of things to affect the running variable (e.g., study harder, do more community service).

So what could go wrong such that continuity does not hold?

Suppose that units have extremely precise control over their value of the running variable such that certain types may cluster just above or just below the threshold. This could potentially be a problem. In our scholarship example, we might worry that more privileged or more ambitious students have better information about the scoring system and can do just enough to exceed the threshold. Or we might worry that the committee has reasons to want to grant scholarships to students with certain characteristics (e.g., children of donors, athletes, particular ascriptive groups) and manipulates the scores or the threshold a little bit to get the desired result. In both of these cases, individuals just above the threshold would not be comparable to those just below. Instead they would have been *sorted* (by themselves or others) around the threshold based on other baseline characteristics that matter for outcomes. If this is the case, regression discontinuity does not provide an unbiased estimate of the causal effect.

Things can go wrong even without sorting around the threshold, simply because things other than treatment status change at the threshold. Here's a pretty interesting real world example. In France (and many other countries), the mayor's salary depends on a city's population. For instance, by law, mayoral salaries jump when a city has a population of more than 3,500 residents.

This seems like an opportunity to use an RD to learn about the effect of mayoral salary on all sorts of outcomes. For instance, we might want to know whether cities are better governed or elections more competitive, when mayors are paid more. For either of these outcomes, the treatment of interest is mayoral pay. The running variable is population.

And we happen to know that, by law, there is a discontinuous jump in the treatment as the running variable crosses the 3,500 resident threshold. Surely cities with 3,400 residents and cities with 3,600 residents are similar on average.

It looks good, no? But there's a problem with continuity. It doesn't come from towns strategically determining their populations to change the mayor's salary. It comes from other policies. You see, mayoral salary is not the only feature of city governance that changes by law at the 3,500 resident threshold. Other things that change include the size of the city council, the number of deputy mayors, the electoral rules, the process for considering a budget, gender parity requirements for the city council, and so on. So any discontinuity in outcomes at the 3,500 resident does not provide an unbiased estimate of the effect of mayoral salary because other characteristics that might matter for those outcome also change discontinuously at the threshold.

Clearly, then, before interpreting RD as an unbiased estimate of a causal relationship, it is important to assess the plausibility of the continuity assumption. There are several ways to do this. The most important is to think substantively. The best way to spot possible violations of continuity is to know a lot of details of the situation, so that you can be alert to any reasons to worry about sorting, bunching, or other things changing at the threshold. In our scholarship example, if you had sat in on a committee meeting or had deep knowledge of the kinds of characteristics the committee was under pressure to make sure were well represented among scholarship recipients, you would be in a better position to assess the plausibility of the continuity assumption than if you had no specific substantive knowledge of the situation. There are also other kinds of analyses one can do to help validate the continuity assumption. For instance, an analyst can look directly at measurable pre-treatment characteristics and see whether they seem to have discrete jumps at the threshold. If many measurable characteristics appear continuous at the threshold, we might be more confident that other, unmeasured baseline characteristics are also continuous.

One can also look at the distribution of the running variable itself. If we find bunching or clumping—i.e., significantly more units whose value of the running variable is just above or just below the threshold than we’d expect based on the overall distribution—then we might be concerned about some manipulation that violates continuity.

Exactly how bad a violation of the continuity assumption is depends on the details of the problem. If there is just a little sorting, or a small discontinuity in baseline characteristics, the RD is biased, but perhaps only a little bit. And if the researcher has a lot of data and, so, can focus on units only extremely close to the threshold, sorting would have to be extremely precise for it to affect the results. For instance, if we are estimating our scholarship RD using data on students with scores in the 940-949 range and students in the 950-959 range, we might be more concerned about sorting than if we have enough data to only consider students with a score of 949 or 950.

Does continuity hold in election RD designs

As we discussed above, elections are a great setting for RD designs since they have a clear running variable and a sharp threshold for winning. Not surprisingly, the election RD has been used in many studies on the effects of election outcomes ranging from campaign donations to drug violence to the study of the effect of extremists versus moderates we discussed earlier. So it is important to think clearly about whether the election RD is in fact a good research design.

Let’s remember what needs to be true for the election RD to provide an unbiased estimate of a causal relationship. We need for everything else that matters for the outcome under study to be continuous at the threshold. This guarantees that places where the relevant candidate (e.g., an extremist) just barely won are on average comparable to places where the relevant candidate just barely lost. In any application of the regression discontinuity approach, including electoral applications, it is always important to ask if this

condition is plausible.

And, indeed, some studies have argued that continuity may be violated in some electoral settings. The concerns have to do with manipulation of election results in close elections. For instance, in the example we just discussed, perhaps the party leadership prefers moderates. If it has ways of intervening (say by putting pressure on officials responsible for recounts) to nudge close election outcomes, it might do so in favor of moderate candidates. For his study, Hall shows that this does not appear to be the case.

But in another setting, the post-WWII U.S. House of Representatives, some evidence suggests that there may be continuity problems. In the relevant studies, scholars are interested in using the RD to estimate the *incumbency advantage*—how much better does the incumbent party do than the out-party, all else equal? To estimate this incumbency advantage, they compare things like the probability the Democrat wins when the incumbent is a Democrat versus when the incumbent is a Republican in places where the previous election (the one that determined incumbency) was just barely won by the Democrat versus just barely won by the Republican. For this to be a valid research design, there must be continuity at the threshold—the probability of the Democrat versus the Republican winning in the next election wouldn't change discontinuously in vote share in the previous election if it weren't for the fact that incumbency status changes. But there is reason to worry this isn't true. In particular, the winners of close elections appear to systematically differ from losers in a variety of ways—they have more campaign money, more experience in office, and are more likely to already be incumbents. For instance, in House elections decided by less than 0.25 percent of the vote, the incumbent party is statistically more likely to win than the challenging party. If this is because parties are able to manipulate close election outcomes, then we might worry that, even very close to the 50% threshold, we aren't making an apples-to-apples comparison when we compare future electoral outcomes in places where one party just barely won versus just barely lost. So, what's going on?

Devin Caughey and Jas Sekhon, who uncovered this fact, argue that the evidence points to electoral manipulation—incumbents have very precise knowledge of expected vote share and act strategically on or before election day in ways that allow them to win very close elections more than half the time. To believe this, however, you must believe that incumbent candidates can distinguish between situations where they expect their vote percentages to fall between 49.75 and 50.0 vs. 50.0 and 50.25. Real life campaigns appear to have nowhere near this level of precision in their election forecasts. Therefore, strategic campaigning is unlikely to be the explanation. What else could explain the imbalance? Most likely, this is a case of noise producing a false positive, much like Paul the Octopus in Chapter 7. When Anthony and four coauthors replicated the same tests that Caughey and Sekhon did, but for 20 different electoral settings across several countries, the postwar U.S. House was the only one for which such an imbalance was present. Thus, we suspect the election RD is in fact a good research design for learning about causal relationships in politics.

Non-Compliance and the Fuzzy RD

Thus far, we’ve talked about using a regression discontinuity design when treatment is completely determined by the running variable and the threshold. When this is the case, we sometimes say we are using a *sharp regression discontinuity designs*.

But, just as in our discussion of experiments, there are sometimes problems of non-compliance in settings that are otherwise suitable for an RD. That is, treatment may be discontinuously affected by which side of the threshold the running variable is on, but not deterministically. In addition to the compliers, there are some never-takers (units with values of the running variable above the threshold but who are untreated) and there are some always-takers (units with values of the running variable below the threshold but who are nonetheless treated).

When there are such non-compliers, we need to combine the regression discontinuity approach with an instrumental variables approach of the sort we discussed in Section 11. We do so by using which side of the threshold the running variable is on as an instrument for treatment assignment. This approach is sometimes called a *fuzzy regression discontinuity design*.

Of course, when combining the RD with an IV strategy, we have to worry about all the IV conditions that we discussed back in Section 11. One of those requirements was exogeneity—that we be able to estimate the effect of the instrument on the treatment and on the outcome without bias. As we will see, the RD requirement of continuity at the threshold insures that exogeneity is satisfied. But there were other IV conditions that we need to keep thinking clearly about. Most especially, we need to worry about the exclusion restriction—the requirement that the instrument have no

To see how fuzzy RD works, let's work through an example.

A classic question in counterinsurgency is whether indiscriminate violence by a counterinsurgent is productive or counterproductive. Melissa Dell and Pablo Querubin shed some quantitative light on this question in the setting of the US bombing strategy during the Vietnam War.

In Vietnam, the United States engaged in a massive bombing campaign in an attempt to suppress the Viet Cong guerilla forces in the north. Dell and Querubin want to ask whether such bombing worked?

One comparison they might make to try to answer that question is whether insurgents were more or less active in places the parts of Vietnam that experienced more bombing. Of course, such a comparison is not apples-to-apples. One might, for instance, worry that the United States was more likely to bomb locations where the insurgents were already quite active, in which case there would be a reverse causality problem.

In order to get at the true causal effect of indiscriminate bombing, Melissa Dell and Pablo Querubin use a regression discontinuity design. The history underlying it is actually quite amazing.

During the Vietnam War, Secretary of Defense Robert McNamara was obsessed with quantification. McNamara had pioneered the use of quantitative operations research during his time as president of Ford Motor Company. And At the department of defense, he surrounded himself with a team of so-called “whiz kids”, and a large team of computer scientists, economists, and operations researchers, with the goal of providing precise, scientific, quantitative guidance to war planners and the military.

One of these efforts was the *Hamlet Evaluation System* (HES). This project collected answers to enormous battery of monthly and quarterly questions about security, politics, and economics. The data were collected by local US and South Vietnamese personnel, who obtained information by visiting hamlets. Question answers were entered by punch card into a mainframe computer and then a complex algorithm converted them into a continuous score, ranging from 1 to 5, that was supposed to characterize hamlet security. These raw scores, however, were never reported out by the mainframe. No human ever say them. Instead, the computer rounded the scores to the nearest whole number, so that all analysts or decision-makers ever saw was a grade of A–E. Better letter grades were understood to correspond to greater hamlet security. These grades helped determine which hamlets should be bombed—with bombing being more often targeted at less secure hamlets.

Dell and Querubin were able to reconstruct the algorithm and, using declassified data, recover the underlying continuous scores. This sets them up for a regression discontinuity design.

Think about hamlets with scores in the 1.45–1.55 range. Some of these hamlets ended up with a score just below 1.5 and received an E. Others of these hamlets ended up with a score just above a 1.5 and received a D. But the difference between, say, a 1.48 and a

1.52, on a score created by a complicated (and largely arbitrary) combination of answers to 169 questions is meaningless. So we should expect that the underlying level of Viet Cong activity in these two types of hamlets is the same—that is, we should expect the potential outcomes to be continuous at the threshold.

But treatment—which, here, means being bombed by the United States—changes discontinuously at the threshold. US war planners did not ever see the underlying continuous score. All they saw was the letter grade. And, so, they perceived hamlets that received a D as meaningfully more secure than Hamlets that received an E (and similarly for D vs. C, C vs. B, and B vs. A). As such, they were more likely to bomb the hamlets with lower letter grades.

Figure 12.7 shows that this is the case. The horizontal axis plots the running variable—the distance of the first decimal of a hamlet’s score from 0.5 during a given quarter (call it quarter t). A hamlet whose value of the running variable is negative (because its score’s first decimal was below .5) received the worse of the two possible grades (e.g., an E instead of a D). A hamlet whose value of the running variable is positive (because its score’s first decimal was above .50) received the better of the two possible grades.

The vertical axis plots the frequency with which a given hamlet was bombed in the subsequent quarter (call it quarter $t+1$). The figure pools all the hamlets together, treating going from E to D, D to C, C to B, and B to A the same, so as to maximize the number of observations. As the figure shows, there is a discontinuous jump down in the frequency of US bombings at the threshold—hamlets that just barely received better grades were bombed less frequently than hamlets that just barely received worse grades.

Given this discontinuous change in treatment, it makes sense to use a regression discontinuity design to estimate the effect of bombing on insurgency. Figure 12.8 illustrates the idea. The horizontal axis is the same running variable as above. But now the vertical axis is the outcome of interest—Viet Cong activity in the hamlet in the next quarter. As the

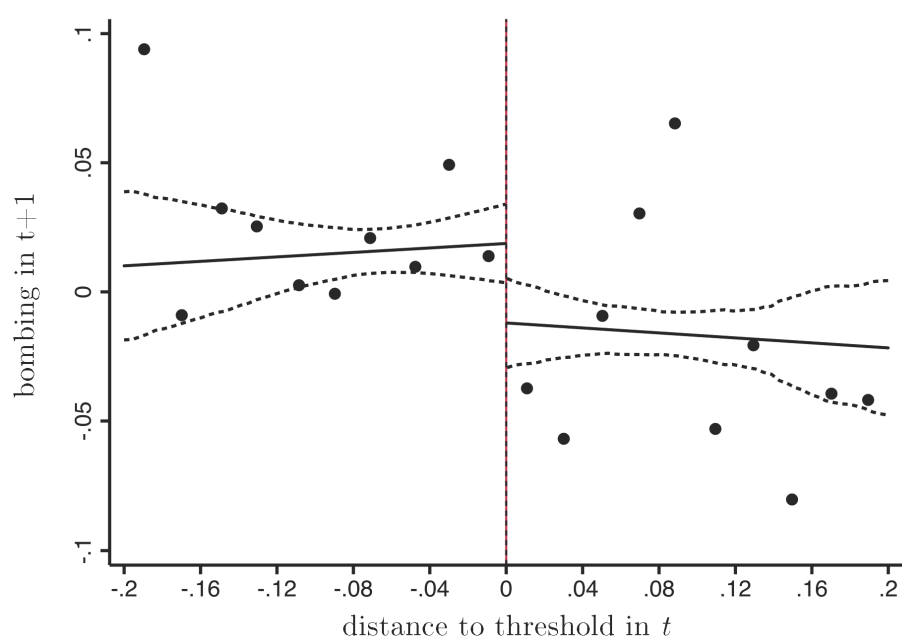


Figure 12.7: Hamlets that just barely received better grades in the Hamlet Evaluation System were bombed less frequently than hamlets that just barely received worse grades. .

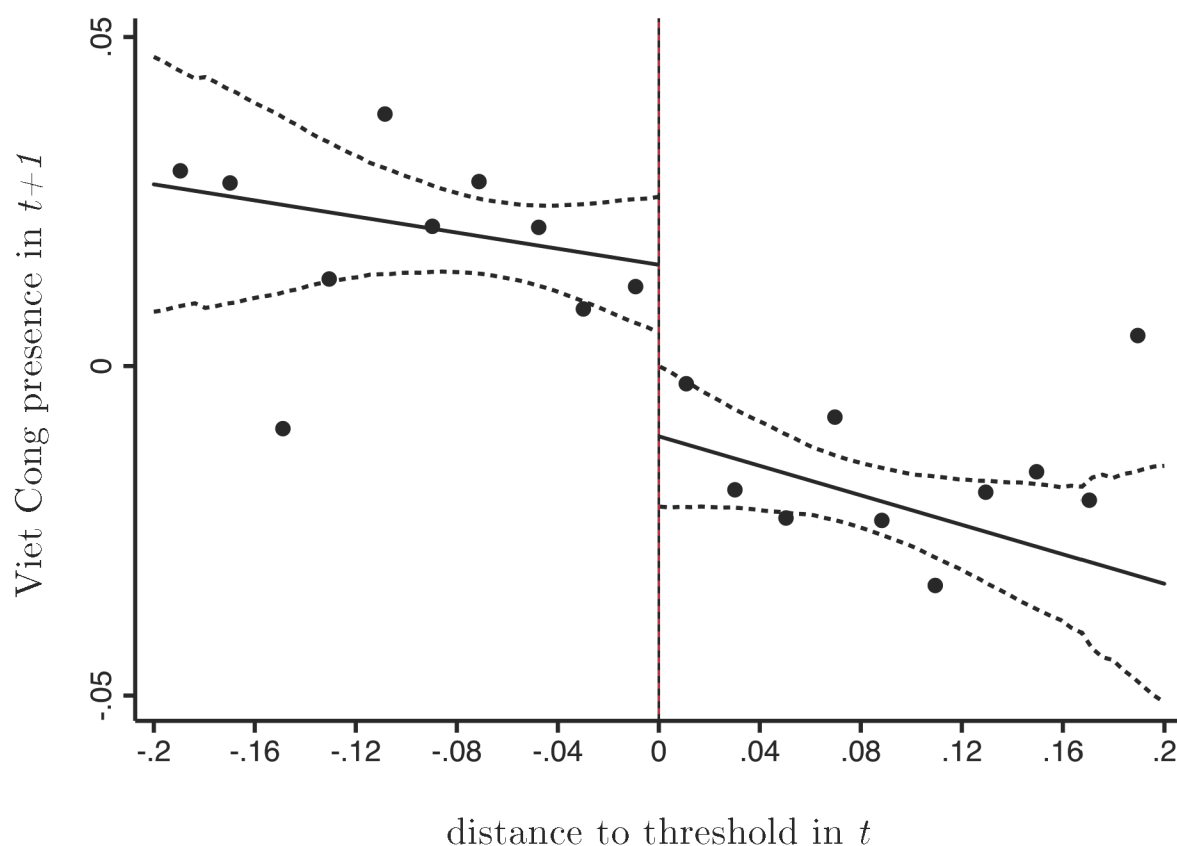


Figure 12.8: Hamlets that experienced more bombing saw more subsequent insurgent activity compared to otherwise similar hamlets that experienced less bombing.

figure shows, indiscriminate bombing appears to have been counter-productive. There is a discontinuous jump down in Viet Cong activity at the threshold. This means that hamlets that were bombed more (those to the left of the threshold) experienced more insurgent activity in the next quarter than otherwise similar hamlets that were bombed less.

But notice there is something a little different here from our normal regression discontinuity story. Going from a better score to a worse score doesn't guarantee increased bombing. The security score was only one input to bombing decisions. So it is not the case

that treatment goes from fully on to fully off at the threshold. That is to say there are some non-compliers—hamlets whose treatment status doesn’t depend on its exact value of the running variable.

But we know what to do about non-compliers. As we discussed in Chapter 11, we can use an instrumental variables approach. Recall, an instrument must satisfy several conditions:

1. *Exogeneity*: The instrument must be randomly assigned or “as if” randomly assigned, allowing us to obtain unbiased estimates of both the first stage and reduced form effects.
2. *Exclusion Restriction*: All of the reduced form effect must occur through the treatment. In other words, there is no other pathway for the instrument to influence the outcome except through its effect on the treatment.
3. *Compliers*: There must be some compliers.
4. *No Defiers*: There must be no units that take treatment only when the instrument says they should not.

How would we apply an instrumental variables approach here? The idea is to use which side of the threshold our running variable is on as the instrument. Let’s see that this satisfies the four conditions needed for an instrument.

The whole point of the regression discontinuity design is exogeneity. If potential outcomes are continuous at the threshold, then the RD allows us to obtain an unbiased estimate of both first stage (the effect of the instrument on bombing, as illustrated in Figure 12.7) and the reduced form (the effect of the instrument on Viet Cong activity, as illustrated in Figure 12.8).

The exclusion restriction requires that which side of the threshold the running variable is on has no effect on the Viet Cong activity other than through its effect on bombing. Here

there are questions to be asked. For instance, we need to worry about whether these grades were used for any other US military or policy decision making. If so, then the instrument will not satisfy the exclusion restriction.

Dell and Querubin provide two kinds of evidence in support of the plausibility of the exclusion restriction. First, they repeat their RD analysis for lots of other kinds of military operations by both the American and South Vietnamese military. They find no evidence of any other kind of military operations changing discontinuously at the threshold. As such, it is unlikely that the effects they find are the result of military actions other than bombing. Second, they review the administrative history of the Hamlet Evaluation System. That review reveals little evidence of the Hamlet Evaluate System scores being used for any other policy decision-making. The one exception is a program aimed at driving the Viet Cong out of the least secure hamlets. But that program had ended before the sample period covered by Dell and Querubin's data.

The requirements that there be compliers and no defiers are the most straightforward. It is clear from both the data and the history that there were compliers. And it seems unlikely that there are defiers—hamlets that would have been bombed if only they had gotten a *better* security score.

Given all of this, Dell and Querubin feel justified in employing a fuzzy RD design—using which side of the threshold a hamlet's security score was on as an instrument for bombing. In doing so, they are estimating an estimand that is a bit of a mouthful since reflects the localness of both the RD and the IV. In particular, they are estimating the local average treatment effect of bombing on insurgent activity for hamlet's with scores close to the threshold (the LATE from the RD) whose level of bombing is responsive to that score (the CATE from the IV). Doing so, they find that bombing was deeply counterproductive. For such hamlets, going from experiencing no bombing to experiencing the average level of bombing increased the probability of Viet Cong activity in the hamlet by 27 percentage

points.

Motivation and Success

Let's end with one last, fun example of a regression discontinuity design. Jonah Berger and Devin Pope implement an RD to estimate the effect of psychological motivation on performance. They analyze over 18,000 professional basketball games to test whether the motivation of being behind and needing to catch up leads to better performance than the complacency of being ahead and simply needing to hold onto a lead. Their running variable is the point margin of the home team at half time, and they test whether the probability of ultimately winning a game changes discontinuously as the half-time point margin crosses the threshold of 0, when the home team goes from being just behind to just ahead.

Figure 12.9 below shows the results. As we would expect, the point margin at half time is correlated with the probability of ultimately winning the game. When the home team is 10 points ahead at half time, they go on to win about 85% of the time, but when they're 10 points behind, they only win 25 percent of the time. This makes sense since some teams are better than others—good teams are both more likely to be ahead at the half and more likely to win the game. More interesting, however, is the comparison when the score is almost tied at the half. Presumably, there is very little quality difference, on average, between teams that are ahead or behind by just 1 point at half time. Yet, the home team is actually more likely to win when they're 1 point behind at halftime than when they're 1 point ahead. Berger and Pope's regression discontinuity shows that being just barely behind increases the probability that the home team wins by 6 percentage points! Maybe those inspirational half-time speeches really do work.

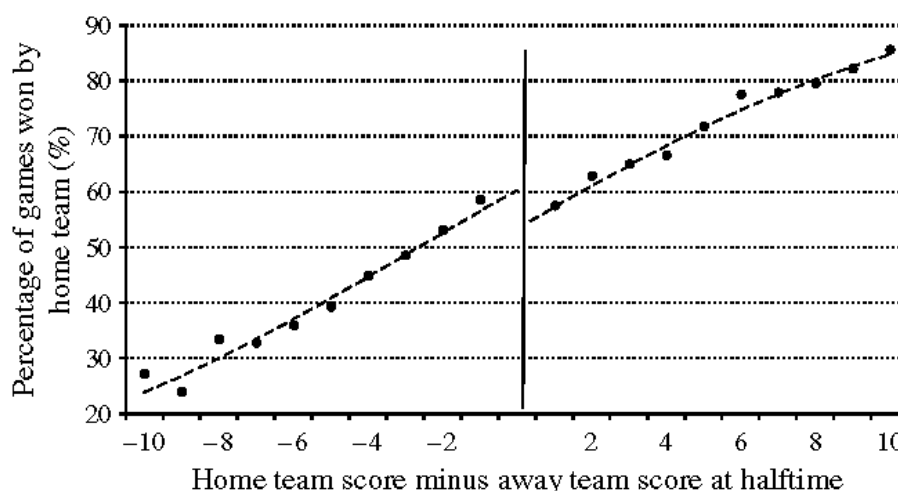


Figure 12.9: The effect of being ahead or behind at half-time on winning the game, for games that were close at half-time. Source: Berger and Pope)

Key Words

- **Running Variable:** A variable for which units' treatment status is determined by whether their value of that variable is on one or the other side of some threshold.
- **Regression Discontinuity Design:** A research design for estimating a causal effects by looking at the discontinuous jump in an outcome on either side of a threshold that determines treatment assignment.
- **Local Average Treatment Effect (LATE):** The average treatment effect for some specific subset of the population. In the case of regression discontinuity designs, the relevant LATE is the one for the subset of the population with a value of the running variable right at the threshold that determines treatment assignment.
- **Continuity at the Threshold:** The requirement that potential outcomes do not change discontinuously at the threshold that determines treatment assignment. If continuity at the threshold doesn't hold, then a regression discontinuity design does

not provide an unbiased estimate of the local average treatment effect.

- **Sharp Regression Discontinuity:** An RD where treatment assignment is fully determined by which side of the threshold the running variable is on.
- **Fuzzy Regression Discontinuity:** A research design that combines RD and IV. The fuzzy RD is used when treatment assignment is only partially determined by which side of the threshold the running variable is on. The researcher, therefore, uses which side of the threshold the running variable is on as an instrument for treatment assignment. In this setting, continuity at the threshold guarantees that the exogeneity assumption of IV is satisfied. But we still have to worry about the exclusion restriction.

Exercises

1. The state of Alaska asks you to estimate the effect of their new automatic voter registration policy. They want to estimate the effect of automatic registration on voter turnout. The policy was first implemented in 2017, but they report to you that, unfortunately, they initially didnt have the resources to roll the policy out to everyone in the state. As a result, they initially just applied automatic registration to people who had moved to Alaska in the last 2 years, but they havent yet applied it to people who moved to Alaska before then. Theyre worried that this might be a limitation for your study, and they apologize that they werent able to implement the policy for everyone, but theyre still hoping that you can help. How would you respond, and how might you go about estimating the effect of automatic voter registration in Alaska?
2. The U.S. federal government subsidizes college education for students through Pell Grants. An individual is eligible for Pell Grants if their family income is less than \$50,000 per year. Knowing this information, how could you potentially utilize this

information and implement an RD design to estimate the effect of college attendance on future earnings? Would this be a sharp or a fuzzy RD design? What data would you want to have at your disposal? What is the running variable? What's the treatment? What's the instrument (if any)? What's the outcome? What assumptions would you have to make in order to obtain unbiased estimates?

3. Download the data from [*link*](#) on crime and temperature in Chicago across different days in 2018. This is the same data we examined in Chapters 2 and 5. Imagine that the Chicago Police Department implemented a policy in 2018 whereby they stopped patrolling on days where the average temperature was going to be below 32 degrees (and suppose they have really good forecasts so they can very accurately predict, at the beginning of the day, the average temperature for that day). Their logic is that it's less pleasant for police officers to be out on the streets when it's cold, and there's less crime on cold days anyway. Use this (fake) information to estimate the effect of policing on crime.
 - (a) A helpful first step when implementing an RD design is to generate your own running variable where the threshold of interest is at 0. Rescale the temperature such that the threshold is at 0 by generating a new variable called "runningvariable" which is simply the temperature minus 32.
 - (b) We'll also need to generate our treatment variable. Generate a variable which takes a value of 1 if policing was in place on that day and 0 if it was not.
 - (c) It's often helpful to look at our data before conducting formal quantitative analyses. Make a scatter plots with crimes as your vertical axis and temperature as your horizontal axis. Focus only on days where the temperature was within 10 degrees of the policy threshold, and draw a line at the threshold. Visually, does it look like there is a discontinuity at the threshold?

-
- (d) There are several different ways to formally implement an RD design. The simplest is to focus on a narrow window around the threshold and simply compare the average outcome on either side. Focusing only on days where the temperature was within 1 degree of the threshold, compute the average number of crimes just above and just below and compute the difference. Notice that you can (if you'd like) do this in one step with a regression.
- (e) What concerns would you have with the simple approach above? Think about the trade-offs you face as you're deciding which bandwidth to select. How does your estimate change if you use a bandwidth of 10 degrees instead of 1 degree? Why?
- (f) Another strategy is to use the local linear approach. For days that were less than 5 degrees below the threshold, regress crime on the running variable and compute the predicted value at the threshold (hint: because you rescaled your running variable, this should be given by the constant term). Do the same thing for days that were less than 5 degrees above the threshold, and compare those two predicted values. (Note that this can also be done with a single regression as described in the text.)
- (g) What benefits does this local linear approach have over the simpler approach of comparing means?
- (h) You might also consider allowing for a non-linear relationship between the running variable and the outcome. Generate new variables corresponding to the running variable squared and the running variable to the third power. Regress crimes on policing, the running variable, the running variable squared, and the running variable to the third power. Only include observations within 10 degrees of the threshold. Interpret the estimated coefficient associated with policing.
- (i) What are the pros and cons of this polynomial approach relative to the previous

ones?

Notes

The study on the effects of electing an extremist vs. a moderate in a primary election is

Andrew B. Hall 2015. “What Happens When Extremists Win Primaries?” *American Political Science Review* 109(1):18–42.

The studies on the validity of electoral regression discontinuity designs are

Devin Caughey and Jasjeet S. Sekhon. 2011. “Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008.” *Political Analysis* 19(4):385–408.

Andrew C. Eggers, Anthony Fowler, Andrew B. Hall, Jens Hainmueller, and James M. Snyder, Jr. 2015. “On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races.” *American Journal of Political Science* 59(1):259–274.

The study on US bombing during the Vietnam War is

Melissa Dell and Pablo Querubin. 2018. “Nation Building through Foreign Intervention: Evidence from Discontinuities in Military Strategies.” *Quarterly Journal of Economics* 133(2), 701–764.

The study on the effect of losing at halftime in basketball is

Jonah Berger and Devin Pope. 2011. “Can Losing Lead to Winning?” *Management Science* 57(5):817–827.

Chapter 13

Difference-in-Differences

What You'll Learn

- Another situations that allow us to estimate causal effects in an unbiased way is when a treatment changes at different times for different units. Here a difference-in-differences design may be appropriate.
- Difference-in-differences designs effectively control for all confounders that don't vary over time.

Regression discontinuity isn't the only creative research design that lets us get a causality in the absence of an experiment. When some units change treatment status over time but others don't we may be able to learn about causal relationships using a strategy called *difference-in-differences*.

The basic idea is pretty simple. Suppose we want to know the effect of a policy. We can find states (or countries or cities or individuals or whatever the relevant unit of observation is) that switched their policies and measure trends in the outcome of interest before and after the policy change. Of course, we may worry that outcomes are systematically changing over time for other reasons. But we can account for that by comparing the change in outcomes for states that changed policy to the change in outcomes for states that did not change policy. If the trends in outcomes for states that did and did not change policy would have been the same if not for the policy change in some states, then we can use the states that did not change policy as a baseline of comparison, to account for the over-time trends. Our estimate of the causal effect of the policy change, then, will come from any change in outcomes in states that did change policy over-and-above that baseline trend that we estimated from the states that didn't change policy. This is called a difference-in-differences design because we first get the difference (or change) in outcomes over time for states that did and did not change policy. Then we compare the difference in those differences.

Like the regression discontinuity design, the power of the difference-in-differences approach is that it allows us to estimate causal effects even when we can't randomize the treatment or control for every possible confounder. But nothing is for free. Difference-in-differences designs come with their own requirements. For regression discontinuity, we needed continuity at the threshold. For difference-in-differences, we need the condition we just described: that the trend in outcomes would have been the same on average across units, but for the change in treatment that occurred in some units. This condition is often called *parallel trends*.

Parallel Trends

It's worth making sure we are thinking clearly about what the parallel trends requirement really means. As we've said, differences-in-differences estimates are unbiased so long as the trends in outcomes would have been parallel, on average, in the absence of any changes in treatment. In other words, the parallel trends requirement is really about potential outcomes. For a binary treatment, can think about each unit's outcome under treatment and under control in each of two time periods. To capture this idea, let's think about there being potential outcomes for each unit in each time period. We will refer to the two time periods as period *I* and period *II*. And let's think about our population being divided into two groups: a group that changes from untreated to treated between the two periods (*UT*) and a group that remains untreated in both periods (*UU*).

To think clearly about parallel trends, we are going to return to our potential outcomes notation. Label the average potential outcome in group *G* in period *p* under treatment status *T* as:

$$\bar{Y}_{T,G}^p.$$

We observe the outcomes for a sample of the members of each group in each period. Let's start with the group that never changes treatment status (*UU*). If we just look at the average change in outcome between the two periods, it gives us an estimate of the difference in outcomes in the untreated condition between the two periods:

$$DIFF_{UU} = \underbrace{\bar{Y}_{0,UU}^{II} - \bar{Y}_{0,UU}^I}_{\text{average untreated trend for } UU} + \text{Noise}_{UU}.$$

The noise, of course, simply comes from the fact that we are looking at a sample.

And analogously for the group that changes treatment status (*UT*) the average change

in outcome between the two periods is

$$DIFF_{UT} = \bar{Y}_{1,UT}^{II} - \bar{Y}_{0,UT}^I + \text{Noise}_{UT}.$$

The *difference-in-differences* is, quite literally, the difference between these two differences:

$$\text{Difference-in-Differences} = DIFF_{UT} - DIFF_{UU}.$$

To see where parallel trends comes in, we are going to cleverly rewrite $DIFF_{UT}$ by adding and subtracting $\bar{Y}_{0,UT}^{II}$ from it. You'll recall we did something similar back in Chapter 9 in order to understand baseline differences. Just like then, while we know this seems kind of weird, we ask that you trust us for a minute. And, remember, at the very least it should be clear that we aren't doing any harm since, by adding and subtracting the same term, we are really just adding zero. When we do that, we get:

$$\begin{aligned} DIFF_{UT} &= \bar{Y}_{1,UT}^{II} - \bar{Y}_{0,UT}^I \\ &= \underbrace{(\bar{Y}_{1,UT}^{II} - \bar{Y}_{0,UT}^{II})}_{\text{average treatment effect for } UT \text{ in } II} + \underbrace{(\bar{Y}_{0,UT}^{II} - \bar{Y}_{0,UT}^I)}_{\text{average untreated trend for } UT} + \text{Noise}_{UT} \end{aligned}$$

Once again, our algebra trick was actually pretty cool. We can now see that the over-time difference for group UT is made up of three things, which correspond to our favorite equation. First, there is the average (period II) treatment effect for group UT . Back in Chapter 9 we learned that this was called the ATT, the average treatment effect on the treated units. We can think of this as our estimand. Second, there is the trend in outcomes that would have happened for group UT even if they had remained untreated. We can think of this as a source of bias that comes from just looking at what happens in the UT group before and after treatment. And there is noise, as always.

With this in hand, we can now rewrite the difference-in-differences in terms of the ATT, the average untreated trends in potential outcomes for both groups, and noise. This will make clear what we are really doing—using the overtime trend in the UU group to try to eliminate the bias that comes from just looking at $DIFF_{UT}$:

$$\begin{aligned} \text{Difference-in-Differences} &= DIFF_{UT} - DIFF_{UU} \\ &= \underbrace{\bar{Y}_{UT,1}^{II} - \bar{Y}_{UT,0}^{II}}_{ATT} + \underbrace{\underbrace{\bar{Y}_{UT,0}^{II} - \bar{Y}_{UT,0}^I}_{\text{average untreated trend for } UT} - \underbrace{\bar{Y}_{UU,0}^{II} - \bar{Y}_{UU,0}^I}_{\text{average untreated trend for } UU}}_{\text{difference in average trends}} + \underbrace{\text{Noise}_{UT} - \text{Noise}_{UU}}_{\text{Noise}} \end{aligned}$$

Now we can see what parallel trends really means in terms of potential outcomes and our favorite equation. The difference-in-differences equals the ATT (estimand) plus the difference between the average trend for group UT and group UU (bias) plus noise. So when does the difference-in-difference gives us an unbiased estimate of the ATT? When the untreated trend is the same for both groups, so that the difference in average trends equals zero.

This is what parallel means. The change in average outcome would have been the same in the treated and untreated group, had everyone remained untreated. When this is the case, by subtracting $DIFF_{UU}$ from $DIFF_{UT}$ we eliminate the overtime trend, leaving an unbiased estimate of the average treatment effect (in period II) for units that switched treatment status.

Notice, this notation highlights another subtle point. Difference-in-differences does not quite estimate the ATE. It estimates the average treatment effect for those units who actually change treatment status—more analogous to the ATT. Whether or not this is a good estimate of the ATE depends on whether treatment effects differ systematically across the units that do and don't switch treatment status. But, in any event, this is a genuine

causal effect and, at least for some applications, may in fact be the quantity of interest.

2 Units and 2 Periods

So far, we've been a bit abstract. Let's talk about a concrete example from classic work by David Card and Alan Krueger on the effect of the minimum wage on employment. This example is nice because it shows how difference-in-differences works in its most simple form. There are only 2 units, 2 periods, and 1 change in treatment status for one of the units.

Card and Krueger wanted to know whether a higher minimum wage increased unemployment. Their idea was to exploit the fact that New Jersey raised its minimum wage in early 1992, while Pennsylvania, which borders New Jersey, did not. They collected data on the average number of full-time equivalent employees (FTE) per fast food restaurant (which tend to pay minimum wage) in both New Jersey and Pennsylvania in January of 1992 (before New Jersey raised its minimum wage) and in November of 1992 (after New Jersey raised its minimum wage). Their data look as follows.

	January 1992	November 1992
New Jersey	20.44	21.03
	FTE/restaurant	FTE/restaurant
Pennsylvania	23.33	21.17
	FTE/restaurant	FTE/restaurant

Table 13.1: Fast food employment from Card and Krueger.

A first comparison we might think to make to learn about the effect of the minimum wage on employment is the difference between the employment levels in New Jersey and in Pennsylvania in November of 1992. After all, by November, New Jersey had raised its minimum wage, while Pennsylvania had not. That comparison shows that Pennsylvania

fast-food restaurants employed only .14 more people, on average, than New Jersey restaurants, suggesting that a higher minimum wage may have almost no impact on employment.

But we already know lots of reasons why we cannot interpret that difference as the effect of minimum wage. New Jersey and Pennsylvania might differ in all sorts of ways besides the minimum wage that matter for employment. For instance, perhaps them two states have different levels of economic prosperity, different tax systems, or differently sized fast food restaurants. And since, in this comparison, the state and the treatment are perfectly correlated, any such difference between New Jersey and Pennsylvania can be thought of as a confounder.

Another comparison we might make is to look at the change in employment in New Jersey between January and November, since the New Jersey minimum wage changed between these two time periods. This comparison shows an increase in employment of 0.59 employees per restaurant, suggesting that perhaps raising the minimum wage slightly increased employment. This approach has the advantage of comparing one state to itself, so we no longer need to worry about any cross-state differences. But now we have a new concern. Maybe January and November differ in terms of fast-food employment for other reasons—e.g., because of seasonality or overall changes to the economy over the course of the year. Any such time trends would be a confounder in this comparison.

Table 13.2 sums up, in terms of our favorite equation, the two ideas we've considered thus far and explains why neither gets us an unbiased estimate of the effect of minimum wage.

But we can do better. Start by thinking about the comparison between New Jersey and Pennsylvania in November. The problem with that comparison is that it reflects both the effect of the high minimum wage (the estimand) and any systematic differences between New Jersey and Pennsylvania (the bias), plus, as always, noise. But suppose the differences between New Jersey and Pennsylvania aren't changing over time. Then the difference

	January 1992 <i>NJ and PA</i> <i>low minimum wage</i>	November 1992 <i>NJ high minimum wage</i> <i>PA low minimum wage</i>	Difference <i>November - January</i>
NJ	20.44 FTE/restaurant	21.03 FTE/restaurant	.59 percentage points <i>effect of high minimum wage</i> + <i>over time trend</i> + <i>noise</i>
PA	23.33 FTE/restaurant	21.17 FTE/restaurant	
Difference <i>NJ - PA</i>		-.14 percentage points <i>effect of high minimum wage</i> + <i>differences between states</i> + <i>noise</i>	

Table 13.2: Two comparisons that do not unbiasedly estimate the causal effect of minimum wage increase.

in employment in New Jersey and Pennsylvania in January, when they both have a low minimum wage, reflects those same across-state differences, but without the effect of the high minimum wage that New Jersey adopted later in the year. So we can use that employment difference in January to estimate the underlying differences between the two states. And then, subtracting the January difference from the November difference (i.e., finding the difference-in-differences) will leave us with an unbiased estimate of the effect of the high minimum wage. (Of course, there is different noise in each comparison, so the noise terms don't just cancel.)

The same procedure works if we start from our comparison of New Jersey in November to New Jersey in January. The problem with that comparison is that it reflects both the effect of the high minimum wage and any other differences between November and January that matter for employment (plus noise). But suppose those over time trends are the same in New Jersey and Pennsylvania. Then the difference in employment in Pennsylvania between year January and November is an estimate of the over time trend, without any

effect of the minimum wage (since Pennsylvania didn't change its minimum wage in 1992). So subtracting the change in employment in Pennsylvania from the change in employment in New Jersey will also leave us with an unbiased estimate of the effect of the high minimum wage.

As shown in Table 13.3, either way we do this calculation, we find the same answer. Surprisingly, the estimate that this procedure leaves us with is that a higher minimum wage appears to increase employment by 2.75 FTE per restaurant. The key is that the Pennsylvania data suggests that there was a big baseline drop in employment from January to November 1992. So the 0.59 FTE increase in NJ was a misleading underestimate of the true effect being masked by an overtime trend.

Importantly, by calculating the difference-in-differences, we were able to effectively account for systematic differences between the states and this overtime trend, without ever observing what those differences or trends were. This is the power of the difference-in-differences approach.

Of course, this wasn't magic. As we've said, in order for this approach to be valid, we need the parallel trends condition—that the over time trend in outcomes (and, thus, confounders) would have been the same across units, but for the change in treatment status—to hold. But this is typically a weaker assumption than assuming we've actually controlled for all possible confounders. For instance, in our example, we're not assuming that New Jersey and Pennsylvania are the same (or that we've directly controlled for any differences) absent any differences in minimum wage. We're also not assuming that there are no time trends. Instead, we're assuming that the trends are parallel: whatever time trends affect employment do so in the same way in both New Jersey and Pennsylvania, at least in expectation.

Difference-in-differences has a lot going for it, and there are a lot of situations where we think this parallel trends condition is quite plausible. This design accounts for all time-

	January 1992 <i>NJ and PA</i> <i>low minimum wage</i>	November 1992 <i>NJ high minimum wage</i> <i>PA low minimum wage</i>	Difference <i>November - January</i>
NJ	20.44 FTE/restaurant	21.03 FTE/restaurant	.59 percentage points <i>effect of high minimum wage</i> + <i>over time trend</i> + <i>noise</i>
PA	23.33 FTE/restaurant	21.17 FTE/restaurant	-2.16 percentage points <i>over time trend</i> + <i>noise</i>
Difference <i>NJ - PA</i>	-2.89 percentage points <i>differences between states</i> + <i>noise</i>	-.14 percentage points <i>effect of high minimum wage</i> + <i>differences between states</i> + <i>noise</i>	Difference-in-Differences $.59 - (-2.16) = -.14 - (-2.89)$ $= 2.75$ <i>effect of high minimum wage</i> + <i>noise</i>

Table 13.3: Difference in Differences for fast-food employment

invariant differences between units that would plague a cross-sectional analysis. It also accounts for all of the time-specific factors that would plague a “before-and-after” analysis. What it does not account for is time-varying differences between units. And these are only a problem if they vary in ways that correspond with the treatment. For example, if New Jersey increased its minimum wage because they thought the economy was about to experience a boom relative to neighboring states, then this would be a violation of the parallel trends assumption.

Of course, even if the parallel trends assumption seems conceptually reasonable, just looking at 2 units is not particularly illuminating. Surely lots of idiosyncratic differences pop up in any two places in any two years, so the noise in the estimate is likely to be large. To do better, we need to extend the intuition we developed in this simple example to situations where we observe more than 2 units over more than 2 time periods in which policy changes occur.

N Units and 2 Periods

To start extending the intuition, suppose there are lots of units (e.g., maybe we have data on employment and minimum wage from all 50 states) but still just two time periods. And suppose that some of the units never received the treatment while other units received the treatment in the second but not the first period. We still want to look at changes for units that experience a change in treatment and compare those to changes for units that did not experience a change in treatment. Here are three different options, all of which are algebraically identical and will, thus, provide the same answer:

1. **By Hand:** Just as we did in the example above, calculate the average outcome in each period separately for those that never received the treatment and those that got the treatment in the second period and calculate the difference-in-difference by hand.
2. **First Differences:** Put the data into a spreadsheet with 1 row per unit (this is called *wide format*). Calculate the change in the outcome and the change in the treatment for each unit, and regress the former on the latter. The change in treatment will be 0 for the units that never changes and 1 for those that did change. So we're just comparing the average change for these two groups.
3. **Fixed Effects Regression:** Put the data into a spreadsheet with 1 row per unit-period (this is called *long format*). Regress the outcome on the treatment while also including dummy variables for each unit and time period. In this example, we would have a dummy variable for time that takes a value of 1 if the observation is in period 2 and 0 if the observation is in period 1. We would also have separate dummy variables for each unit. So the dummy variable for unit i would take the value 1 if the observation involved unit i and 0 if it involved a different unit (there would be one such dummy variable for each unit). We often call these dummy variables *fixed effects*. For instance, if an analyst says they included *state fixed effects* in a

regression, they just mean that they included a separate dummy variable for each state. Including these fixed effects insures that we're removing all average differences between units and all average differences over time, and once we've done that, the coefficient associated with the treatment variable is just the difference-in-differences.

Let's look at a fun example with multiple units. Matthew Gentzkow and Jesse Shapiro were interested in how watching television as a pre-schooler affects future academic performance (among other questions). The problem, of course, is that how much television a kid watches is affected by all sorts of factors that also affect future school performance. So a simple comparison of TV watchers to non-TV watchers isn't apples-to apples. To get at the causal relationship more credibly, they used variation in the timing with which TV originally became available in different locations in the United States. We are going to simplify what they did so you can their basic idea.

Broadcast television first became available in most US cities between the early 1940s and the early 1950s. Happily, in 1965, there was also a major study of American schools (called the Coleman Study) that, among other things, recorded standardized test scores for over 300,000 6th and 9th graders. A 9th grader in 1965 was in pre-school in approximately 1955. A 6th grader in 1965 was in pre-school in approximately 1958. Gentkow and Shapiro use both the over time roll out of TV and the Coleman data to learn about the effect of pre-school television watching on test scores.

Let's imagine that we have the Coleman data on test scores for the 6th and 9th graders in two types of towns. Towns in group A first got TV in 1953. So they had TV when both the 6th and 9th graders in the Coleman study were in pre-school. Towns in group B didn't get TV until 1956. So they had TV when the 6th graders were in pre-school but not when the 9th graders were. Overall, then, Table 13.4 summarize the way the observed data look.

	9th graders in 1965 <i>pre-school in 1955</i>	6th graders in 1965 <i>pre-school in 1958</i>
A Towns <i>TV in 1953</i>	Avg Test Scores 9A	Avg Test Scores 6A
B Towns <i>TV in 1956</i>	Avg Test Scores 9B	Avg Test Scores 6B

Table 13.4: TV and Test Scores Data Structure

If you want to learn about the effect of having access to TV as a preschooler on future academic achievement, a first comparison you might think to make is to compare the test scores of the 9th graders in the B towns (who couldn't watch TV in pre-school) to the test scores of the 9th graders in the A towns (who could watch TV in preschool). You could do that by simply subtracting the average test score of a 9th grader in a B town from the average test score of a 9th grader in an A town, as represented in Table 13.5.

	9th graders in 1965 <i>pre-school in 1955</i>	6th graders in 1965 <i>pre-school in 1958</i>	Difference
A Towns <i>TV in 1953</i>	Avg Test Scores 9A	Avg Test Scores 6A	
B Towns <i>TV in 1956</i>	Avg Test Scores 9B	Avg Test Scores 6B	
Difference	$9A - 9B$ <i>effect of TV</i> + <i>town differences</i> + <i>noise</i>		

Table 13.5: Comparing 1965 9th graders in towns that did and didn't have TV in 1955.

But we already know lots of reasons why we cannot interpret that as an unbiased estimate of the causal effect of having access to TV in preschool. These two types of towns might be different in all sorts of ways, besides when broadcast TV showed up, that matter for academic performance. For instance, maybe they have different average quality schools, different industries, or what have you. And since, in this example, the type of town and the treatment are perfectly correlated, any such difference between the towns can be thought of as a confounder.

Another comparison we might make is to look at the difference in test scores in the B towns between the 9th graders and the 6th graders, since the 6th graders had access to TV in pre-school but the 9th graders did not, as illustrated in Table 13.6.

	9th graders in 1965 <i>pre-school in 1955</i>	6th graders in 1965 <i>pre-school in 1958</i>	Difference
A Towns <i>TV in 1953</i>	Avg Test Scores 9A	Avg Test Scores 6A	
B Towns <i>TV in 1956</i>	Avg Test Scores 9B	Avg Test Scores 6B	$6B - 9B$ <i>effect of TV</i> + <i>cohort differences</i> + <i>noise</i>
Difference			

Table 13.6: Comparing 1965 6th graders to 1965 9th graders in towns that got TV in 1953.

This approach has the advantage of holding fixed the type of town, so we no longer need to worry about systematic cross-town differences. But now we have a new concern. Maybe the 9th grade and 6th grade cohorts differ in their test performance for other reasons—e.g., because 9th graders are older, or because of cohort-specific differences. Any systematic

over-time or cohort differences would be a confounder in this comparison.

Table 13.7 sums up the two ideas we've had thus far and explains why neither gets us a credible estimate of the true effect in terms of our favorite equation.

	9th graders in 1965 <i>pre-school in 1955</i>	6th graders in 1965 <i>pre-school in 1958</i>	Difference
A Towns <i>TV in 1953</i>	Avg Test Scores 9A	Avg Test Scores 6A	
B Towns <i>TV in 1956</i>	Avg Test Scores 9B	Avg Test Scores 6B	6B – 9B <i>effect of TV</i> + <i>cohort differences</i> + <i>noise</i>
Difference	9A – 9B <i>effect of TV</i> + <i>town differences</i> + <i>noise</i>		

Table 13.7: Two comparisons that do not result in unbiased estimates of the effect of TV.

But, just as with the minimum wage example, we can do better. Start by thinking about the comparison between 9th graders in the two types of towns. The problem with that comparison is that it reflects both the effect of TV exposure and any other systematic differences between the types of towns. But suppose those baseline differences between the types of towns aren't changing over time. Then the difference in academic performance between the 6th graders in the two types of towns, all of whom had access to TV in pre-school, reflects those same cross-town differences, but without the effect of TV. So we can use that difference between the 6th graders to estimate the cross-town differences. And then, subtracting the difference between the 6th graders from the difference between the 9th graders (i.e., calculating the difference-in-differences) will leave us with just the effect of TV exposure in pre-school (plus noise).

The same procedure works if we start from our comparison of 9th graders and 6th graders from the B towns. The problem with that comparison is that it reflects both the effect of exposure to TV in pre-school and any baseline differences between the 6th and 9th grade cohorts that matter for academic performance (plus noise). But suppose those over time or cohort trends are the same in the A towns and B towns. Then the difference in academic performance between 6th and 9th graders in A towns is an estimate of the over time or cohort trend without any effect of TV (since both sets of kids had access to TV in pre-school in Town A). So subtracting the difference in test scores in the A towns from the difference in test scores in the B towns will again leave us with an unbiased estimate of the effect of pre-school TV exposure.

As shown in Table 13.8, either way we do this calculation, we find the same answer.

	9th graders in 1965 <i>pre-school in 1955</i>	6th graders in 1965 <i>pre-school in 1958</i>	Difference
A Towns <i>TV in 1953</i>	Avg Test Scores 9A	Avg Test Scores 6A	$6A - 9A$ <i>cohort differences</i> $+ \text{noise}$
B Towns <i>TV in 1956</i>	Avg Test Scores 9B	Avg Test Scores 6B	$6B - 9B$ <i>effect of TV</i> $+ \text{cohort differences}$ $+ \text{noise}$
Difference	$9A - 9B$ <i>effect of TV</i> $+ \text{town differences}$ $+ \text{noise}$	$6A - 6B$ <i>town differences</i> $+ \text{noise}$	Difference-in-Differences $(6B - 9B) - (6A - 9A)$ $= (9A - 9B) - (6A - 6B)$ <i>effect of TV + noise</i>

Table 13.8: How difference-in-differences might give an unbiased estimate of the effect of TV.

For those interested in the answer, Gentzkow and Shapiro find evidence that, during the 1950s, having access to TV in pre-school was actually beneficial for average test scores, especially for kids from poorer families. Of course, this was at a time when kids watched shows like *Howdy Doody*. So you might not want to immediately extrapolate to the effects of YouTube.

More importantly, for our purposes, is seeing the power of the difference-in-differences approach. By calculating the difference-in-differences, we were able to effectively account for systematic differences between towns and over time (or cohorts), without ever observing what those differences or trends were.

N Units and N Periods

Suppose you have more than two periods and suppose that the treatment is changing at different times for different units. What do you do?

Much of the logic from the above discussion still applies. Of course, option 1 above (calculating the difference-in-differences by hand) no longer works. But you can still use option 2 (first differences) or option 3 (fixed effects). However, first differences and fixed effects are no longer mathematically identical, and will not necessarily give you the same answers, once you move beyond 2 periods. What's the difference? With first differences, you're regressing period-to-period changes in the outcome on period-to-period changes in the treatment. With fixed effects, you're regressing the outcome on the treatment while controlling for all fixed characteristics of units and time periods. Both are doing the same basic thing, but they are using slightly different kinds of variation.

Which specification makes more sense depends on the specific context. In general, the fixed effects strategy is more flexible. For instance, it allows you to include additional time-varying control variables in the regression (if necessary), and it also allows you to conduct

some helpful diagnostics. Importantly, the first-differences specification requires that you know the precise timing of the effect. If, for example, it takes some time for the effect of the treatment to set in, or if the effect size decays over time, first-differences can give you a highly misleading answer while the fixed effect strategy will still give you a meaningful estimate of the average effect.

To better understand the logic of the fixed effects strategy, it might help you to think back to Chapter 10 and what it means to control for a variable in a regression. By including controls for each time period and each unit, we're effectively removing the fixed differences in average outcomes between units and over time. That's also exactly what we were doing in the tables above when we calculated difference-in-differences by hand. When we include those fixed effects in a regression, the only remaining variation in outcomes are deviations from the time trends and deviations from each unit's average.

Even if you don't understand the technical details of fixed effects regressions, the intuition of difference-in-differences designs should be clear from our examples. And it is an important intuition. If someone shows you that some treatment of interest is correlated with an outcome of interest, you are already skeptical because of what we learned in Chapter 9. Difference-in-differences allows you to check whether changes in the treatment are also correlated with changes in the outcome. If they are, then that might be more compelling evidence of a causal relationship. And if they aren't, then the original correlation was likely the result of confounding.

Useful Diagnostics

As we've said, for difference-in-differences to yield an unbiased estimate of an average treatment effect, we need parallel trends. That is, in the counterfactual world where the treatment did not change, the difference in average outcomes would have stayed the same between

the units where the treatment did in fact change and the units where it did not. Since we don't observe that counterfactual world, we can't know if that's true. So a careful analyst always wants to do whatever is possible to probe the plausibility of parallel trends.

One conjecture is that if parallel trends holds, we should see similar trends in outcomes in earlier periods, before any units changed treatment status. We can check these pre-treatment trends (often called *pre-trends*) directly by comparing the trend in outcomes for units that do and do not change treatment status later on. We can also do it in a regression framework by including a *lead treatment* variable—that is, a dummy variable indicating the treatment status in the *next* period. If the trends are indeed parallel prior to the change in treatment, the coefficient on the lead treatment should be zero and the coefficient on the treatment variable should not change when we include that lead treatment variable in the regression.

We can also relax the requirement of parallel trends a bit by allowing for the possibility that different units follow different linear trends over time to see if this changes our results. The specific details for how you implement this are not important for now (you can read about them in a more advanced book). But you can see that there are various strategies for probing a difference-in-differences analysis to see whether parallel trends seem plausible.

But remember that diagnostic tests of this sort are a complement, not a substitute, for clear thinking. The most important defense of an assumption like parallel trends must be a substantive argument. Why did the treatment change in some units and not in others? Does that reason seem likely to be related to trends in the outcome, or independent of trends in the outcome? Can you think of reasons that units might have changed their treatment right as they expected the outcome to change for other reasons? These are critical questions whose answers require deep substantive knowledge of your context, question, and data. Good answers are absolutely essential to assessing how convincing the estimates that come out of a difference-in-differences are.

To get a better sense of the kinds of questions that can be answered using a difference-in-differences design, and of how one thinks through questions about parallel trends, let's look at a couple examples.

Compulsory Voting in Australia

There's a big debate in democracies about the extent to which low turnout biases election results and public policies away from public preferences. In the United States, for instance, rich, educated, old, and white citizens are much more likely to turn out to vote. So we might worry that election results and public policies are quite different than what they would be in the counterfactual world in which everyone votes. How could we try to assess these questions empirically?

One strategy would be to survey voters and ask them how they would vote if we forced them to. Of course, they may not tell us the truth or even know themselves how they would vote if they actually took the time to learn about the candidates and go to the polls.

Alternatively, we could compare policy in places with high turnout (like Minnesota) versus low turnout (like Mississippi). But presumably that's not an apples-to-apples comparison. The places with high turnout might just be wealthier, more progressive, colder, or what have you. We'd have no way of separating the effects of turnout from these baseline differences.

Ideally, we'd like to run an experiment. Perhaps we could force everyone to vote in some places and not do so in others to see how turnout affects electoral and policy outcomes. This kind of experiment is probably not feasible. But maybe we can find a natural experiment that comes close.

Australia has a compulsory voting law, meaning that there is a financial penalty for failing to vote. It is reasonably well enforced, over 90% of eligible citizens vote in ev-

ery election. Therefore, Australia is pretty close to that counterfactual world of universal turnout. Maybe we could go to Australia to learn about the effect of their compulsory voting policy. Anthony has done just this in one of his own studies.

Fortuitously, Australia's 6 states implemented compulsory voting for their state assembly (like a state legislature in the U.S.) elections at different times. Of course, it's not random which states adopted the policy first. But it might be reasonable to assume that the trends in baseline electoral and policy outcomes are parallel, on average, across states, and that states did not strategically pass compulsory voting laws just as they otherwise expected political outcomes to shift for other reasons.

One potentially relevant fact is that compulsory voting was passed unanimously by all parties in 5 of the 6 states, so it's not as if compulsory voting was a partisan power grab or some such. Also, there's no evidence that the passage of compulsory voting followed differential trends, and compulsory voting doesn't appear to correspond with changes in observable characteristics of the state, such as demographics.

As it turns out, difference-in-differences results suggest that compulsory voting had significant electoral and policy effects in Australia. Compulsory voting dramatically increased turnout (as expected) which in turn increased the vote share and the seat share of the Labor Party. It also looks like the adoption of compulsory voting for federal elections caused a significant increase in government provided pensions.

Is Obesity Contagious?

Humans are social animals. We live embedded in a complex web of relationships. Increasingly, we are told, our networks define who we are. A growing body of research claims to measure exactly how our thinking, tastes, and behavior are determined by our social networks.

Perhaps the most well known of this research is authored by Nicholas Christakis and James Fowler. What is striking about Christakis and Fowler's work is that they find that behaviors and personal characteristics that many of us think of as profoundly personal appear to be network characteristics. For instance, Christakis and Fowler report that smoking, drinking, happiness, and obesity all spread in social networks. Or, to use their more colorful language, "obesity is contagious" (Connected, p. 111). Christakis and Fowler believe the evidence shows that if your friends (or the friends of the friends or your friends) become happier or fatter, this causes you to be more likely to become happy or fat, perhaps because you imitate the behavior of your social network, perhaps because a new norm emerges about acceptable body types, or perhaps for other reasons.

It seems indisputable that your behavior is affected by those with whom you interact, that their behavior is affected by those with whom they interact, and so on. In this sense, we are entirely with Christakis and Fowler—we are all influenced by our social networks. But these authors, and many other scientists who study network effects, are making a claim stronger than just the commonsensical observation that our interactions affect how we behave. They are claiming to measure and quantify that effect.

In a study of the spread of obesity in social networks, published in the *New England Journal of Medicine*, Christakis and Fowler ask whether changes in a person's weight are correlated with changes in their friends', family members', or neighbors' weight. They make these comparisons controlling for personal characteristics like age, gender, and education.

What do they find? The chance that a person became obese is 57% higher if that person has a friend who became obese than if that person does not have a friend who became obese. Friendship seems to matter more than familial ties when it comes to weight gain. If a person has a brother or sister who becomes obese, that person's chances of becoming obese increased by 40%. If a person's spouse becomes obese, that person's chances of becoming obese increase by 37%. Having obese neighbors has no effect. Based on these findings, the

New York Times declared in a front page article, “the way to avoid becoming fat is to avoid having fat friends.” Christakis and Fowler didn’t love this particular interpretation. Instead, the Times reported, Christakis suggested, “why not make friends with a thin person. . . and let the thin person’s behavior influence you and your obese friend?”

Christakis and Fowler’s approach is effectively a difference-in-differences design. They test how changes in one person’s obesity correspond to changes in another person’s obesity. So if we want to think clearly about whether these are credible estimates of the causal effect, we need to think about whether we find the assumption of parallel trends plausible.

Recall what parallel trends says here. It requires that, in the counterfactual world where there was no change in treatment (i.e., no one’s friends became more or less obese), the trend in outcomes (personal obesity) would have been the same on average among people who in fact experienced a change in treatment (i.e. whose friends’ obesity changed) and people who did not experience a change in treatment (i.e., whose friends’ obesity did not change). If parallel trends holds, then Christakis and Fowler’s difference-in-differences yields an unbiased estimate of the causal effect of your friends’ obesity on your obesity. But if the trends are not parallel, then their estimates are biased, since some of the difference-in-differences they are observing and attributing to network effects would have happened even if your friends’ obesity hadn’t changed.

One concern often raised about studies of network effects is what medical researchers call homophily. People with similar characteristics tend to group together. Suppose you find that people whose friends are smokers are more likely to be smokers themselves. Social network researchers might want to interpret this as evidence that having friends who smoke causes you to become more likely to smoke. But, for that conclusion to be warranted, we’d have to be comparing apples to apples. That is, other than how their friends behave, the people in social networks of smokers and the people in social networks of non-smokers would have to be essentially the same. If members of networks of smokers were already more likely

to be smokers than networks of non-smokers, we'd be comparing apples to oranges.

It seems entirely plausible (indeed likely) that people who are members of networks of smokers are, independent of their friends, already more likely to be the sort of people who smoke because of homophily. Smokers might well meet their friends in bars that allow smoking, in the outside area at work or school where people gather to smoke, or in other smoker-friendly environs. Put differently, it might not be that having smoking friends causes you to smoke. It might be that being a smoker causes you to have smoking friends. Because people don't choose their social networks randomly, when we compare people in smoking networks to people in non-smoking networks, we aren't comparing apples to apples.

Now, by comparing changes, Christakis and Fowler are accounting for the possibility that obese people tend to be friends with each other and smokers tend to hang out together. This is one of the great things about difference-in-differences! Their finding is more compelling than that. They show that when one person *becomes* obese, their friends are also more likely to *become* obese. But we still might worry about parallel trends. Homophily might suggest that people who are on the path to becoming obese (perhaps because they have similar diets, exercise habits, genetic predispositions, cultural pressures, etc.) are more likely to be friends with each other.

We can't know whether that is the case. But there is some evidence that points toward the possibility that difference-in-differences is not unbiased here. Ethan Cohen-Cole and Jason Fletcher conducted a study of the spread of two individual characteristics—height and acne—in social networks. Using the same difference-in-differences approach that Christakis and Fowler use to argue for the social contagion of divorce, loneliness, happiness, obesity, and many other things, Cohen-Cole and Fletcher find that both height and acne appear contagious in social networks. Knowing what we do about height and acne, it is pretty hard to believe that their spread is actually caused by social interactions within a network. This is Cohen-Cole and Fletcher's point. Height and acne likely don't spread in a social

network. Instead, their apparent social contagion almost surely results from violations of parallel trends, perhaps due to homophily. Having friends with acne doesn't give you acne; people at high-risk for acne tend to hang out together. The same may well be true for obesity, divorce, happiness, and so on.

To be clear, we're not saying that we think there are no causal network effects. Indeed, we're certain there are. Furthermore, Christakis and Fowler's study is surely more convincing because they compared changes to changes, rather than simply showing that obese people are more likely to be friends with each other. But there are lots of ways in which parallel trends could be violated. So we must be cautious and think clearly about those possibilities before interpreting the results of a difference-in-differences design as an unbiased estimate of the true causal effect.

Key Terms

- **Difference-in-differences:** A research design for estimating causal effects when some units change treatment status over time but others do not.
- **Parallel trends:** The condition that potential outcomes under no treatment follow the same trend in the units that do and do not change treatment status. This says that outcomes would have followed the same trend had it not been for some unit's changing treatment status. If parallel trends doesn't hold, difference-in-differences does not provide an unbiased estimate of an average treatment effect.
- **First Differences:** A statistical procedure for implementing difference-in-differences. It involves regressing the change in outcome for each unit on the change in treatment for each unit.
- **Wide format:** A description of a data set where each row corresponds to a unique

unit.

- **Fixed effects regression:** A statistical procedure for implementing difference-in-differences. It involves regressing the outcome on the treatment while also including dummy variables (the *fixed effects* for each time period and for each unit.
- **Pre-trends:** The trend in outcomes before any unit changes treatment status. If pre-trends are not parallel, it is harder to make the case that the parallel trends condition holds.
- **Lead treatment variable:** A dummy variable indicating that treatment status in a unit will change in the next time period.

Exercises

1. For years, the state of Illinois has administered the Illinois State Aptitude Test (ISAT) to 3rd, 5th, and 8th graders. For most of this time, the test was relatively low stakes—not tied to promotion to the next grade, teacher compensation, school resources, and so on. The stakes changed in 2002, when the ISAT became the test that the Chicago Public School’s used to comply with the federal No Child Left Behind law.

Consider two cohorts of students: students who were fifth graders in 2001 and students who were 5th graders in 2002. Both of these groups of students took the ISAT in 3rd grade when it was low stakes. The students who were in fifth grade in 2001 also took the ISAT in 5th grade when it was low stakes. But the students who were in fifth grade in 2002 took their second ISAT when it was high stakes. Make a two-by-two table showing how we could learn about the average effect of high-stakes testing on student test scores using a difference-in-differences design if we had data on the average test scores of these two cohorts of students when they were fifth and third graders.

2. Suppose we want to estimate the extent to which the policy positions of Democratic and Republican candidates for Congress diverge. In other words, we'd like to know how differently the Democratic and Republican candidate would represent the same set of constituents.
 - (a) Suppose we measured how conservatively each member of Congress voted on bills and ran a regression of roll-call voting on an indicator for being a Republican. Would this be a satisfying way to estimate divergence? What kinds of bias would you worry about?
 - (b) Download data from [INSERT LINK HERE](#) on congressional elections and roll-call voting behavior. Using only the variables available in the provided data set, try to estimate divergence by controlling for confounders. If it helps, you may want to only analyze one congressional session at a time.
 - (c) Using the data available, now estimate divergence using a regression discontinuity design. Again, you might find it helpful to focus on one congressional session at a time.
 - (d) Finally, estimate divergence using a differences-in-differences design.
 - (e) Compare and contrast these three different approaches. Which one estimates divergence with the most defensible assumptions? How much do your estimates depend on your design?

Notes

The study on the effect of increasing the minimum wage in New Jersey is

David Card and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American*

Economic Review 84(4):772–793.

The study on television and academic performance is

Matthew Gentzkow and Jesse M. Shapiro. 2008. “Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study.” *Quarterly Journal of Economics* 71(3):279–323.

The study on compulsory voting in Australia is

Anthony Fowler. 2013. “Electoral and Policy Consequences of Voter Turnout.” *Quarterly Journal of Political Science* 8(2):159–182.

The study of contagious obesity is

Nicholas A. Christakis and James H. Fowler. 2007. “The Spread of Obesity in a Large Social Network over 32 Years.” *New England Journal of Medicine* 373:370–379.

The study on legal access to alcohol and motor vehicle deaths is

Christopher Carpenter and Carlos Dobkin. 2011. “The Minimum Legal Drinking Age and Public Health.” *Journal of Economic Perspectives* 25(2):133–156.

Chapter 14

Assessing Mechanisms

What You'll Learn

- Estimating an average causal effect doesn't tell us why or how that effect arises.
- Learning about the mechanisms underlying a causal effect is harder than it seems.
Generally, we can't just measure potential mechanisms and learn which ones are most important.
- Combining theory, measurement, and clear thinking can help.

As we discussed in Chapter 4, when we say that a treatment affects an outcome, all we mean is that a change in the treatment changes the outcome. But that effect need not be direct or proximate—the effect of some event on some outcome could be the result of a long chain of relationships. So, in many settings, even once we’ve credibly estimated a causal relationship, we might remain uncertain why or how the treatment affects the outcome.

For example, suppose we found that attending charter schools rather than regular public schools causes an increase in the likelihood students go to college. That’s an interesting, policy-relevant finding. It tells us that, on average, charter schools are helping students. But it doesn’t tell us *how*. Maybe the curriculum is more innovative, maybe students benefit from having more motivated peers, maybe discipline is stricter, maybe the facilities are nicer, maybe there are more advanced placement classes, maybe the students are better prepared for standardized tests, maybe there are more opportunities for after school enrichment, or maybe the student themselves work harder.

Randomizing students into charter schools versus public schools, or employing some other compelling research design, allows us to assess the average effect of going to a charter school. But it won’t unpack which mechanisms are doing the work. Sometimes, understanding the mechanisms is important. For instance, if we are going to try to build more charter schools, we would like to know which features of existing charter schools are the most important to replicate. Should we make sure that there are nicer facilities, more advanced placement (AP) classes, more standardized test prep, or stricter discipline? In this chapter, we consider some of the issues surrounding trying to learn not just about effects, but about the mechanisms underlying those effects.

Causal Mediation Analysis

One approach that some researchers take to try to get at causal mechanisms is called *causal mediation analysis*. In causal mediation analysis, the goal is for the data to directly tell us how important a role some mechanism plays in driving a particular effect.

Suppose, for instance, that we want to know how much of the effect that attending a charter school has on college admission is due to charter school student's taking more AP classes. In our earlier language, attending a charter school is treatment and college admission is the outcome. We call taking AP classes a *mediator*—a mechanism by which charter schools have their effect on college admission. The idea is that attending charter schools affects taking AP classes and taking AP classes affects college admission. So some of the effect of attending a charter school on college admission runs through the AP class mechanism. We'd like to know how much of the effect is due to AP classes.

Here is an intuition for how you might be tempted to use the techniques from Chapter 10 to try answer this question if you weren't thinking clearly. Start by using an admissions lottery to estimate the effect of attending the charter school on college admission. You can do so by regressing college admission on winning the charter school admissions lottery for the set of students who entered the lottery (i.e., wanted to attend the charter school in the first place). Then measure the number of AP classes each student took. Now re-run the regression of college admission on winning the charter school lottery, but controlling for AP classes. The thought is that, if the estimated effect of charter schools shrinks once we control for AP classes, then that portion of the effect that disappeared is attributable to the AP class mechanism since, by controlling for AP classes, we are effectively holding them constant so that they are statistically removed from our estimate of the effect.

This idea seems sensible on its face. We are trying to figure out how much of the effect of charter schools on college admission runs through AP classes. So we'd like to compare

the effect of charter schools on college admissions to the effect of charter schools on college admissions *purged of the effect of taking AP classes*. So why won't it work?

The basic problem is that controlling for AP classes is not the same as purging their effect. You can see this just by thinking conceptually. To purge the effect of AP classes from the effect of attending a charter school on college admission, surely we must have a way of estimating the effect of taking AP classes on college admissions. But we have described no research design to do so. Let's try to think that through as clearly as we can.

The charter school admissions lottery lets us estimate the effect of charter school attendance on college admissions. It also lets us estimate the effect of charter school attendance on taking AP classes. But what lets us estimate the effect of taking AP classes on college admissions? We are sure that you can think of lots of confounders for the relationship between how many AP classes a student takes and whether that student is admitted to college. So just regressing college admissions on AP classes surely won't do it.

To see why this is a problem, let's think about an extreme version of what could go wrong. Suppose that charter schools indeed cause students to take more AP classes, but that AP classes have no effect whatsoever on college admissions. (This is, of course, just for the sake of argument. You should take hard classes.) So, if our causal mediation analysis works to identify the importance of the mechanism, then we should find no difference between the effect of charter schools attendance on college admissions, whether or not we control for AP classes (since it is not in fact one of the mechanisms). But suppose taking AP classes hard happens to be correlated with academic talent (which we don't have a measure of) and that academic talent affects college admissions. Now, when we run the regression of college admissions on charter school attendance and AP classes, we will find that the estimated effect of going to a charter school is indeed lower when we control for AP classes. This is because AP classes are proxying for academic talent. From this statistical result, we will wrongly conclude that allowing students to take AP classes is, therefore, an important

mechanism by which charter schools cause college admissions. But, in fact, we've stipulated that AP classes have no effect. Taking AP classes just happens to be correlated with things like talent that are also correlated with getting into college. Our causal mediation analysis, therefore, was misleading. If we don't think clearly about this type of analysis, we could end up making bad decisions about how to allocate resources or design schools.

A literature delves into what kinds of conditions you would need to have causal mediation analysis work. Without going into the technical details, it boils down to something along the lines of having research designs that allow you to separately estimate the effect of the treatment on the outcome, the effect of the treatment on the mechanism, and the effect of the mechanism on the outcome. If you can estimate these quantities, you can net out the effect of the treatment on the outcome that runs through the mechanism. The key takeaway, of course, is that there is no magical technical or statistical way to identify which mechanisms matter. If we want to learn about mechanisms, just like when we wanted to learn about causal effects, we are going to have to work hard and think clearly.

Intermediate Outcomes

One thing an analyst can do is test for the effect of their treatment on other, perhaps more intermediate outcomes which might provide some hints about mechanisms. Once you have a research design that allows you to estimate the effect of some treatment, it can, in principle, be applied to any downstream outcome that you can measure.

So, to partially assess mechanisms, we can see which intermediate outcomes appear to be affected by the treatment. Returning to our example, we can use the charter school lottery to assess the effect of going to a charter school not just on college admissions, but on intermediate outcomes like study habits, extracurricular participation, standardized test scores, taking AP classes, and so on. Of course, this doesn't give us a way to assess the

effect of those intermediate outcomes on the final outcome (here, college admissions). For that, we'd need a separate research design. So, for reasons just discussed in the last section, this approach won't tell us exactly how much of an effect is explained by one mechanism or another. But it can allow us to make some progress in thinking about which mechanisms likely do or do not help us understand the effect. For instance, if it turns out that attending a charter school has no effect on taking AP classes, it seems unlikely that AP classes are one of the mechanisms by which charter schools affect college admissions.

One real-world example of this comes from research by Chris Blattman, Julian Jamison, and Margaret Sheridan. The authors randomly assigned subjects in Liberia to behavioral therapy with the hope of improving economic outcomes and reducing crime and violence. The therapy appears to have worked well, significantly improving both outcomes. That is good news for the therapy program Blattman, Jamison, and Sheridan were studying. But in addition to knowing that the therapy worked, it would be nice to know *why* it worked. The treatment lasted as long as eight weeks and included work on a variety of skills, from self control to personal appearance. Some of the interventions even included monetary compensation. So there are a lot of different features of the treatment that could explain the result. And if you were going to try to transport this evidence elsewhere, you'd want to know what features of this program are important and which ones are incidental.

As we've discussed, there's no way to know with certainty which features of the intervention led to the overall effect. But the authors did measure a number of intermediate outcomes that might help to elucidate the mechanism and aid practitioners who would like to apply the lessons elsewhere. Interestingly, the authors found little evidence that the therapy affected subjects' self-control skills, such as impulsiveness, perseverance, or conscientiousness. Therefore, they conclude that these self-control skills are unlikely to be an important mechanism through which the therapy reduced violence or improved economic outcomes. By contrast, they do find that their intervention had a large effect on other in-

intermediate outcomes, such as social networks and attitudes toward violence. This suggests that these mechanisms are more likely to explain the success of the treatment and might be more important to replicate in other settings.

To be clear, this study does not show that self-control skills have no effect on economic well-being or violence. They might have very large effects. The research instead shows that the particular intervention under study had little effect on self-control skills and therefore self-control skills are unlikely to explain the large effects of the intervention on economic well-being or violence. Furthermore, if future practitioners were hoping to replicate the success of this behavioral therapy in Liberia, they might learn from this that, for whatever reason, they won't have much success changing self-control skills, so they might be better off focusing more on other factors like social networks and attitudes toward violence, which seem to be the places where this particular approach has the most bite.

Independent Theoretical Predictions

One way to get at specific mechanisms is to think theoretically and generate independent tests that might help adjudicate between different potential mechanisms. In Chapter 7, we discussed how we could use independent theoretical predictions to probe whether some estimated effect was the result of noise (i.e., a false positive). There, we gave the example of re-examining the putative effect of college football games on elections. The idea here is similar. But now, instead of testing whether an observed relationship is genuine or spurious, we want to look for the mechanisms underlying an estimated effect that we believe is genuine.

A study by Sarah Anzia and Chris Berry illustrates how this can work. There is much concern about the possibility of discrimination in the electoral process. For example, are voters biased against female candidates? Answering this question compellingly is difficult for obvious reasons. Discriminatory voters may not reveal their prejudices in surveys. And

in real elections, there could be factors other than discrimination that could explain why women do better or worse than men.

Some scholars have noticed that in the United States, when female candidates run for office, they perform similarly to male candidates on average. This, they argue, suggests there may be little discrimination by voters. However, Anzia and Berry point out that if there is discrimination, we might expect that only really high-quality women will decide to run for office, which could also explain why women do about as well as men even in a world with discrimination. So the fact that women perform as well as men, *when they run*, doesn't necessarily imply that voters aren't discriminating.

Continuing with this line of thinking, Anzia and Berry try to generate theoretical predictions that should hold if there indeed is discrimination against women in elections. One prediction is that if there is discrimination, all else equal, the women who are elected to office should be *better at their jobs* than the men who get elected. Because of discrimination, they will have to be better in order to get elected. Of course, we don't have perfect measures of job performance. But Anzia and Berry look at several measures of how members of Congress perform, including the number of bills they sponsor and also the amount of federal spending they bring home to their districts. The result is exactly in line with the theoretical prediction. Using a difference-in-differences design, they show that, on average (accounting for differences across districts and time periods), women perform better in Congress than men, consistent with the possibility that women have to clear a higher hurdle in order to get elected because of voter discrimination.

Having uncovered an interesting and compelling phenomenon, Anzia and Berry push further still. The evidence is pretty clear that, on average, female members of Congress are more productive than male members of Congress, and discrimination is a potential explanation. But are there other mechanisms that could potentially explain this effect? What if, for example, women are just better at some parts of the job than men, regardless

of any selection or discrimination? Or what if women are treated differently once they get to Congress, not because of their ability but because they are viewed as a token minority?

To address these kinds of questions, Anzia and Berry keep thinking. If it's really discrimination that explains this result, are there further theoretical predictions that they can test? One prediction is that the gap in performance between elected women vs. men should be greater the more discrimination women face. Of course, we don't know for sure which Congressional districts discriminate more. But one reasonable hypothesis is that more conservative districts will discriminate against female candidates more than more liberal districts. Therefore, Anzia and Berry test whether the difference in performance in Congress between men and women is greater in more conservative districts, as measured by how the district votes in presidential elections. The answer is *yes*.

To provide further evidence of their purported mechanism, Anzia and Berry note that one large group of female Congressional representatives—those who gained office because they were the widow of a recently deceased member of Congress—likely did not have to overcome the same type of discrimination as other female candidates. As such, we would not expect them to be of higher quality on average. And sure enough, when they study the legislative productivity of widows, they do not perform better than male members of Congress and they are notably worse than female candidates who were elected independent of their spouse.

The compellingness of the Anzia and Berry study comes not from a single, airtight research design or statistical test demonstrating the presence of discrimination. Rather, they elucidate an interesting and important mechanism by generating a theory of discrimination in elections and testing multiple, independent predictions that follow from that theory.

Of course, this still doesn't settle the matter. Other mechanisms might also account for the observed patterns. For instance, a variety of scholars point to evidence that women may under-estimate their abilities or be averse to putting themselves forward as candidates.

These mechanisms might also account for female representatives performing better in their jobs, conditional on winning election. So there is still lots of work to do in figuring out which mechanisms are at work. But, in our view, this study provides a model for using a combination of clear thinking and data analysis to try to provide a better understanding of causal mechanisms.

Testing Mechanisms by Design

In some special circumstances, we can design studies in ways that isolate particular mechanisms. Take, for example, a clever study on voter turnout by Alan Gerber, Don Green, and Christopher Larimer. These researchers mailed postcards to a randomly selected group of registered voters. The postcards informed the recipients which of their neighbors had and had not voted in recent elections. (You may not have known this, but whether or not you voted is a matter of public record. Only how you voted is secret.) It also indicated that another, similar postcard would be sent to neighborhood residents after an upcoming election. The implication was that, if the recipient didn't vote in the upcoming election, all their neighbors would find out about it. This unusual (and perhaps invasive) postcard increased voter turnout dramatically; people who received the postcard were 8 percentage points more likely to vote than a control group.

Having seen these experimental results, we might wonder why the postcards had such large effects. Was the social aspect of the treatment important? Do people really not want their neighbors to know that they don't vote. Are people mobilized just because the postcard reminded them about the election? Or are people perhaps just trying to impress researchers and so they turn out once they know they are being studied?

To learn about the importance of the social mechanism, the researchers designed their experiment to include another randomly assigned postcard. This postcard mimicked every

feature of the previously discussed one, with one exception. Instead of containing voter turnout information about all of the recipient's neighbors, it only contained information about members of the recipient's household. Recipients of this type of postcard would no longer worry that all of their neighbors were going to find out if they didn't vote. Now it would just be the people they lived with who would learn about their voting behavior. And those people presumably already had a pretty good guess. The thought is that this small change takes a lot of the social pressure mechanism out of the intervention. And, indeed, this postcard also increased voter turnout, by about 5 percentage points relative to a control group.

The clever part of this design is that, by including multiple treatments in the experiment, the authors were able to estimate how much the social aspect of the first postcard matters. In particular, having voter turnout information made public across a whole neighborhood, instead of just within a household, appears to account for 3 percentage points of the overall 8 percentage point effect.

Disentangling Mechanisms

Sometimes, we can do a similar pulling apart of mechanisms even when we don't get to design the study ourselves. To do so, of course, we must have multiple research designs that make it possible to separately estimate the effects of different mechanisms. Let's see an example.

For decades, scholars have studied economic conditions, violent conflict, and the causal relationships between the two. It's hard to think of a topic where the stakes are higher. We would love to improve economic conditions and reduce violent conflict around the world, but we don't seem to know how to do so.

If you remember, we already discussed the difficulty of empirically assessing the effect

of the economy on conflict in Chapter 9. When we observe a strong correlation between conflict and poor economic conditions, it is unclear whether the former causes the latter, the latter causes the former, some confounding factors cause both, or some combination of all these possibilities.

To gain more traction on one part of the problem, many scholars have tried to find research designs that allow us to more credibly estimate the effect of economic conditions on violent conflict. One common strategy involves using commodity price shocks as part of a differences-in-differences design. The basic idea is as follows.

Poppy farming is a major industry in parts of Afghanistan. One might think that Afghan farmers and farm workers making money in the poppy industry would be less willing to go fight because they'd be giving up relatively good economic opportunities. But, clearly, we can't just regress the amount of violence in different parts of Afghanistan against the amount of poppy farming to learn about the relationship between conflict and economic opportunity. Confounders abound. Poppies are the key raw ingredient to heroin, so poppy farming goes along with the drug trade, which might affect violence independently. Moreover, poppies grow in mountainous areas. And the terrain might also affect the amount of violence.

Another thought is that the willingness of poppy farmers and farm workers to fight might go down when the poppy business is particularly good and go up when it is bad. But, again, such an overtime comparison has potential confounders. Perhaps a surge in demand for poppies happens to also coincide with features of the season, US troop deployments, or other factors that also affect violence.

But a difference-in-differences strategy might address both concerns. That is, we'd like to compare the amount of violence in poppy-producing vs non-poppy-producing places at times when the poppy business is good vs. bad. The idea being that by comparing poppy-producing to non-poppy producing areas, we'll take out any systematic differences over time. And then the remaining comparison within poppy-producing areas will tell us how

changes to the poppy business affect violence.

To pull off a strategy like this, of course, we need some measure of when the poppy business is good vs. bad. For that, researchers use changes to the world price of poppies (or the world price of heroin). The idea is that, for most commodities, most countries are small players. So the world price of that commodity is unlikely to be strongly influenced by things that are happening in that country. And therefore, perhaps we can use changes in the world price to estimate the effect of local economic prosperity (remember instrumental variables from our discussion of noncompliance in Chapter 11).

The idea is a pretty nice one. Using differences-in-differences designs, we can estimate the effects of economic shocks on violence more credibly than we could just by comparing violence levels in rich and poor places, or even over time.

Rather than doing this for just one commodity and one country, scholars have done the painstaking work of measuring which and how much of each of hundreds of commodities each country produces. From this, they create an index, for each country, of its *commodity bundle*. They have also collected the world price of each commodity for each year, so they can measure how much the value of each country's commodity bundle changes each year. From this, they can do a giant difference-in-differences analysis to see how violence changes in response to economic shocks the world over.

Interestingly, when scholars did this, they got a bunch of conflicting results. Sometimes it looked like positive economic shocks reduced violence, sometimes they appeared to increase violence, and sometimes they had no detectable effect. These contradictory and inconsistent findings were somewhat disconcerting. Better data and research designs are supposed to provide more, not less, definitive answers.

What is going on? Ernesto Dal Bo and Pedro Dal Bo suggested one possible theoretical explanation. They pointed out that there are at least two mechanisms through which

economic conditions might influence conflict. And those two mechanisms pull in opposite directions! On the one hand, good economic conditions create more and better jobs, and workers with those good jobs might be less willing to leave them to fight. In other words, good economic conditions increase the *opportunity costs* of fighting. If you're unemployed and hungry, then you might join a revolution, but a good-paying job might be enough to deter that inclination. On the other hand, armed groups often fight over control of economic resources. Good economic conditions mean there is more to fight over. In other words, good economic conditions increase the benefits of *predation*. If you live in a desolate place with no economic value, what's the point in fighting over it? But if there's money to be made from controlling a booming economy, maybe it's worthwhile to fight. Perhaps the fact that a shock to economic conditions can activate these opposing forces explains the murky results of previous studies.

So where do we go from here? Knowing that there are competing forces is theoretically illuminating, but it's not enough to inform policy. Fortunately, Dal Bo and Dal Bo thought more about conditions under which one force might dominate over the other. Their idea was that in labor-intensive industries and economies, the *opportunity cost* mechanism should dominate because economic shocks create more need for labor, higher wages, and better jobs. But in capital-intensive industries and economies, the predation mechanism should dominate because better economic conditions create more to fight over without meaningfully increasing wages or employment.

In a 2013 study, Oeindrila Dube and Juan Vargas found a way to empirically test these ideas using the sort of difference-in-differences strategy we've already described. They studied armed conflict in Colombia, and they focused most of their attention on two major industries: coffee and oil. Coffee is relatively labor intensive, requiring lots of workers to farm and process. Oil is relatively capital intensive, once an oil well is drilled, oil producers do not need lots of workers on hand. Importantly, some locales in Colombia

have coffee-intensive economies, while others have oil-intensive economies. Consistent with theoretical predictions, positive shocks to the world price of coffee appear to reduce conflict in coffee-intensive locales relative to non-coffee intensive locales. This is evidence in favor of the opportunity costs mechanism. By contrast, positive shocks to the world price of oil appear to increase conflict in oil-intensive locales relative to non-oil intensive locales. This is evidence in favor of the predation mechanism.

This evidence suggests that economic conditions do indeed influence violent conflict through multiple mechanisms. As such, economic improvements can either mitigate or exacerbate conflict, depending on which mechanism dominates in a given context. This means that if we find no average relationship between economic shocks and conflict world-wide, that doesn't mean economic conditions don't matter for conflict. Rather, the average relationship might be masking multiple off-setting effects that we can only understand if we disentangle the mechanisms. So, perhaps economic aid can mitigate conflict, but only when it corresponds with job opportunities. Economic aid that simply increases the size of the economic pie that warring factions can fight over, but without providing employment opportunities, will likely make matters worse. This is an important insight could only have been discovered through the interplay of data, research design, theory, and clear thinking. The data alone wasn't enough. Nor was a good research design. It took all of these tools together.

Key Terms

- **Mediator:** A feature of the world that is affected by the treatment and affects the outcome.
- **Causal Mediation Analysis:** Techniques for trying to estimate how much of the effect of a treatment on an outcome is the result of the treatment's effect on a mediator

and the mediator's effect on the outcome.

Exercises

In the 1990s, the U.S. Department of Housing and Urban Development ran a large-scale field experiment called Moving to Opportunity. They randomly selected some households living in high-poverty public housing projects and offered them housing vouchers (i.e., money that could be used to pay rent) if they moved to a low-poverty neighborhood. Other households were given nothing. The goal was to learn whether moving to a low-poverty neighborhood would be beneficial for economic, mental, and physical well-being.

Researchers have examined the data and found that households in the treated condition (receiving a voucher to move to a low-poverty neighborhood) experienced better physical health, mental health, and subjective well-being than those in the untreated condition (receiving nothing). There were not significant differences on economic outcomes.

1. This result is strong evidence that a treatment—receiving a housing voucher that you can use only by moving to a low-poverty neighborhood—causes improved well-being. Does this imply that the treatment works through the mechanism of moving people to a low-poverty neighborhood? Offer at least one other mechanism that might explain the result?
2. How could you modify or add to the experiment in order to better elicit the effect of moving to a low-poverty neighborhood?
3. There was actually one additional treatment considered in the experiment. Another group of individuals were randomly assigned to receive a housing voucher to move anywhere they liked (not just to a low-poverty neighborhood). In light of this information, what comparison(s) would you make in order to separate the effect of

moving to a low-poverty neighborhood from the potential effects of moving in general or receiving the financial benefit of a voucher?

4. *Bonus challenge:* Not surprisingly, there was some non-compliance in this experiment—some people who were assigned to the voucher treatment chose not to move. And as you can imagine, the rate of compliance was different between the two treatments. 63 percent of households took the voucher when there was no restriction on where they could move, but only 48 percent took it when it required that they move to a low-poverty neighborhood. Further complicating matters, some of the households in the unrestricted condition moved to low-poverty neighborhoods even though they didn't have to. What would you do if you wanted to estimate the effect of moving to a low-poverty neighborhood in light of these non-compliance issues? (There's no easy answer, but we hope it's illuminating to think through all the complications and appreciate how hard it is to learn about causal mechanisms).

Readings

If you are interested in learning more about causal mediation analysis, you could start with

Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309–334.

The study on behavioral therapy in Liberia is

Christopher Blattman, Julian C. Jamison, and Margaret Sheridan. 2017. Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review* 107(4):1165–1206.

The study on discrimination against women in elections is

Sarah F. Anzia and Christopher R. Berry. “The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?” *American Journal of Political Science* 55(3):478–493.

The experiment on social pressure and turnout is

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. 2008. “Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment.” *American Political Science Review* 102(1):33–48.

The studies on economic prosperity and conflict are

Ernesto Dal Bo and Pedro Dal Bo. 2011. “Workers, Warriors, and Criminals: Social Conflict in General Equilibrium.” *Journal of the European Economic Association* 9(4):646–677.

Oeindrila Dube and Juan F. Vargas. 2013. “Commodity Price Shocks and Civil Conflict: Evidence from Colombia.” *Review of Economic Studies* 80:1384–1421.

Part IV

From Information to Decisions

Chapter 15

Turn Statistics into Substance

What You'll Learn

- Statistics are often reported in ways that are misleading or unhelpful for decision making.
- If you think clearly about the question at hand, you can often translate a statistic into a piece of more substantively useful information.
- Quantitative evidence, on its own, can't tell you what to believe. Your beliefs depend on a combination of the new evidence and your prior beliefs. The way in which they get combined is called Bayes' Rule.
- Quantitative evidence, on its own, can't tell you what to do. For that, you must carefully combine your evidence-based beliefs with your values.

Quantitative analysis should provide information that helps us make better decisions. The ideas we've emphasized thus far—how to establish whether a relationship exists, mean reversion, the difference between causation and correlation, tools for estimating causal effects, and so on—are important inputs to that process. But they are not the end point.

Suppose you've estimated that some intervention has a positive causal effect on some outcome. Does that mean you should do the intervention? You can't know from a quantitative analysis alone. You need to translate that statistical information into a substantive answer to your question.

People frequently get confused on this point. It's easy to stop thinking clearly once you've got some precise and authoritative sounding quantitative finding, reaching incorrect conclusions even from correct information. In this chapter, we explore how to avoid such mistakes. The key is to turn statistics into substance, so as to make sure you are asking and answering the question you really care about.

What's the right scale?

There is more than one precise and accurate way to represent a piece of quantitative information. But they are not all equally good. How the information is presented can have an important effect on its perceived substantive meaning. For instance, changing scales can dramatically alter whether a relationship seems large or small, important or unimportant, a good reason for taking action or not. So it is important, when presented with such information, to think about whether the scale on which it is presented corresponds with the substantive question you are trying to answer, or whether reframing the relationship on some other scale might provide a better guide. To see what we mean, consider a couple examples.

Miles-per-gallon vs. gallons-per-mile

Suppose you worked for the Environmental Protection Agency (EPA) regulating emissions from automobiles. Your team brings you two proposals for new regulation to evaluate. One regulation will result in a two mile-per-gallon improvement in the fuel efficiency of small sedans. The other will result in a two mile-per-gallon improvement in the fuel efficiency of large SUVs. There are the same number of these two kinds of automobiles on the road and, on average, each gets driven an average of 10,000 miles per year. The sedans get 30 miles-per-gallon (which the regulation would improve to 32 miles-per gallon), while the SUVs get 10 miles-per-gallon (which the regulation would improve to 12 miles-per gallon). The SUV regulation will cost a little bit more to implement. And since the two regulations each offer a 2 mile-per-gallon improvement for the same number of vehicles driven the same number of miles per year, your team recommends regulating the sedans. Does this make sense?

Let's start by remembering the substantive question you want to answer. Your job is to reduce the amount of gas used by automobiles. Does improving fuel economy by 2 miles-per-gallon on sedans and SUVs translate into the same reduction in gas used? Let's turn the statistics into substance to check.

The SUVs get 10 miles-per-gallon, which means that, since the average driver drives 10,000 miles per year, on average SUVs use 1,000 gallons of gas per year ($\frac{10,000}{10}$). If you implement the regulation that improves fuel efficiency to 12 miles-per-gallon, then on average SUVs will use about 833 gallons per year ($\frac{10,000}{12}$). The 2 miles-per-gallon improvement saves 167 gallons of gas per SUV per year.

What about for the sedans? The sedans get 30 miles-per-gallon, which means that, since the average driver drives 10,000 miles per year, on average sedans use about 333 gallons of gas per year ($\frac{10,000}{30}$). A regulation that improves fuel efficiency to 32 miles-per-gallon

results in sedans using about 313 gallons of gas per year ($\frac{10,000}{32}$). The 2 miles-per-gallon improvement saves only about 20 gallons of gas per sedan per year.

It wasn't obvious until we translated the statistics into substance, but now we can see that your team's recommendation looks wrong. The same 2 miles-per-gallon improvement in fuel economy has a much larger effect on gas consumption when applied to a gas guzzling SUV than when applied to an already relatively fuel efficient sedan. So you should regulate the SUVs unless doing so is much more expensive.

Figure 15.1 shows how much gas a vehicle that drives 10,000 miles a year uses as a function of the miles-per-gallon it gets. As expected, gas consumption decreases as miles-per-gallon increase. Less intuitive, but critical for understanding the results we just showed, is the fact that the slope of this curve is really steep for low values of miles-per-gallon (less efficient cars) and much less steep for high values of miles-per-gallon (more efficient cars). You get a lot more bang for your buck improving miles-per-gallon for inefficient cars than you do for efficient cars. This has interesting implications beyond our example. In particular, moving people out of very fuel inefficient vehicles into somewhat more fuel efficient vehicles does a lot more to reduce emissions than moving people already in relatively fuel efficient vehicles into very fuel efficient vehicles like hybrids.

Returning to our example, there was nothing wrong with the quantitative information your team used to form its recommendations. Yet their recommendation was incorrect. Why? The problem came from the scale that was used to present the quantitative information. Miles-per-gallon is the most commonly reported metric for fuel efficiency in the US. But it is not a particularly helpful statistic for substantive decision making.

The substantive question we care about is how much gasoline a car burns, given how far it is driven. But miles-per-gallon tells you how far a car drives, given how much gasoline is burned. That's backwards for answering our question. You can obviously do the math, as we did just now, to turn this statistic into substance. But most people won't. In fact,

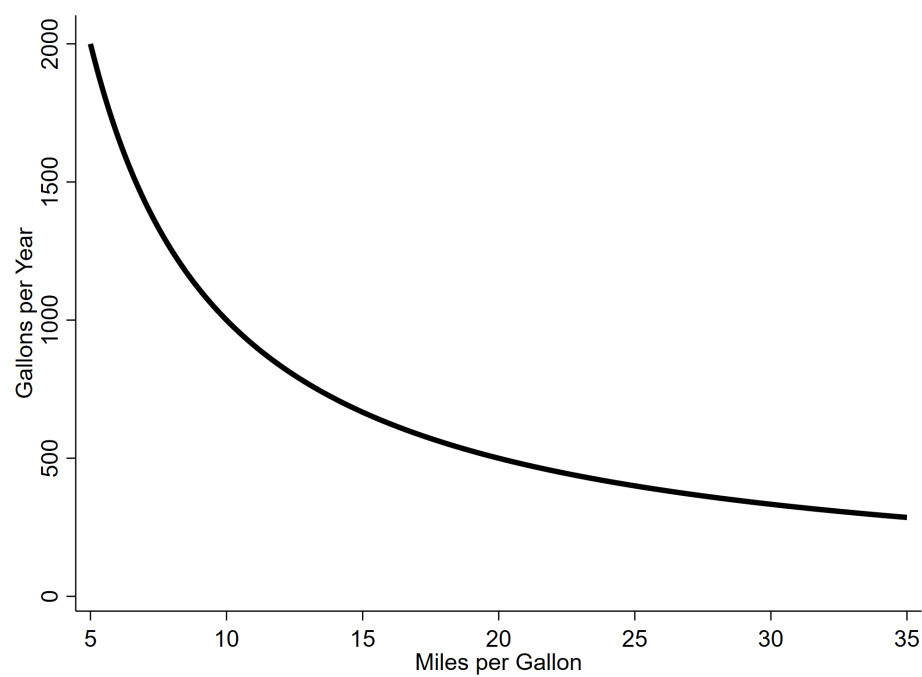


Figure 15.1: Gas consumed by driving 10,000 miles as a function of the miles-per-gallon.

most people won't even notice the distinction. And therefore, consumers and regulators alike may be confused (or tricked) into making bad decisions.

If you wanted to do a better job, you would report the quantitative information on a substantively meaningful scale, something like gallons-per-hundred-miles, instead of miles-per-gallon of gas. As we just saw, the same 2 miles-per-gallon improvement results in very different improvements in gallons-per-mile, depending on the automobile's baseline fuel efficiency. You'd have had no trouble making the right decision if your team had come to you with two regulations, one of which saved about 8.3 gallons-per-hundred-miles (the SUV regulation) and the other of which saved about 3.1 gallons-per-hundred-miles (the sedan regulation).

Indeed, in a 2008 study, Richard Larrick and Jack Soll show that the way statistics are reported can be quite consequential for important decisions. They quote an automotive expert who seems to think it's not worthwhile to try to make marginal improvements in the miles-per-gallon of large SUVs when that's, in fact, where engineers and policymakers would likely get the biggest bang for their buck in terms of mitigating emissions. Furthermore, they show that consumers' willingness to pay for improvements in fuel economy varies linearly with miles-per-gallon, despite the fact that, as we've just seen, consumers should be willing to pay a lot more for improvements on low fuel economy vehicles. Consumer willingness to pay does a much better job of tracking actual costs when efficiency is reported as gallons per hundred miles.

Percent vs. percentage point

Often, in evaluating the substantive importance of some effect, we want to know how big the effect is. There are at least two ways the size of an effect might be reported: the percent change in the outcome it induces or the percentage point change in the outcome it induces. The *percentage point* change is the simple numerical difference between two

percentages. The *percent* change is the ratio of the percentage point change to the initial value. So, for instance, moving from 20 percent to 22 percent is a 2 percentage point increase ($22\% - 20\%$) but a 10 percent increase ($\frac{2}{20}$). Which of these one reports can lead to very different perceptions of the magnitude of an effect. So it is important to check your intuitions by translating back and forth and thinking clearly about which matters for your question. Here's an example.

The Wall Street Journal reported on a medical experiment showing that a new drug reduced the “risk of heart-related death, heart attacks, and other serious cardiac problems by 44%.” A 44 percent reduction sounds big. This, coupled with the headline, “Cholesterol Drug Cuts Heart Risk in Healthy Patients,” makes it sound very important that people have access to the treatment.

But let's stop to think clearly about the substantive question we are interested in when evaluating a quantitative result like this. To determine whether it is worthwhile to give a particular treatment to a large population, we'd like to know how much the treatment costs and how many people the treatment will save. Knowing that a treatment reduces heart attacks among otherwise healthy people by 44 percent doesn't actually tell you how many people it saves. To determine that, you also need to know how frequent heart attacks are in that population in the first place.

Later in the article we learn that 250 out of the 9000 people randomly assigned to the control group, which received a placebo pill, had heart attacks over the course of the study. This suggests a baseline heart attack risk of about 2.8 percent ($\frac{250}{9000}$). A 44 percent decrease in heart attacks means going from about 2.8 percent of people having a heart attack to about 1.6 percent having a heart attack. Because heart attacks are so rare in this population, the 44 *percent* reduction in heart attacks translates into about a one *percentage point* reduction—not such a huge difference. Indeed, if the drug is expensive, you might well conclude that the treatment is not worthwhile.

Here, again, we see the value of translating statistics into substance. The article uses statistics that answer one question—does the drug cause a large percent decrease in heart disease—to which the answer is yes. But the headline makes it seem as though it is answering a much more important question—will the drug save a lot of lives?—to which the answer is probably not. By translating the statistics (the percent reduction) into substance (number of heart attacks prevented per 100 people treated), we can easily spot the difference and answer the question most relevant for decision making.

From Statistics to Beliefs: Bayes' Rule

In 1964 in Los Angeles, an elderly woman named Juanita Brooks was walking down an alley, pulling a basket of groceries with her purse resting on top, when she was pushed to the ground from behind and her purse was stolen. She didn't get a good look at the perpetrator. Around the same time, an eye witness saw a woman run out of that same alley and enter a yellow car. The witness also didn't get a great look. But he did note that the woman running was white and had a ponytail and that the driver of the car was a black man with a beard and mustache. Based on this eyewitness testimony, police later arrested Malcolm and Janet Collins and charged them with the robbery. Malcolm was a black man with a beard and mustache. Janet was a white woman with a ponytail. And they drove a yellow car.

As Jonathan Koehler relates in an article, prosecutors brought in a mathematician to testify regarding the chances that, based on this evidence alone, Malcolm and Janet were guilty of the murder. The mathematician concluded that there was only about a 1 in 12 million chance that the couple was innocent. Here was the reasoning.

If we just arrested an innocent couple at random, it's very unlikely that the husband would be black with a beard and mustache, that the wife would be white with a ponytail,

and that they would drive a yellow car. Why is this?

The argument starts with some quantitative facts. If we just picked a man at random from the population, there's a 10 percent chance that he would be black, because only about 10 percent of the U.S. population is black. Suppose that 10 percent of all men have beards, so there's also a 10 percent chance that he'd have a beard. Perhaps there's a 20 percent chance that he'd have a mustache. And there's only a 0.5 percent chance that he'd drive a yellow car, given the number of yellow cars on the road.

How do we take these numbers and turn them into an overall probability of a randomly selected innocent couple having this confluence of characteristics? To get an intuition, think about an analogy to a deck of cards. What is the probability that a randomly drawn card is the 4 of hearts? The probability that a randomly drawn card is a four is 1 in 13. And the probability that a randomly drawn card is a heart is 1 in 4. Since being a four and being a heart are independent (i.e., knowing that a card is a four doesn't tell you anything about how likely it is to be a heart and vice versa), the probability of the two characteristics occurring together is simply the product of the probability of each occurring individually. So if we draw a random card from a deck, the probability it is the four of hearts is $\frac{1}{13} \times \frac{1}{4} = \frac{1}{52}$. This makes sense. There are 52 cards in a deck. Only one of them is the four of hearts.

The prosecutor applied the same logic to the Collins couple. He argued that the chances that a randomly selected person would be black, have a mustache and beard, and drive a yellow car is the produce of the probabilities of the individual characteristics: $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{5} \times \frac{1}{200} = \frac{1}{100,000}$. He continued to add characteristics (being married, being an interracial couple, the woman having blond hair, the woman having a ponytail, etc.), eventually arriving at a probability of 1 in 12 million. Indeed, as the prosecutor pointed out, even this was an underestimate, since the couple had many other characteristics that had not been accounted for, so that the probability of innocence was probably more like 1 in 1 billion! A jury found the Collins guilty, and newspapers praised prosecutors for making

such a quantitatively compelling case.

But this analysis does not reflect clear thinking. Indeed, there is so much wrong, it is hard to know where to start. But start we must.

So, first, these characteristics (unlike being a heart and being a four in a deck of playing cards) are not independent of one another. So you can't just multiply the probabilities of each individual characteristic together to get the probability of the confluence of characteristics. For instance, having a beard is positively correlated with having a mustache. As such, the probability of having a beard and a mustache is much higher than the probability of having a beard times the probability of having a mustache. That is, if 1 in 10 men have a beard and 1 in 5 have a mustache, since many of these are the same men, many more than 1 in 50 men have a beard and a mustache. Indeed, the probability is likely much closer to 1 in 10, since almost everyone with a beard also has a mustache. So, if we took into account all the relevant correlations, maybe we wouldn't conclude that the probability that a randomly selected couple fit the eyewitness description was 1 in 12 million. But we would still get a pretty low probability (maybe 1 in a million). That still still seems like good evidence on which to convict, no?

No. It really isn't. We haven't even yet talked about the main thing that's gone wrong in the analysis. The main thing that's gone wrong is that we've been answering the wrong question. If we think clearly about the right question, we reach a very different conclusion.

The jury has to decide whether or not to convict the Collins. They don't want to do so if the Collins are sufficiently likely to be innocent and they do want to do so if the Collins are sufficiently likely to be guilty. So the right question for the jury is how likely the Collins are to be innocent, given the evidence. The evidence, of course, is that the Collins match the eyewitness description. So the right statistic to answer the jury's substantive question is the probability that the Collins are innocent given that they match that description. Write this as $\text{Pr}(\text{innocent} \mid \text{match})$. This is called a *conditional probability*, since it is the

probability of one thing conditional on another. It is read in one of two ways. People either say, “the probability they are innocent, conditional on them matching the evidence” or they say “the probability they are innocent, given that they match the evidence.” Either one is fine. The probability they are guilty conditional on them matching the evidence is just $\Pr(\text{guilty} \mid \text{match}) = 1 - \Pr(\text{innocent} \mid \text{match})$.

The mathematical analysis we’ve discussed thus far has *not* told us this probability and, so, has not answered the right question. The analysis thus far tells us how likely it is that a randomly selected couple would match the eyewitness description. That is, it tells us $\Pr(\text{match} \mid \text{innocent})$, which is read “probability a couple would match conditional on them being innocent.” While this statistic may be useful for answering the jury’s question, it is not itself the answer. The jury wants to know $\Pr(\text{innocent} \mid \text{match})$. The prosecutor has told them $\Pr(\text{match} \mid \text{innocent})$. But the jury (and the press) failed to notice the difference because they weren’t thinking clearly.

Let’s see why this matters. In the analysis above, we agreed that $\Pr(\text{match} \mid \text{innocent})$ is approximately equal to 1 in 1,000,000. We need to figure out $\Pr(\text{innocent} \mid \text{match})$. Can we do it?

Table 15.1 will help. It categorizes couples in Los Angeles county according to two characteristics: whether they match the eyewitness description or not and whether they are guilty or not. We know that there is exactly one guilty couple and that couple matches the eyewitness description. So the guilty column is easy to fill in. The innocent column is a little trickier. We’ve agreed that the prosecutor’s analysis, along with a bit of conjecture, suggests that the probability an innocent couple matches the eyewitness description is about 1 in 1,000,000. If we approximate that there are roughly 2 million innocent couples in LA County, we’ll conclude that there are approximately 2 innocent couples who also match the description. The remaining 1,999,998 couples in LA County fall into the last cell: innocent and don’t match.

	innocent	guilty
don't match	1,999,998	0
match	2	1

Table 15.1: LA couples by innocence or guilt and whether or not they match the evidence.

So, how likely is a couple to be innocent, given that they match the description—i.e., what is $\Pr(\text{innocent} \mid \text{match})$? Well, there are 3 couples that match the description. Exactly one of them is guilty. So the true probability that a couple is innocent, given that they match the eyewitness description, is not anything like 1 in one million. It is 2 in 3! That means the probability the couple is guilty given that they match the eyewitness description is only 1 in 3. Based on the eyewitness evidence alone, the Collins were twice as likely to be innocent as they were to be guilty.

The discrepancy arises not because the mathematician, the prosecutor, and the press looked at incorrect quantitative information, but because they looked at quantitative information that answered the wrong question. As the mathematician and prosecutor said, it is very unlikely that a randomly selected innocent couple would match the description of the criminals. But that doesn't mean it is very unlikely that a couple that matches the description of the criminals is innocent. Only one innocent couple in a million matches the description. But two couples out of three who match the description are innocent. If the jury had been able to think more clearly about the quantitative information, we suspect the Collins couple would not have been convicted. Few jurors want to send people to jail just because there's a 1 in 3 chance that they committed a crime.

Bayes' rule

The analysis we've just done is an example of a general approach to figuring out what we should believe, given some evidence, known as *Bayes' Rule*. It is named after Thomas Bayes, an 18th century philosopher and statistician.

Bayes' rule tells us the correct formula for how likely a claim is to be true, given the available evidence. It goes like this. Suppose we want to know the probability that a claim C is true, given evidence E . That is, we want to know $\Pr(C \mid E)$. In our example, the claim was that the Collins were innocent and the evidence was that they matched the eyewitness decision. Bayes' rule says

$$\Pr(C \mid E) = \frac{\Pr(E \mid C) \Pr(C)}{\Pr(E)}.$$

Let's go back to the Collins case to unpack this a bit. We want to know the probability the Collins are innocent, conditional on them matching the eyewitness description. In this case, Bayes Rule says

$$\Pr(\text{Innocent} \mid \text{Match}) = \frac{\Pr(\text{Match} \mid \text{Innocent}) \Pr(\text{Innocent})}{\Pr(\text{Match})}.$$

We can use Table 15.1 to find the values to plug in to see how this works.

What is $\Pr(\text{Match} \mid \text{Innocent})$? It is the probability a couple matches, given that they are innocent. There are 2 million innocent couples. Two of them match. So $\Pr(\text{Match} \mid \text{Innocent}) = \frac{2}{2,000,000}$.

What is $\Pr(\text{Innocent})$? It is the overall probability that a random couple is innocent. There are 2,000,001 couples in LA county. Of them, 2,000,000 are innocent. So $\Pr(\text{Innocent}) = \frac{2,000,000}{2,000,001}$.

Finally, what is $\Pr(\text{Match})$? Again, there are 2,000,001 couples, of which 3 match the eyewitness description. So $\Pr(\text{Match}) = \frac{3}{2,000,001}$.

Putting these together we have:

$$\begin{aligned}\Pr(\text{Innocent} \mid \text{Match}) &= \frac{\Pr(\text{Match} \mid \text{Innocent}) \Pr(\text{Innocent})}{\Pr(\text{Match})} \\ &= \frac{\frac{2}{2,000,000} \cdot \frac{2,000,000}{2,000,001}}{\frac{3}{2,000,001}} \\ &= \frac{2}{3}.\end{aligned}$$

Notice, we were able to figure this out earlier without knowing Bayes' rule, just by looking at the table. So there isn't much need to memorize the formula. But it is important to know how to calculate beliefs from evidence and to make sure you are thinking clearly about what question you want to ask and answer. Because it is really easy to convince yourself that $\Pr(\text{Match} \mid \text{Innocent})$ is the same as $\Pr(\text{Innocent} \mid \text{Match})$. But, as we've now seen, they can be really different.

Information, beliefs, priors, and posteriors

Bayes' rule is useful anytime we receive new information and want to update our beliefs about how likely some claim is to be true. Before we get the new information, we have what we call a *prior belief* about the claim. That is, our belief about the probability that the claim is true, without knowing the new evidence. In the formula, this prior belief is represented by $\Pr(C)$ —the probability the claim is true, without reference to the evidence. After we incorporate the new information, Bayes' rule gives us what we call a *posterior belief*: $\Pr(C \mid E)$.

In *People v. Collins*, the prior belief is the baseline probability that the Collins couple was innocent, before hearing about the eyewitness testimony. At that point, there was no reason to suspect them more than any other couple living in LA, so the prior was very close

to 1—something like $\frac{2,000,000}{2,000,001}$, since all but one couple was innocent.

We learned that the Collins couple matches the description of the criminals. In fact, the chances that an innocent couple matches that description was only 1 in 1,000,000, which might make us think that they are almost certainly guilty. But Bayes' rule tells us to hold off before jumping to conclusions. On the one hand, the evidence seems pretty damning. It's extremely unlikely that an innocent couple would match the description. On the other hand, the prior pushes in the other direction. It's extremely unlikely that any given couple is guilty. To figure out how likely it is that the Collins are guilty, given both of these facts, we have to ask about the relative likelihood of each one. If we ignore either our prior belief or the new evidence, we arrive at the wrong conclusion. Incorporating both, we see that, while the Collins are way more likely to be guilty than a randomly selected couple, there's still a good chance that they are innocent.

One way of thinking about the problem with the prosecutor's argument was that he talked only about the new evidence, ignoring the prior. This is a common mistake that people make when they aren't thinking clearly about quantitative evidence.

Abe's celiac revisited

If you remember, way back in Chapter 1, we told you the story of Ethan's son, Abe, being incorrectly diagnosed with celiac disease. In case you don't remember, here are the highlights of the story.

As a little kid, Abe was small for his age, which is an indicator for celiac. His pediatricians administered two blood tests. One came back positive (evidence that he had the disease), the other negative (evidence that he did not have the disease). The doctors concluded that Abe probably had celiac, because the positive test was "over 80 percent accurate."

The test on which Abe came up negative (let's call this Test 1) for celiac disease had quite low false negative and false positive rates, about 5 percent each. We can write this in our new notation. The false negative rate is the probability you get a negative test result given that you have the disease, i.e., $\Pr(\text{Negative on Test 1} \mid \text{celiac}) = 0.05$. The false positive rate is the probability you get a positive test result given that don't have the disease, i.e., $\Pr(\text{Positive on Test 1} \mid \text{No celiac}) = 0.05$.

The test on which Abe came up positive (let's call this Test 2) for celiac disease had a false negative rate of about 20 percent, i.e., $\Pr(\text{Negative on Test 2} \mid \text{celiac}) = 0.2$. This, we suspect, is where the 80 percent accurate claim came from. That test has a false positive rate of 50 percent, ie., $\Pr(\text{Positive on Test 2} \mid \text{celiac}) = 0.5$.

Prior to the blood tests, a reasonable guess about the probability of Abe having celiac disease, given his small stature, was maybe 1 in 100. That is Ethan's prior: $\Pr(\text{Abe has celiac}) = 0.01$.

Let's ignore Test 1 for a second, and just apply Bayes' rule to Test 2. Imagine a group of 10,000 kids, all of whom were similarly small in stature. Our prior tells us that, of those 10,000 kids, about 100 (1%) will have celiac. Test 2's false negative rate tells us that, of those 100 kids with celiac, about 20 (20%) will nonetheless test negative, while 80 will test positive. And Test 2's false positive rate tells us that, of the 9,900 kids without celiac, about 4,950 (50%) will nonetheless test positive and 4,950 will test negative. Table 15.2 provides a summary.

	celiac	No celiac
Negative on Test 2	20	4,950
Positive on Test 2	80	4,950

Table 15.2: Outcomes of a celiac test on 10,000 kids.

So what is the probability Abe has celiac, given that he was small in stature and tested positive on Test 2? Well, a total of $4,950 + 80 = 5,030$ kids test positive. Of those, 80 have celiac. So the probability that one of these kids has celiac given a positive result on Test 2 is $80/5030$, or approximately 1.6 percent.

Notice, now that we know Bayes' rule, we could have done this without making the table:

$$\Pr(\text{celiac} \mid \text{Positive on Test 2}) = \frac{\Pr(\text{Positive on Test 2} \mid \text{celiac}) \Pr(\text{celiac})}{\Pr(\text{Positive on Test 2})}.$$

We know enough to calculate each of these quantities. $\Pr(\text{Positive on Test 2} \mid \text{celiac})$ is 1 minus the false negative rate, i.e., 0.8. $\Pr(\text{celiac})$ is our prior belief, i.e., 0.01.

Calculating $\Pr(\text{Positive on Test 2})$ is a bit more involved. Here's how you do it. There are two kinds of people who test positive. Kids with celiac who get a correct test result. And kids without celiac who get a false positive. One percent of kids have celiac and of these 80 percent get a positive test result. 99 percent of kids do not have celiac and of these 50 percent get a positive test result. So

$$\begin{aligned} \Pr(\text{Positive on Test 2}) &= \Pr(\text{celiac}) \Pr(\text{Positive on Test 2} \mid \text{celiac}) \\ &\quad + \Pr(\text{No celiac}) \Pr(\text{Positive on Test 2} \mid \text{No celiac}) \\ &= 0.01 \times 0.8 + 0.99 \times 0.5 \\ &= 0.503. \end{aligned}$$

Now we can calculate Ethan's posterior beliefs directly:

$$\begin{aligned}\Pr(\text{celiac} \mid \text{Positive on Test 2}) &= \frac{\Pr(\text{Positive on Test 2} \mid \text{celiac}) \Pr(\text{celiac})}{\Pr(\text{Positive on Test 2})} \\ &= \frac{0.8 \times 0.01}{0.503} \\ &\approx 0.016.\end{aligned}$$

Of course, Abe actually had two tests. What happens if we add in the fact that Abe tested negative on the more accurate Test 1? If we assume that false positives and false negatives on these two tests are independent, then we can just multiply to get the relevant quantities.

$$\Pr(\text{celiac} \mid \text{Neg on Test 1 \& Pos on Test 2}) = \frac{\Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{celiac}) \Pr(\text{celiac})}{\Pr(\text{Neg on Test 1 \& Pos on Test 2})}$$

What is the probability that a kid with celiac gets a negative result on Test 1 and a positive result on Test 2? Well, Test 1 returns a negative for a kid with celiac (i.e., a false negative) only 5 percent of the time. Test 2 returns a positive for a kid with celiac 80 percent of the time. So, if the false negatives and false positives are independent across the two tests, then

$$\begin{aligned}\Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{celiac}) &= 0.8 \times 0.05 \\ &= 0.04.\end{aligned}$$

The prior belief, $\Pr(\text{celiac})$, remains the same, 1 percent. And, again, there are two kinds of kids who might get a negative on Test 1 and a positive on Test 2. First, the kid might have celiac (that's true of 1 percent of these kids). That kid would then need to get a

false negative on Test 1 but a correct result on Test 2. As we've just seen, the probability of this is $0.8 \times 0.05 = 0.04$. Second, the kid might not have celiac (that's true of 99% of these kids). That kid would then need to get a correct result on Test 1 and a false positive on Test 2. This happens with probability $0.99 \times 0.5 = 0.495$. Now we can calculate the overall probability of these two test scores using the fact that 1 of these kids have celiac and our knowledge of the false positive and negative rates for each test.

$$\begin{aligned}\Pr(\text{Neg on Test 1 \& Pos on Test 2}) &= \Pr(\text{celiac}) \Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{celiac}) \\ &\quad + \Pr(\text{No celiac}) \Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{No celiac}) \\ &= 0.01 \times 0.04 + 0.99 \times 0.495 \\ &= 0.49045.\end{aligned}$$

Plugging all of this into Bayes' rule, we get:

$$\begin{aligned}\Pr(\text{celiac} \mid \text{Neg on Test 1 \& Pos on Test 2}) &= \frac{\Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{celiac}) \Pr(\text{celiac})}{\Pr(\text{Neg on Test 1 \& Pos on Test 2})} \\ &= \frac{0.05 \times 0.01}{0.49045} \\ &\approx 0.001\end{aligned}$$

The probability that Abe had celiac given the two tests results was 1 in 1000!

Now that you know Bayes' rule, you can see that the doctors were not thinking very clearly about what the evidence really meant. But there is actually another layer to this story. And, in fact, it is one that you've had the tools to think through since all the way back in Chapter 8, when we learned about reversion to the mean.

Remember what Abe's doctors recommended after the blood tests came in? Put him on a gluten free diet and see whether, over the course of the next several months, his weight percentile goes up. If so, the doctors argued, the evidence for the celiac diagnosis would be

even stronger. Before reading further, ask yourself why mean reversion is a problem for this suggestion.

Abe was very small for his age. The timing with which a perfectly healthy kid grows is a function of lots of things, many of them random genetic and environmental factors that we know nothing about. At any given point in time, if a kid is particularly small for their age, that kid is likely to have experienced a particularly small draw of these random factors. Over the next period of time, those random factors are expected to be average, so the kid is likely to catch up in size, just by mean reversion alone. If Ethan had taken the doctor's advice, it's likely Abe's weight percentile would have increased over the following months. And this would have been taken as even stronger evidence for the celiac diagnosis. But, in fact, all of this would have been a mistake due to unclear thinking at every step of the diagnostic process.

Finding terrorists in an airport

In the years following the terrorist attacks of September 11, 2001, the United States government poured resources into airport security. One of the major new programs was called Screening of Passengers by Observation Techniques (SPOT).

The idea of SPOT was to use behavioral cues to catch potential terrorists before they boarded a plane. Behavior Detection Officers watched people in the security line at airports, looking for indicators that a person was nervous or otherwise suspicious. Different kinds of suspicious behaviors were assigned different numbers of points. If a person exhibited a cluster of suspicious behaviors that rose above some point threshold, that person was targeted for additional questioning, searching, and screening.

By the year 2010, about 5 percent of the Transportation Security Administration's (TSA) annual budget, hundred of millions of dollars per year, went to fund the SPOT

program. Let's use Bayes' rule to see why this wasn't a very good plan.

The TSA needs to be able to answer questions like: "Given a set of behaviors and characteristics, how likely is it that a person in question is a terrorist?" In other words, the TSA is trying to form a posterior belief about the probability that a traveler is a terrorist, given some evidence gleaned by observing the traveler's behavior. To form such posterior beliefs correctly based on a program like SPOT, the TSA needs to know at least three pieces of information:

1. How likely is a random traveler to be a terrorist?
2. How likely is a terrorist to appear suspicious to a Behavior Detection Officer?
3. How likely is a non-terrorist to appear suspicious to a Behavior Detection Officer?

Unfortunately, according to the General Accountability Office (GAO)—an independent, nonpartisan agency that works for Congress and is charged with investigating how the federal government spends taxpayer dollars—the TSA doesn't know the answers to any of these questions. No existing scientific research confirms, much less quantifies, the usefulness of behavioral observation for identifying terrorists. What we do know is that, even according to the TSA's own report intended to show the efficacy of the SPOT program, it seems that no terrorists have ever been caught by it. Indeed, the GAO reports that undocumented immigration status was by far the most common reason for detention of a person identified for additional screening by a Behavioral Detection Office.

So the government doesn't have the data that we need to calculate posterior beliefs based on the evidence collected by the SPOT program. But we can see that this program was never going to work even without this information. The way we do this is by asking how well the program would work in something like the best case scenario. That is, we make up some data, being extremely generous to the SPOT program in all of our assumptions,

and see whether the program would be a good idea under these assumptions. If the answer is no even under these generous assumptions, then we can be sure the answer is also no under more realistic assumptions.

According to the GAO, there are approximately 2 billion passenger-trips through American airports each year. For convenience, let's say there are two billion and one hundred. Presumably the vast majority of those people are innocent travelers. Very few travelers are trying to hijack planes or engage in other forms of terrorism. Let's be generous to the government and suppose that each year, 100 would-be terrorists are in American airports attempting to hijack airplanes. So that's our prior: $\Pr(\text{Terrorist}) = \frac{100}{2,000,000,100}$.

Second, we need to know how likely these terrorists are to exhibit the suspicious behaviors that the Behavior Detection Officers are looking for. Of course, we have no idea. But all the scientific evidence suggests that these kinds of behavioral cues are extremely noisy at best. Again, let's be generous and stack the deck in favor of SPOT. Suppose that 99 percent of all terrorists exhibit the behavior that the TSA is looking for, that is, $\Pr(\text{Suspicious} \mid \text{Terrorist}) = 0.99$. In reality, this number is much much lower.

Finally, we need to know how likely innocent travelers are to exhibit the suspicious behavior. As we've already said, these behaviors are very noisy, so at least some innocent people will exhibit them. But we want to be generous to SPOT. So suppose that only 1 percent of innocent people exhibit suspicious behavior, such that $\Pr(\text{Suspicious} \mid \text{Not Terrorist}) = 0.01$. Again, in reality, this number is much much higher. We are assuming SPOT is an incredibly accurate behavioral screening program.

How likely is a person who behaves suspiciously to be a terrorist? Even under these extremely generous assumptions, the answer is not very likely. Table 15.3 shows you the data you'd get based on our assumptions

Of the two billion and one hundred passenger-trips, 100 involve terrorists. 99 of them

	not terrorist	terrorist
not suspicious	1,980,000,000	1
suspicious	20,000,000	99

Table 15.3: How many terrorists and non-terrorists appear suspicious.

will exhibit suspicious behavior. The remaining 2 billion trips involve innocent travelers. Just one percent of them will exhibit suspicious behavior. But this one percent amounts to 20 million people! A total of 20,000,099 people act suspiciously. Of them, 99 are terrorists. So the probability that someone is a terrorist given that they acted suspiciously is $\frac{99}{20,000,099}$. That is, approximately 0.000005—about 1 in 200,000.

We could have similarly calculated this directly from Bayes' rule.

$$\begin{aligned}
 \Pr(\text{Terrorist} \mid \text{Suspicious}) &= \frac{\Pr(\text{Suspicious} \mid \text{Terrorist}) \Pr(\text{Terrorist})}{\Pr(\text{Suspicious})} \\
 &= \frac{\frac{99}{100} \cdot \frac{100}{2,000,000,100}}{\frac{20,000,099}{2,000,000,100}} \\
 &= \frac{99}{20,000,099}.
 \end{aligned}$$

Remember the numbers above come from assumptions that are extremely generous to the government. There is no way that terrorists actually exhibit the behavior the SPOT program looks for 99 percent of the time. And there is no way that innocent people actually exhibit the behavior the SPOT program looks for only 1 percent of the time. So the probability a suspicious person is a terrorist is actually way lower than 1 in 200,000. Indeed, if terrorists exhibit suspicious behavior 75 percent of the time and innocent people 10 percent of the time, the probability of being a terrorist given suspicious behavior becomes about 1 in 37 million:

$$\begin{aligned}
\Pr(\text{Terrorist} \mid \text{Suspicious}) &= \frac{\Pr(\text{Suspicious} \mid \text{Terrorist}) \Pr(\text{Terrorist})}{\Pr(\text{Suspicious})} \\
&= \frac{\frac{75}{100} \cdot \frac{100}{2,000,000,100}}{\frac{200,000,075}{2,000,000,100}} \\
&= \frac{75}{200,000,075} \\
&\approx \frac{1}{37,000,000}.
\end{aligned}$$

Remarkably, even with this number, we are being too generous. According to a study by the National Academy of Sciences, screeners looking for just one facial characteristic (rather than the many things SPOT screeners are looking for) in perfect conditions gets their assessment right only about 60 percent of the time. In more realistic conditions, they get their assessment right only about 30 percent of the time. With this level of accuracy and the tiny proportion of people who are terrorists, we think it is safe to say that the over one billion dollars allocated to the SPOT program was not money well spent.

This is easy to see when we ask the right questions.

Let us end this unpleasant tale with one more distressing tidbit that harkens back to the key lesson from Chapter 4, correlation requires variation. The Government Accountability Office is a watchdog organization that is supposed to make sure that government agencies spend money appropriately. After investigating a program it may also provide the relevant government agency with advice about how to improve. This is precisely what the GAO did after evaluating the SPOT program.

One of the areas that the GAO was concerned about was the lack of a scientific basis for the behavioral characteristics that TSA has its SPOT screeners looking for. According to GAO, the TSA has no idea whether terrorists are actually more likely to exhibit the behaviors they are looking for or not. (Of course, as we just saw, even if they are, this

program is a waste.) And so, here is what the GAO recommended to the TSA to improve accuracy:

Studying airport video recordings of the behaviors exhibited by persons waiting in line and moving through airport checkpoints and who were later charged with or pleaded guilty to terrorism-related offenses could provide insights about behaviors that may be common among terrorists.

Suppose you watched these videos and found that, for example, all the people who turned out to be terrorists wore sunglasses and looked agitated waiting in the security line. Do you want to start arresting everyone who meets that description? We hope not. As we've known since Chapter 4, correlation requires variation. If you want to get better at the (hopeless) task of identifying characteristics that predict a person is a terrorist, at the very least you must compare the characteristics of terrorists and non-terrorists. You can't just study the terrorists.

Bayes' rule and quantitative analysis

One particularly interesting application of Bayes' rule is thinking about how confident we should be about the truth of some scientific hypothesis, in light of the evidence presented in a scientific study. Of course, we already discussed one approach to this issue in Chapter 6. There we learned that the p-value tells us how likely a given estimate is to have occurred by chance alone. But if you think about it clearly, that really isn't an answer to the right question. In fact, when an analyst finds a low p-value and concludes that the finding must be true, they've made the same mistake as the mathematician and the prosecutor in *People v. Collins*. They've calculated the probability they would have found a relationship in their data, even if there is no real relationship in the world—i.e., $\Pr(\text{result} \mid \text{relationship not real})$. But what they really want to know is how likely it is that there is no real relationship,

given their result—i.e., $\Pr(\text{relationship not real} \mid \text{result})$. The probability there *is* a real relationship given the finding is just 1 minus this.

Let's use Bayes' rule to think about this a little more clearly. Suppose we collect some data, test for an relationship, and obtain a statistically significant result at the .05 level, i.e., $p < .05$. What's the probability that the estimated relationship reflects a real relationship in the world (as opposed to appearing in the data due to noise)? Bayes' rule tells us.

$$\Pr(\text{relationship real} \mid \text{result}) = \frac{\Pr(\text{result} \mid \text{relationship real}) \Pr(\text{relationship real})}{\Pr(\text{result})}.$$

And, of course, like before, we can break down $\Pr(\text{result})$ into two components. One way we might have found the result is that the relationship is real and the test correctly identified it. The probability of this is $\Pr(\text{relationship real}) \times \Pr(\text{result} \mid \text{relationship real})$. The other way we could have found the result is that the relationship is not real and the test spuriously identifies it due to noise. The probability of this is $\Pr(\text{relationship not real}) \times \Pr(\text{result} \mid \text{relationship not real})$. So we can write Bayes' rule as follows:

$$\Pr(\text{relationship real} \mid \text{result}) = \frac{\Pr(\text{result} \mid \text{relationship real}) \Pr(\text{relationship real})}{\Pr(\text{relationship real}) \Pr(\text{result} \mid \text{relationship real}) + \Pr(\text{relationship not real}) \Pr(\text{result} \mid \text{relationship not real})}.$$

We know the $\Pr(\text{result} \mid \text{relationship not real})$. This is just the *significance* level used in our hypothesis test. If we would declare a statistically significant result if $p < 0.05$, then $\Pr(\text{result} \mid \text{relationship not real}) = 0.05$.

The other numbers are more complicated. The quantity $\Pr(\text{relationship real})$ is our *prior* belief that a genuine relationship exists, before seeing any of our new evidence. The quantity $\Pr(\text{result} \mid \text{relationship real})$ —i.e., the probability you find a result in your data given that the relationship really exists in the world—is called the statistical *power* of the

test. Assuming the relationship is real, how likely is it that we would obtain a statistically significant estimate? There are ways of estimating the statistical power once we know more details about the data and test. For instance, one might conduct computer simulations to determine how likely they would be to statistically detect an effect of a certain magnitude.

Now we can rewrite the formula for Bayes' rule one more time in terms of these substantively interpretable quantities:

$$\begin{aligned} \Pr(\text{relationship real} \mid \text{result}) &= \frac{\Pr(\text{result} \mid \text{relationship real}) \Pr(\text{relationship real})}{\Pr(\text{relationship real}) \Pr(\text{result} \mid \text{relationship real}) + \Pr(\text{relationship not real}) \Pr(\text{result} \mid \text{relationship not real})} \\ &= \frac{\text{Power} \times \text{Prior}}{\text{Power} \times \text{Prior} + \text{Significance} \times (1 - \text{Prior})}. \end{aligned}$$

Suppose we have a hunch about some causal effect in the world. It's a bit of a long shot. We think there's a 5 percent chance that this effect exists (prior = .05). So we run a randomized experiment. We want to be confident in the answer, so we get a big sample size, such that the statistical power of our test will be .8 (we'll have an 80 percent chance of detecting an effect if one really exists). And following convention, we use a 0.05 threshold for statistical significance. Now, we can ask, conditional on obtaining a statistically significant result, what should our posterior beliefs be about the probability that the effect is real?

Plugging these numbers into the above equation, we get

$$\begin{aligned} \Pr(\text{effect real} \mid \text{result}) &= \frac{.8 \times .05}{.8 \times .05 + .05 \times .95} \\ &\approx .46. \end{aligned}$$

What happened? Even conditional on getting a result that is statistically significant at the 95 percent level, there's still only a 46 percent chance that the effect we believe we are

estimating exists at all! That may not sound right, but it is. And the logic is the same as that underlying the conclusion that the Collins were more likely to be innocent than guilty even though the probability a random couple matched the description was only 1 in a million. The p-value, just like that 1 in a million, is just one of the numbers we need in order to form our posterior beliefs. If the power is low or our prior beliefs are low, our posterior beliefs are likely to be low as well.

This kind of thinking also helps us to better understand the replication crisis in so many scientific disciplines that we described back in Chapters 7 and 8. Remember the ESP study? What was your prior belief about humans having ESP before you saw the results from that study? Probably pretty low, right? So your correct posterior belief that the effect is real, even given the statistically significant evidence, isn't that high. Figure 15.2 gives you a sense of this. The vertical axis is the posterior probability that an observed relationship is real. The horizontal axis is the prior probability it is real. The curve plots the correct posterior belief as a function of your prior belief, given that a study with statistical power of 0.8 and a significance threshold of 0.05 generated statistically significant evidence of the relationship.

Obviously, our prior beliefs are hugely important for our posterior beliefs. Indeed, if you have really low priors about ESP, like we do, then it might well not even make sense to study ESP, because the results of the study will have virtually no effect on your beliefs.

Figure 15.3 shows how the change in beliefs in response to new evidence relates to the prior. That is, it plots your posterior belief that a real relationship exists minus your prior belief that a real relationship exists, for different values of the prior, given that you saw statistically significant evidence in favor of the relationship. As you can see, if your prior is already very close to 0 or 1, it is very hard to move your beliefs. The effect of new evidence is largest for moderately surprising results (i.e., results where your prior was around .2).

Figure 15.3 also illustrates that two people can (and should) react quite differently to

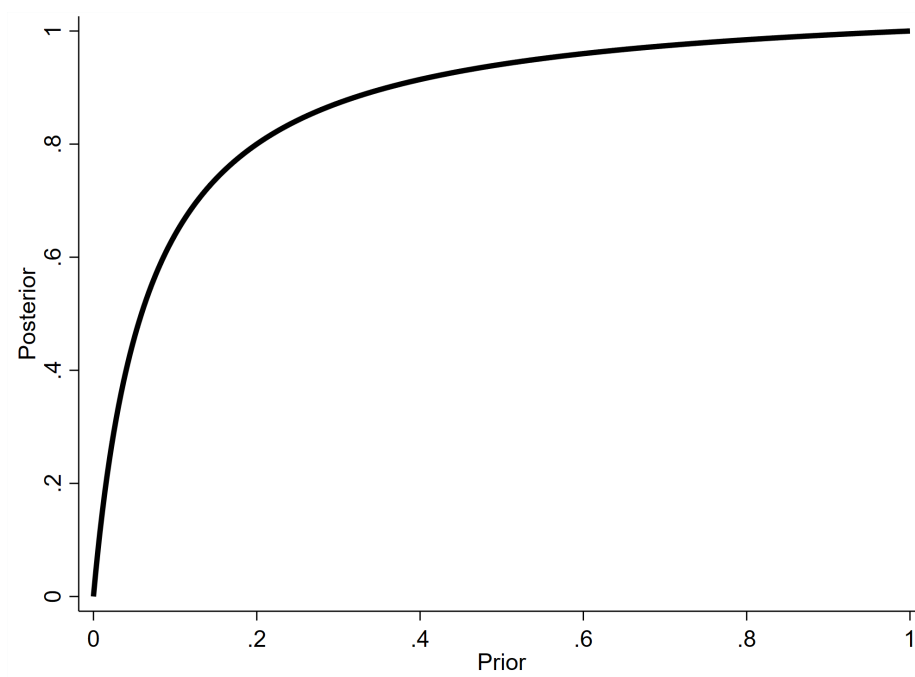


Figure 15.2: Posterior belief that effect is real given statistically significant evidence, as a function of prior belief.

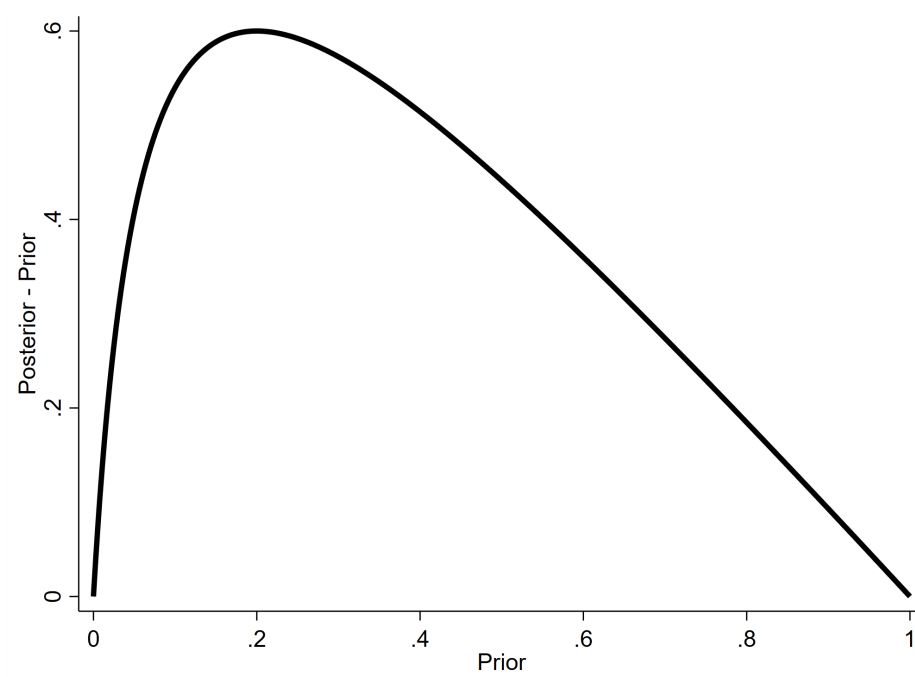


Figure 15.3: How much posterior beliefs change in response to new evidence as a function of the prior.

the same piece of information if they have different prior beliefs. Some people might see a piece of evidence about ESP, the consequences of global warming, or Russian interference in American elections and shift their beliefs dramatically, while others see the same piece of evidence and barely shifts their beliefs at all. When we experience this in our day-to-day lives, we often conclude that people who reacted differently than we did are unreasonable or irrational. But Bayes' rule tells us that it is perfectly understandable that different people react differently to the same information if at the outset they didn't agree about their prior beliefs at.

Some of this discussion might make you uncomfortable. As data analysts, aren't we supposed to let the data speak for itself without imposing our own prejudices? And where do these priors come from, if not from data? These are tough questions. But there's no way around them. If you want to say something about the probability there is a genuine relationship in the world, given some piece of evidence, you need to have prior beliefs about the likelihood of that relationship. You can't just ignore your priors. Because, as we've seen, $\Pr(\text{result} \mid \text{relationship not real})$ and $\Pr(\text{relationship not real} \mid \text{result})$ can be very different.

Here's another wrinkle. Most of the time, we're not really interested in the probability that some phenomenon exists or doesn't exist (though we probably are in the ESP example). Typically we want to know how substantively important an effect or relationship is, not just that it exists. That is, instead of just wanting to know if there is a real effect of, say, campaign strategy on vote share, we want to know how big the effect of campaign strategy on vote share is. How many supporters can a campaign turn out by launching a door-to-door canvassing campaign? Will turnout increase by 0.1, 1, or 10 percentage points? We can also incorporate Bayesian reasoning in such situations, but it's complicated. When thinking about the magnitude of a relationship, your prior belief is not just a single number as it was when thinking about the probability that a relationship exists. Instead, it is a belief about

how likely each possible relationship size is. And when you update your beliefs, you have to update your beliefs about each of these probabilities. Some analysts do this formally, specifying the whole prior distribution of beliefs about all the possible magnitudes and then doing complicated computations to estimate their posteriors. An alternative approach is to continue using conventional statistics like those described in Chapter 6 (these are often called *frequentist* statistics, in contrast to *Bayesian* statistics), while also being careful when interpreting the results.

Expected Costs and Benefits

Your beliefs about effects are, of course, only one input to a decision. Even once you have made sure that things are in the right scale, you are answering the right question, and you have formed correct posterior beliefs based on the evidence and your prior beliefs, quantitative information still doesn't speak for itself. To use information and evidence to improve decision making, you have to combine your evidence-based beliefs with your values and desires to figure out how to act.

There is a sense in which this is obvious. Suppose a really well designed series of studies convinces you that a certain kind of school intervention increases the likelihood that students will attend college by 30 percentage points. That's a big effect. But that alone doesn't tell you that the intervention is a good idea. To answer that question, at the very least, you have to know the value of college and how much the intervention costs.

It is easy, in the midst of forming beliefs based on sometimes complicated data analysis, to lose track of thinking about costs, benefits, values, and desires. A giant effect may seem compelling, just on its own. But it is important not to fall into this trap. Because, what seem like obvious implications of a piece of evidence, may turn out not to be so obvious. Let's see an example.

Amniocentesis and age

Pregnant women face a lot of medical decisions. How extensively to screen for potential genetic disorders is one of the most difficult. The most accurate way to screen is amniocentesis. And so, for a variety of reasons, women whose fetuses are at elevated risk for genetic disorders are encouraged to consider amnio. For instance, the Mayo Clinic advises women to consider amniocentesis if they had abnormal results from prenatal screening, had a previous pregnancy in which the fetus had a genetic disorder, have a family history of such disorder, or are over the age of 35.

In an interesting analysis, Eduardo Fajnzylber, Joseph Hotz, and Seth Sanders question this this last piece of advice—that amniocentesis makes more sense for older women than younger women. A woman deciding whether or not to get amniocentesis must weigh the benefits against the costs. So, if we want to assess whether older women should have amniocentesis and younger women shouldn't, we want to know how the benefits and costs of amniocentesis differ for older versus younger women. Let's break this down into two parts: (1) how do the expected benefits compare for older versus younger women and (2) how do the expected costs compare for older versus younger women.

Let's start by thinking about the benefits. A woman can only benefit from amniocentesis if the test results matter for her behavior in some way. So, let's imagine a woman who would change her behavior in response to a test result indicating a serious disorder, while leaving her behavior as is in response to a test result indicating no serious disorder. Of course, the test result could be a false positive or false negative. But let's assume that, in expectation, the behavioral change is beneficial. So the expected benefit of getting amniocentesis is the probability of a result indicating a serious disorder times the expected benefit of the behavioral change following such a test result.

On this logic, older women have a significantly larger expected benefit from getting

screened. The reason: older women are much more likely to have a fetus with a genetic disorder. So older women are more likely to be in a situation where they get a test result that leads them to alter their behavior in a beneficial way. Younger women who get screened very rarely get a result indicating a serious disorder (since their fetuses rarely have genetic disorders), so the expected benefit of getting screened is very small.

Now let's think about the costs. Amniocentesis carries real risks. Somewhere between half a percent to one percent of pregnant women who undergo amniocentesis miscarry as a result of the procedure. So, the expected costs (beyond any monetary costs) of amniocentesis are determined by at least two factors: (1) the chance of miscarriage as a result of the procedure and (2) how costly a miscarriage would be to the mother.

The standard advice, for older women to get screened but younger women not to, is based on the following logic: Amniocentesis has the same risk of causing a miscarriage regardless of the mother's age. So the expected costs of screening are the same for all women. And, as we've just seen, the expected benefits are larger for older women. So amniocentesis is worth the risk for older women but not for younger women.

Compelling as this logic may sound, it isn't quite right. The answer we've been given on the expected benefits side makes sense—the expected benefits are larger for older women because they are more likely to find out information that leads them to beneficially change their behavior from what they would have done absent the test. But the answer we've been given on the expected cost side is, “the probability of miscarriage is the same regardless of age.” This answer, based on quantitative information alone, would be the end of the story if the likelihood of miscarriage was the only factor determining the costs of amniocentesis. But it isn't. There is a second factor: how costly the miscarriage is for the mother. With respect to this factor, women of different ages may differ and, so, just knowing the probabilities of miscarriage alone does not suffice.

There are many costs to having a miscarriage—emotional loss, physical risk, not having

a desired child. Focus on this last one. Suppose a woman wants to have a child. The cost of a miscarriage, then, depends at least in part on how easily she can get pregnant again and have that additional child in the future. For a young woman, the risk of not being able to have another child is low. She has lots of time. But for a somewhat older woman the risk is higher. Given this, it is not right to say that the expected costs of amniocentesis are the same for younger and older women. The probability of the procedure inducing a miscarriage may be the same for both groups. But the costs of a miscarriage, should it occur, are higher on average for older women than they are for younger women.

So where does that leave us? The standard advice appears to be based on the answer to the wrong question. A more thoroughgoing analysis shows that, while the expected benefits of amniocentesis are indeed larger for older women, so too are the expected costs. So it is not obvious that older women should be more inclined to undergo screening through amniocentesis than younger women. To assess that, you'd need to know a lot more about how much an individual woman wants a child, how likely it is that she could become pregnant again in the event of a miscarriage, and how much she believes she would benefit from knowing about potential genetic disorders in advance.

For what it is worth, there has been some interesting research on this question. Here are a few important facts for thinking about these trade-offs. First, even for women at the very upper age limits of fertility, the vast majority of fetuses do not have genetic disorders. Thus, the expected benefits of amniocentesis, even for these women, may be small. Second, as an empirical matter, women in their mid- to late-thirties are more likely to choose amniocentesis than younger women. However, the rate of amniocentesis drops precipitously for women in their forties. Even if the medical profession hasn't figured out that it is giving advice based on an answer to the wrong question, women in their forties, thinking through a very important personal decision, seem to have worked it out on their own. This is especially true of educated women.

Given all this, it seems possible that the medical profession really is giving the wrong advice to women of all ages. Younger women, who have a long time horizon, might actually benefit from undergoing amniocentesis more often. And older women, who have a shorter time horizon, might benefit from undergoing amniocentesis less often.

This example illustrates that statistics alone are never enough to make good decisions. They can provide useful information and enable better decisions. But without thinking clearly about the particular costs, benefits, and preferences of each pregnant woman, we cannot say which ones should choose amniocentesis and which ones should not. Doctors can responsibly convey facts like the probability of a miscarriage induced by amniocentesis, the probability that the amniocentesis will detect different kinds of disorders, or the probability that the patient will be able to become pregnant again. But they cannot know how each patient should weigh these costs and benefits. Different patients will feel differently about the costs of a miscarriage and the benefits of knowing about a genetic disorder. So they might use the same statistical information to come to different decisions.

Key Words

- **Percentage point change:** The simple numerical difference between two percentages
- **Percent change:** The ratio of the percentage point change to the initial value.
- **Conditional probability:** The probability of an event happen conditional on some other information. We write the probability of the claim C conditional on the evidence E as $\Pr(C \mid E)$.
- **Prior belief:** Your belief about some thing before learning new evidence.
- **Posterior belief:** Your belief about some thing after incorporating new evidence.

- **Bayes' Rule:** A formula for calculating your posterior belief conditional on new evidence and your prior belief. In particular: $\Pr(C | E) = \frac{\Pr(E|C)\Pr(C)}{\Pr(E)}$.
- **Statistical Power:** The probability of finding a result in the data given that the relationship really exists in the world.

Exercises

1. A newspaper reports "Economic growth was 20 percent higher in Country A than in Country B last year."

The typical way that economists measure economic growth is the percent change in GDP from one year to the next. So we'd say economic growth was 3 percent in a particular country and year if the GDP was 3 percent higher at the end of the year than at the beginning.

- (a) Suppose GDP growth in Country B was 10 percent. What was GDP growth in Country A?
- (b) Suppose GDP growth in country B was 0.1 percent. What was GDP growth in Country B.
- (c) What's an alternative way to write the headline so that you don't misleadingly mask the difference between the scenarios described by (a) and (b)?

Now consider two other countries C and D. Suppose that growth in Country C is 1 percent while growth in Country D is 0.1 percent.

- (d) What is the percent difference in growth rate. What is the percentage point difference.
- (e) Write two headlines, each including a true statistical fact about the two countries.

One should make the difference in their economic growth sound like a really big deal. The other should not.

- (f) Now suppose that upon a statistical review, growth in Country D turns out to be just 0.001 instead of 0.1 percent. What now is the percentage difference in growth rates between the two countries? What is the percentage point difference? Which of these two statistics more accurately reflects the shift from 0.1 percent to 0.001 percent? Why?

2. During the coronavirus pandemic, governments and private organizations around the world rushed to create diagnostic tests. Those tests varied in their accuracy. Let's think about one of those tests, which was reported to have a one percent false positive rate and a ten percent false negative rate.

We don't know the underlying rate of coronavirus in the asymptomatic population. Suppose the probability an asymptomatic person has coronavirus is some number p —i.e., the prior probability of being sick is $\Pr(\text{sick}) = p$.

- (a) Using the information above about the false negative rate, what is the probability a person gets a positive result given that they really do have coronavirus (written, $\Pr(+ \mid \text{sick})$)? (Hint: You don't need Bayes' Rule to answer this question.)
- (b) There are two ways to get a positive test result. A person can have coronavirus and get a correct test result. Or a person can not have coronavirus and get a false positive. Calculate the overall probability that an asymptomatic person gets a positive test as a function of p :

$$\Pr(+) = \Pr(\text{sick}) \times \Pr(+ \mid \text{sick}) + \Pr(\text{not sick}) \times \Pr(+ \mid \text{not sick}).$$

- (c) Now use Bayes' Rule to calculate the probability that an asymptomatic person who tests positive—written $\Pr(\text{sick} \mid +)$ —in fact has coronavirus, as a function

of p ?

- (d) We, of course, don't actually know p . Let's think about different scenarios.
- i. Calculate $\Pr(\text{sick} \mid +)$ if $p = .005$, i.e., if half a percent of the asymptomatic population has coronavirus.
 - ii. Calculate $\Pr(\text{sick} \mid +)$ if $p = .01$, i.e., if one percent of the asymptomatic population has coronavirus.
 - iii. Calculate $\Pr(\text{sick} \mid +)$ if $p = .05$, i.e., if five percent of the asymptomatic population has coronavirus.
 - iv. Draw a figure with p on the horizontal axes (going from 0 to 1) that graphs $\Pr(\text{sick} \mid +)$.

3. Discrimination against certain groups in the job market is a major societal and policy concern. Many studies seek to bring quantitative evidence to bear on the extent of such discrimination.

Let's think through a very simple example. Let's imagine a society with two equally-sized and equally-qualified groups: the privileged and the unprivileged.

Using the conditional probability notation we developed earlier we will express the probability that a person gets a job given their group membership as, $\Pr(\text{hired} \mid \text{group})$. Similarly, we will express the probability that a person is a member of a given racial group, given that the person got a given job $\Pr(\text{group} \mid \text{hired})$.

- (a) Suppose you wanted to know whether, if they both apply for the same job, a member of the privileged group is more likely to be hired than a member of the unprivileged group. So you want to know whether, among those who apply for the job, if the following is true:

$$\Pr(\text{hired} \mid \text{privileged \& applied}) > \Pr(\text{hired} \mid \text{unprivileged \& applied})?$$

- i. Use Bayes' Rule to rewrite $\Pr(\text{hired} \mid \text{privileged} \ \& \ \text{applied})$ as a function of three terms, $\Pr(\text{privileged} \mid \text{hired} \ \& \ \text{applied})$, $\Pr(\text{hired} \mid \text{applied})$, and $\Pr(\text{privileged} \mid \text{applied})$.
 - ii. Use Bayes' Rule to rewrite $\Pr(\text{hired} \mid \text{unprivileged} \ \& \ \text{applied})$ as a function of three terms, $\Pr(\text{unprivileged} \mid \text{hired} \ \& \ \text{applied})$, $\Pr(\text{hired} \mid \text{applied})$, and $\Pr(\text{unprivileged} \mid \text{applied})$.
- (b) Suppose a study shows that people in a given job are equally likely to be privileged and unprivileged. Express that using our notation. Which two terms from your answer to (a) does that mean you know.
 - (c) Is the information in (b) sufficient to determine whether if they both apply for the same job, a member of the privileged group is more likely to be hired than a member of the unprivileged group? Using your answer to part (a), what additional piece of information would you need to know to provide an answer?
 - (d) Suppose you learned that the same number of members of the two groups applied for the job, now would you know the answer?

Readings

The study on the reporting of fuel efficiency statistics is

Richard P. Larrick and Jack B. Soll. 2008. The MPG Illusion. *Science* 320:1593-1594.

The Wall Street Journal article on a cholesterol drug is

Ron Winslow. "Cholesterol Drug Cuts Heart Risk in Healthy Patients." *Wall Street Journal*, Nov. 10, 2008. Available at <https://www.wsj.com/articles/SB122623863454811545>

The story of the statistical errors made in the trial of the Collins is relayed in

Jonathan J. Koehler. 1995. "One in Millions, Billions, and Trillions: Lessons from People v. Collins (1968) for People v. Simpson (1995)" *Journal of Legal Education* 47(2): 214–223.

For more on the SPOT program have a look at two reports by the General accountability office

The 2010 report available at <https://www.gao.gov/assets/310/304510.pdf>

The 2013 report available at <https://www.gao.gov/assets/660/658923.pdf>

The analysis of amniocentesis can be found in:

Eduardo Fajnzylber, V. Joseph Hotz, and Seth G.Sanders. 2010. "An Economic Model of Amniocentesis Choice" *Advances in Life Course Research* 15(1):11–26.

Chapter 16

Measure Your Mission

What you'll learn

- It is important that you measure outcomes and treatments that correspond to your mission.
- If you measure the outcome in an incomplete way, apparent improvements may be misleading.
- Data always comes from a particular context. When applying the lessons drawn from data to a new context, it is important to think clearly about whether the contexts are sufficiently similar that the lessons will continue to hold.
- One common way for this to go wrong is when the data are drawn from a sample selected because it has certain characteristics.
- Sometimes, there is a relationship in the world that would help you achieve your goal. But once you actually use that relationship to try to do so, the relationship itself disappears, so that it is no longer helpful.

When you use evidence to inform your decisions, you have some goal in mind. That goal is your mission. Why is it important to measure it?

Suppose you have evidence about a causal relationship; some action affects some outcome in some predictable way. If changing that outcome corresponds to achieving your goal—that is, if in measuring the outcome you measured your mission—then knowledge of that causal relationship is straightforwardly useful. But what if changing the outcome you measured doesn't quite correspond to achieving your goal, or what if it only corresponds to one part of your goal? Then it might not be so clear what action the evidence really suggests will further your mission.

The same goes for correlations. Suppose your mission is to predict some outcome, but you've measured a related, though different, outcome. Are you sure that the correlates of the outcome you measured will help you predict the outcome of interest?

In this chapter, we will explore a bunch of ways in which things can go wrong when we have good evidence about what might turn out to be the wrong outcome. Each of these examples will illustrate the reasons why it is important, as best as possible, to measure your mission when trying to use evidence to make better decisions.

D You Measure the Right Variable?

The most straightforward way that you might fail to measure your mission is by measuring an outcome or treatment that doesn't quite correspond to what you are really interested in. Here we consider three ways in which this commonly happens.

Partial measures

Often our mission is to change some outcome—say educational achievement, national security, or health—that is hard to measure in its entirety. For instance, we might not have

an encompassing measure of overall educational achievement, but perhaps we can measure whether students improve on standardized test scores in a few areas. Such partial measures can be helpful. But we have to be careful about interpretation because improving test scores is not our mission. Our mission is improving education.

In many settings, there are good reason to think that performance across dimensions of a problem are going to be negatively correlated. That is, if we get better at one part of a problem, we might expect to get worse at other parts of the problem. A simple reason for this is resource constraints. Suppose your overall mission is to make a local park more beautiful. You have a budget to support your mission. If you spend more resources on trash pickup, you have less money to spend on landscaping. So improving on one dimension, means getting worse on another. And if you just have a partial measure of your mission (say, the amount of trash on the ground), then as you spend more money on trash pickup, you might be tempted to conclude you are doing a better job achieving your mission. But, because things are getting worse on the landscaping dimension *as a result of getting better on the trash pickup dimension*, this is misleading.

There are, of course, reasons besides limited resources for the negative correlation across dimensions of a problem. Perhaps the most interesting is strategic adaptation of behavior. This too can make partial measurements problematic. Let's see how this plays out in some examples.

Starting in the mid-1960s, hijacking became a serious problem in American civil aviation. Over 80 airplanes were taken by hijackers in 1969 alone. The hijackers included Americans, Croatians, Cubans, Japanese, North Koreans, Palestinians, and many others. Their motivations ranged from simple ransom to nationalist, leftist, and other global political causes. In the early 1970s, in response to this growing assault on air safety, the United States increased airport security. Most importantly, metal detectors were installed in every major American airport in early 1973.

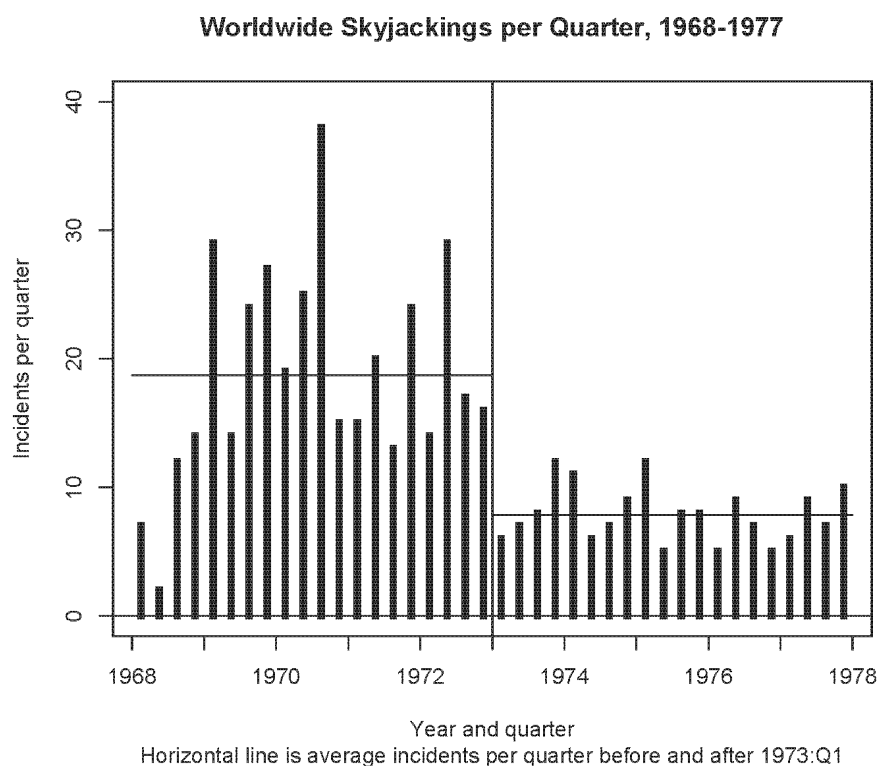


Figure 16.1: Quarterly hijackings 1968-1978. (Based on analysis in Enders and Sandler)

Imagine you were a government official tasked with evaluating the efficacy of these heightened security measures. A natural question you might ask is whether they resulted in a significant decrease in hijackings. Figure 16.1, showing hijackings per quarter from 1968-1978, suggests the answer is yes. Prior to 1973 (represented by the vertical line), there was an average of almost 20 hijackings per quarter. But after 1973, that number drops to fewer than 10 per quarter. (These averages are represented by the horizontal lines.)

Let's think about whether we've measured our mission. One possibility is that the mission is to increase security from *hijackings*. In that case, hijackings are the right outcome to study and this looks like a success. But another possibility is that the mission is to increase security from *terrorism*, not just hijackings. In that case, hijackings are only a

partial measure of the mission because there are lots of other kinds of terrorism.

Moreover, this is just the kind of setting where we might worry that improvements on one dimension of a problem (here, hijackings) is negatively correlated with improvements on other dimensions of the problem (here, other kinds of terrorist attacks). The reason is strategic adaptation. As it becomes harder to hijack an airplane, we might worry that terrorists substitute to attacking hotels, convention centers, shopping malls, and so on. If this is the case, the apparent reduction in hijackings might be misleading as a measure of how successful increased airport security was in terms of the overall counterterrorism mission.

And, indeed, this appears to be the case. As Figure 16.2 shows, after metal detectors were installed in airports, other kinds of terrorist hostage-takings became more frequent. Work by Walt Enders and Todd Sandler shows that during the course of the 1970s, there was an almost one-for-one relationship between the decline in hijackings and the rise in other hostage-taking type terrorist attacks. And so, if we have a more encompassing, rather than partial, measure of our mission, we reach very different conclusions.

Of course, this doesn't mean that the metal detector policy was a failure. Perhaps hijackings are worse, on average, than other kinds of hostage takings. So this might still be a counterterrorism win. But it is not nearly so dramatic a win as one might have thought looking only at the impact on hijackings, rather than a more encompassing measure of counterterrorism.

Intermediate outcomes

Often, measuring the outcome associated with your mission is difficult, takes a long time, or just doesn't yield enough data. A common solution is to measure intermediate outcomes, steps along the path of the mission, that we hope will be indicative of the longer-term

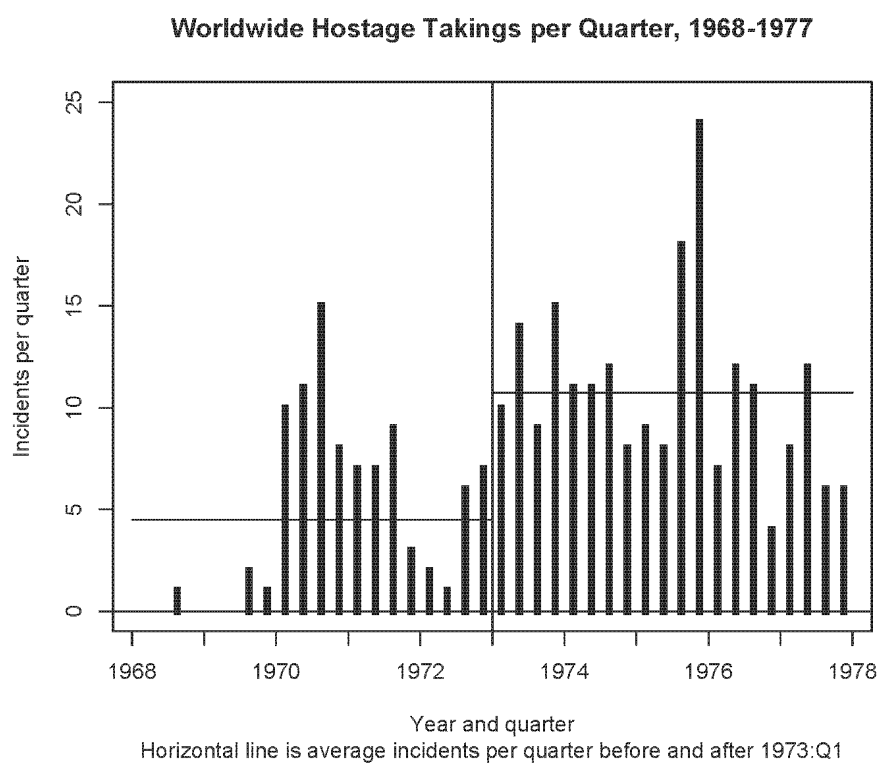


Figure 16.2: Quarterly hostage takings not on airplanes, 1968-1978. (Based on analysis in Enders and Sandler)

objective.

Suppose you're running a political campaign and you're trying to maximize the probability that your candidate wins. You want to test a few different ads to see which ones are the most effective. You could run the ads in different media markets and see where you do the best on election day. But that won't do much good. You need to know which ad to run before the election happens. So, instead, you have to measure some outcome that will give you some sense of which ad is best while the campaign is ongoing.

One natural candidate is opinion polls. Instead of running ads randomly and seeing what happens to vote totals, you could run ads randomly, conduct opinion polls, and see which ad appears to help you in the polls. There's not necessarily anything wrong with doing this. It's a good idea. But you have to keep in mind that you don't care about polls per se. You care about votes. So to the extent that changing poll numbers is a step along the path to changing votes, learning about this intermediate outcome is informative about your mission. But, for example, it's possible that your ad changes poll numbers by changing which people are willing to respond to the poll or whether people tell the truth to pollsters, without changing people's actual vote choices. In such a case, impacting the intermediate outcome might not matter at all for the final outcome you care about. So, whenever you use an intermediate outcome instead of a measure of your ultimate mission, you want to think about how sure you are that the intermediate outcome really is a step on the path to your actual goals.

This problem is particularly acute in lots of medical studies. Suppose the goal of a new drug is to reduce heart attacks. Unfortunately for research purposes (but fortunately for other reasons) heart attacks are rare. So relatively few people in any given sample will have heart attacks during the course of a drug trial. As such, it is very hard to learn directly about whether a drug reduces heart attacks even in a well-designed experiment that randomizes who gets the drug and who doesn't.

So what do medical researchers do? One alternative to waiting 20 years to see whether patients who were assigned the drug are less likely to have heart attacks is to study a surrogate or intermediate outcome, like blood pressure. Since blood pressure predicts heart attacks, the thinking goes, if a drug reduces blood pressure, it is likely to reduce heart attacks.

But we have to be careful. We learned in Part II that correlation need not imply causation. Playing basketball is correlated with height but experimentally increasing basketball playing does not increase height. Similarly, just because blood pressure and heart attacks are correlated doesn't mean that a drug that reduces blood pressure will reduce heart attacks. For that, you'd have to have compelling evidence that blood pressure has a causal effect on heart attacks.

Now, there are good reasons to believe that blood pressure really does have a causal effect on heart attacks. But for many surrogate outcomes used in medical studies, the causal linkage may be far less clear.

In a 1994 review of the evidence, Thomas Fleming illustrates the point in a discussion of research on cancer. Often, when studying cancer treatments, scientists cannot wait long enough to look at the affect on, say, mortality. So, instead, they study the effect of the treatment on an intermediate outcome. One popular such surrogate is tumor size.

For instance, Fleming describes a medical trial for a drug intended to treat prostate cancer. The researchers determined that if they examined mortality as their outcome of interest, they would need a sample size between 40,000 and 100,000 to detect a reasonably sized effect because death from prostate cancer is rare and slow. Since they could only recruit 18,000 men for their trial, they decided to use tumor size, as measured by a prostate biopsy, to assess the effectiveness of the drug instead.

One problem, as Fleming discusses, is that prostate tumor size is only a very weak proxy

for the actual mission, which is presumably something like not dying from cancer. Thirty percent of men over the age of fifty test positive for prostate tumors. But only 3 percent actually die from prostate cancer. Many prostate tumors grow very slowly. So other things, like heart attacks, get people first. The experiment showed that the drug being tested significantly reduced tumor size. But it is entirely possible that much of the reduction in tumor size was for the kinds of tumors that were never going to harm the patients in the first place. So we really don't know that progress on this intermediate outcome contributed much, if anything, to progress on the mission of avoiding death by prostate cancer.

Of course, we don't mean to suggest that studying intermediate outcomes is a bad idea. Indeed, it is often the best that can be done, given other constraints. But in interpreting the finding of a relationship between some action and an intermediate outcome, it is important that we think clearly about what we know about the relationship between the intermediate outcome and our actual mission.

Ill-defined missions

Often, your mission may be slightly tricky to pin down. In particular, there is sometimes more than one reasonable way to measure what may seem to be the same mission. But which choice you make can matter a lot. So it is important to think hard about what outcomes and treatments really define your mission.

Suppose you're a college student considering your educational and career choices with the goal of maximizing your future earnings. The first thing you might think to do is study the Forbes list of the richest people in the world and try to follow in their footsteps. One thing you might infer is that the way to maximize your earnings is to drop out of college and start a tech company. This was the strategy taken by Bill Gates and Mark Zuckerberg, two of the eight richest people in the world at the time of this writing. But you won't make that mistake because we learned in Chapter 4 that correlation requires variation. To know

if dropping out of college and starting a tech company is correlated with success, we can't just study the most successful people.

Suppose you pushed further and tried to get a sense of how many people in the underlying population dropped out of college and started their own tech company. You'd surely find that less than 0.01 percent of all people dropped out of college and started their own tech company, and yet 25 percent of the people with over 60 billion dollars did so. So there appears to be a strong correlation. People who drop out of college and start their own tech company are much more likely to have over 60 billion dollars than people who stay in college or never start a tech company.

Having identified a correlation, there are still some reasons you might not want to make a rash decision and drop out of college today. First, we might have just inadvertently engaged in something akin to p-hacking, a focus of Chapter 7. We studied a small population of extremely wealthy individuals, we looked for lots of commonalities, and we eventually found something that a few of them have in common. But that might just be a coincidence. Maybe the correlation we observe today won't hold in the future, in which case dropping out and starting a tech company might be a bad idea.

Yet another reason we wouldn't recommend dropping out and starting a tech company is that we're not comparing apples to apples. The kinds of people who drop out of college and start a tech company are likely different from those who don't, and we have little way of knowing whether they would have been equally successful had they not dropped out. That is, following the lessons of Chapter 9, this correlation is not an unbiased estimate of the causal relationship.

But even setting all of these reasons aside, there's a fundamental problem with this line of thinking that has to do with correctly measuring your mission. What outcome do you really care about? Is it your expected earnings or is it your probability of becoming a multi-billionaire? To the extent that dropping out of college and starting your own tech

company makes you more likely to be one of the richest people in the world, it probably also makes you more likely to be in serious debt. And for all we know, it might significantly reduce your expected earnings, even if it increases your chances of becoming very wealthy. Is that a gamble you're willing to make?

We're not here to tell you what your particular objectives should be. Some people may have a deep desire to become a billionaire making them willing to take significant risks to maximize their chances of achieving that goal. But we suspect most people are more averse to risk and would rather maximize their expected earnings or perhaps even minimize their chances of being in poverty. Whatever your particular objective is should inform the analyses you conduct. If your goal is to maximize expected earnings, it might be a huge mistake to examine the correlates of being on the Forbes list of wealthiest individuals. Instead, you'd want to collect data on earnings to see how various educational and career choices correspond with earnings, on average. We suspect you'd find that graduating from college and perhaps even going to professional school is a better predictor of earnings than dropping out and starting your own company.

This mistake of studying the wrong outcome can be made in the other direction as well. If you're managing a political campaign, you don't really care per se about your expected point margin or vote share. What you care about is winning, so you should choose strategies that maximize that objective. If your political candidate is polling badly and there's only a week left in the campaign, you might be willing to gamble on an otherwise ill advised strategy to give yourself a chance of winning. Maybe you decide to roll out a really aggressive new policy proposal which the voters probably won't like. So, in expectation, such a strategy reduces your vote total. But there is a small chance they will love your wild idea and you will win. If you don't really care about vote share (losing by 5 points or 10 points is still losing), but only care about winning, a strategy that hurts your expected vote share may be optimal.

And, of course, this measurement problem doesn't only apply to outcomes. It also applies to measuring treatments. This is perhaps most clear when the variables we measure are meant to represent abstract concepts. We have to think clearly about what, exactly, are we measuring when we rank some countries as more or less democratic than others or some classes more or less difficult than others. But this concern can also emerge even when we are measuring more concrete quantities in the world. Here's an example.

Many people are interested in the long-run effects of climate change on economic growth. Climate change, of course, happens over a long period of time and, thus, is hard to measure and study. But related phenomena, such as weather and temperature, vary frequently. So scholars sometimes use variation in the weather to try to learn about the effects of climate change.

For instance, Marshall Burke, Solomon Hsiang, and Edward Miguel estimate the effect of unexpected temperature fluctuations on GDP growth using a difference-in-differences design. That is, they compare the GDP within a country in years when it is exposed to warmer- versus cooler-than-average temperatures due to naturally occurring atmospheric variation. They find that economic productivity is maximized at an annual average temperature of 13 degrees Celcius and that it declines precipitously as the temperature rises. They conclude that “[i]f future adaptation mimics past adaptation, unmitigated warming is expected to reshape the global economy by reducing average global incomes roughly 23% by 2100.”

This is an important study and an important conclusion. But the caveat offered by the authors “[i]f future adaptation mimics past adaptation” is important and points to a critical measurement issue.

The authors are interested in the effects of *climate change*. But the treatment they measure is *temperature fluctuations*. Climate change happens slowly, giving people and society time to adapt. Temperature fluctuations happens quickly, making adaptation difficult.

Thus, in important ways, temperature fluctuations do not measure the right treatment. And, in particular, they don't measure the right treatment in ways that are relevant for the question of adaptation. In light of these measurement concerns, we should be careful in how seriously we take that 23 percent estimated effect.

To appreciate the distinction, consider the difference between the effects of a hot day on economic productivity versus the effects of a hot century. We live in Chicago, which can be a pretty cold place. If we were pleasantly surprised by an especially warm day, Anthony might be tempted to leave work early to play golf. But if climate change meant that every day was warmer, he wouldn't quit his job and play golf every day. The fact that unexpected hot days decrease productivity does not straightforwardly tell us the long-run effects of climate change because we haven't measured and studied the right thing.

Do You Have the Right Sample?

Studying the right outcome isn't all there is to measuring your mission. We almost always have to extrapolate when we apply evidence to decision making. That is, we take knowledge gleaned in some place and time, and try to apply that knowledge to understand what will happen in another place and time. Essentially we are making an analogy between the contexts in which the evidence was generated and the contexts in which we now wish to apply the lessons we learned from that evidence. So we always have to ask whether those contexts are in fact sufficiently similar that such an analogy is valid. Otherwise, we may take actions that are consistent with achieving our mission, but only in a very different context from the one in which we are acting.

External validity

The basic problem here is that relationships can differ from context-to-context. We've spent a lot of time so far in this book on what is sometimes called *internal validity*. Internal validity is about credibly estimating the estimand (e.g., is the estimator unbiased?). But even if you've done everything right with respect to internal validity, you still need to be able to think clearly about whether that relationship is likely to also exist in the context where you hope to apply it. Broadly, this is the problem of *external validity*. External validity is about whether there are good reasons to believe that a relationship estimated on data in one context will hold in some other context. An example will help to illustrate the point.

In the 1980s, the World Bank implemented the Tamil Nadu Integrated Nutrition Project (TINP) in a region of southern India where malnourishment was endemic. While the project included some resources for supplementary nutrition, the main focus was on helping mothers, the main household decision makers concerning food purchasing and preparation, make better use of the resources already at their disposal. The TINP is viewed as a major success by the World Bank. And, while there is some debate, it is widely credited with making a major difference in reducing malnourishment and malnutrition in Tamil Nadu.

This apparent success inspired the Bangladesh Integrated Nutrition Project (BINP) in the 1990s. By that time, Bangladesh, which borders India to the east, was among the most malnourished countries on earth. Evidence suggests that, in the early 1990s, almost two-thirds of Bangladeshi children under the age of 5 had growth stunting due to malnourishment.

Because the TINP had been rigorously evaluated and shown to have made a significant and meaningful dent in malnutrition, the BINP was modeled quite directly on the TINP. And so scholars and practitioners alike were surprised when the BINP's impact did not

live up to the promise of the TINP. Despite being designed to replicate perhaps the most successful malnutrition intervention in history, rigorous evaluation shows little to no impact of the BINP on malnutrition. What went wrong?

There are, of course, many possible answers. And it is virtually impossible to know for sure. But one important factor seems to have been a cultural difference between Tamil Nadu and Bangladesh. As we've mentioned, in Tamil Nadu, mothers are typically the chief decision maker regarding food purchasing and preparation. Thus, it made sense to target mothers for the TINP's nutritional education efforts.

This focus on mothers was ported directly from the TINP to the BINP. But in Bangladesh mothers are only one decision maker regarding food purchasing and preparation. And often they are not the most important. In many households either the father or the mother-in-law (i.e., the father's mother) has authority over either food purchasing or preparation. But these important decision makers were not targeted by the BINP. Thus, the BINP may have failed, at least in part, because a targeting decision that made perfect sense in one setting, was no longer so sensible in another.

This example is particularly interesting to us because it points to the potential for a complementarity between quantitative evidence and qualitative knowledge. Assessing the impact of the TINP required a quantitative approach. But attempting to extrapolate that knowledge to the Bangladeshi context went wrong because of a lack of knowledge about key cultural and institutional differences between Tamil Nadu to Bangladesh. A team that combined both people with expertise in quantitative assessment who could think clearly about the causal effect of the TINP and people with deep qualitative knowledge of the two relevant contexts might have resulted in a better outcome than either alone could hope to achieve.

Selected samples

A particularly common way for people to end up measuring their mission in the wrong context is by studying selected samples. A selected sample is a sample of observations that wasn't drawn at random from the population of interest, but rather was selected to be studied because it possessed some set of characteristics. The problem, of course, is that a *selected sample* may not be representative of the population as a whole. And relationships that hold in that selected sample may not hold in the broader population. If your mission is to predict or affect the behavior of the broader population, things can really go wrong if you rely on evidence from a selected sample.

College admissions

Here's an example that is near and dear to our hearts. Standardized test scores, for better or worse, have been an important part of the college admissions process for decades. However, our own university recently announced that it would no longer require applicants to submit such scores. One (among several) of the rationales for going "test optional" was evidence-based. They looked at the students who attended the university and found little correlation between test scores and performance. So, the argument goes, maybe test scores aren't very good predictors of college performance.

The mission of a college admissions office is multifaceted. But part of that mission is to identify the most academically talented students from the pool of applicants. To fulfill this mission, the admissions office would like to know whether some characteristic of *applicants* (here, their test scores) is correlated with academic performance in college. But that is not what the exercise described above does. Rather, the exercise described above asks whether some characteristic of *enrolled students* (namely, their test scores) is correlated with academic performance in college. But the answer to those two questions need not be

the same.

The set of enrolled students is a *selected sample* of the set of applicants. Students were admitted to college based on their test scores and other factors like writing ability, teacher recommendations, grades, community service, and overcoming adversity. The fact that test scores were used in admissions can lead to a fundamentally different correlation between test scores and academic performance in the selected sample of enrolled students and the broader set of applicants.

To see how, think about students with low test scores who were nonetheless admitted to the university. Those students must have had some other characteristics that led the admissions office to overlook their low scores. Maybe they wrote stellar essays, had particularly strong recommendations from teachers, or made great grades in high school. Similarly, students with particularly high test scores were likely to have been admitted even with somewhat weaker performance on these other dimensions. For this reason, in the set of admitted and enrolled students, we might expect a negative correlation between test scores and other markers of academic quality.

Now, it's quite plausible that test scores are a good predictor of college performance among all applicants, but that writing ability, teacher recommendations, and high school grades are also good predictors. Therefore, once we look at the selected sample of enrolled students, we'll find a weak correlation between test scores and performance. But that's because the only people with low scores who got in are people who are really strong on other dimensions. And so that weak (or non-existent) correlation in the selected sample of enrolled students does not mean that test scores are a bad predictor of academic performance among applicants.

This issue of studying selected samples is not isolated to major research universities. So let's talk through one more example in a different context: baseball.

Why can't major league pitchers hit?

Major League Baseball fans know that pitchers tend to be the worst hitters on their teams. In the National League, where pitchers are required to bat, managers will typically have their pitchers hit last in the lineup to minimize their times at bat. And if a pitcher is coming out of the game the manager will always replace them with a pinch hitter. In the 2017 major league baseball season, the batting average for the average pitcher was .125. The batting average for the average non-pitcher was .259. This is a massive difference. The American League has a designated hitter rule specifically so that pitchers don't have to bat.

So why are major league pitchers so bad at hitting? If you ask a baseball expert, they'd probably tell you that pitchers spend so much time practicing their pitching that they don't have time to practice their hitting. And there might even be something about great pitchers that makes them weaker hitters. Perhaps the kind of strength, flexibility, or body type that's good for pitching is bad for hitting.

These explanations sound pretty compelling, and they're probably right to some extent. But they also probably are not the whole story. One way to start to see this is to notice that this pattern does not hold for high school baseball.

We collected data on four Chicago-area high school baseball teams from the 2018 season and calculated the average batting average for non-pitchers and pitchers (defined as players who pitched more than 10 innings in the season). Unlike the pros, among these high school players, the pitchers actually have slightly higher batting averages than the non-pitchers: .322 vs. .317.

How can this be? Why is the correlation between pitching and hitting ability slightly positive for high school baseball players but negative for seasoned professionals? It's not as if pitching practice doesn't crowd out batting practice for young pitchers. And you'd think the arguments about physical specialization would apply in high school as well as in the

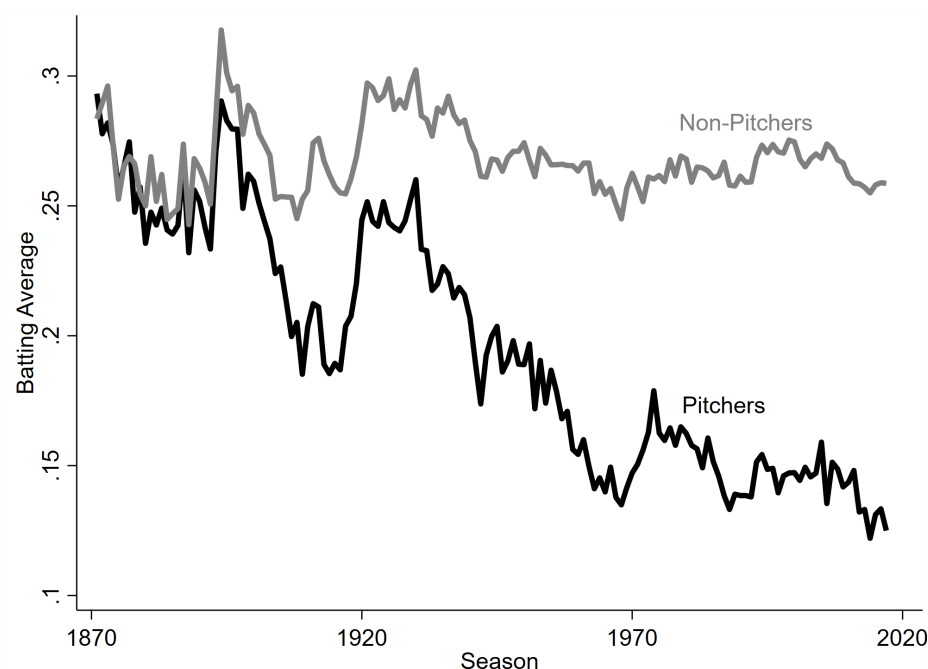


Figure 16.3: The negative correlation between pitching and hitting skill in the major leagues has gotten stronger over time.

major leagues. So how does the correlation seem to change so dramatically and even flip signs as players age?

Even at the professional level, we can see that there wasn't always a negative correlation between pitching and batting ability. Figure 16.3 shows the batting average for the average pitcher and the average non-pitcher in major league baseball from 1871 to 2017. In the 19th century, pitchers and other position players had comparable batting averages. But starting in the 20th century, the pitchers appear to get worse at hitting relative to other players, with the gap gradually increasing over time. And in the modern era, as we already discussed, non-pitchers get approximately twice as many base hits per at bat as do pitchers.

We suspect that the explanations for the changing correlation over time and the difference in correlation between high school and the professionals is one and the same. And it

has to do with selected samples.

Start by thinking about the correlation between pitching and batting ability in the entire population. Suppose we just randomly sampled individuals (say teenagers and older) from the whole world and asked them to play baseball so we could measure their pitching and batting abilities. What do you think we would find? We suspect that you'd find a pretty strong positive correlation. Some people are athletic and have experience playing baseball. They are likely to be good at hitting and pitching. Other people are uncoordinated and inexperienced. They are likely to be bad at hitting and pitching. So, in the population as a whole, you're likely to find exactly the opposite correlation from what you find among professionals.

To see why this should be, think about how a person becomes a major league baseball player. They almost surely start by playing high school ball. A high school coach is trying to assemble the best team possible. That involves choosing the players, from the set of players available, who offer the best combination of batting and pitching ability. To make your high school team, you have to be pretty good at some combination of hitting and pitching. But you don't have to be amazing. That means being a good hitter (even if you are a bad pitcher), a good pitcher (even if you are a bad pitcher), or an at least passable pitcher and hitter.

For most players, the path to the major leagues runs through the minor leagues. Minor league coaches are also trying to assemble the best team possible. And so, they too need the best combination of hitting and pitching available. So, to make a minor league team, you have to be really really good at some combination of hitting and pitching. That means being a great hitter (even if you are a bad pitcher), a great pitcher (even if you are a bad hitter), or an at least good pitcher and solid hitter.

Finally, to make the major leagues (at least the National League where pitchers hit), the test is even more stringent. You've got to be a truly amazing hitter (even if you are

a bad pitcher), a truly amazing pitcher (even if you are a bad hitter), or a pretty great pitcher who can also hit.

Below is a simple demonstration (with hypothetical data) of what these ever more stringent selection criteria do to the correlation between hitting and pitching ability in different samples. Suppose (for simplicity) that we can give every potential baseball player a score that separately summarizes their pitching and batting abilities, and teams want to recruit players with the highest possible sum of both pitching and batting ability. How high that sum needs to be is increasing as you move up the ranks of baseball.

In the top-left panel of Figure 16.4, we've drawn a scatter plot of some data with a strong positive correlation between pitching (horizontal axis) and batting ability (vertical axis). This is meant to represent the entire population. If we just let everyone on the team (as an entry-level team for kids might do), we'd see a pretty strong positive correlation between pitching ability and batting ability. This seems right to us. Our memory of youth sports is that the kids who were the best at one aspect of the game were the best at every aspect of the game.

The upper right panel of Figure 16.4 is meant to represent high school or the 19th century. In high school, the coach is only willing to accept players who are above average in the population, which translates into having a sum of batting and pitching ability above 0. Similarly, in the 19th century, baseball wasn't that popular, so professional coaches couldn't be so selective. You can make a team like this by being a good batter (say a 3) and a weak pitcher (say a -2); a good pitcher (say, a 3) and a weak hitter (say, a -2); or a slightly above-average pitcher and hitter (say .5 on both). But this level of selectivity eliminates people who are bad at both. And so, among the selected sample of high school or 19th century baseball players, there's little correlation between batting and pitching ability.

The panel on the bottom left might represent the minor leagues or the early twentieth century. Selectivity has increased. Players need at least a 3 to make the team. So bad

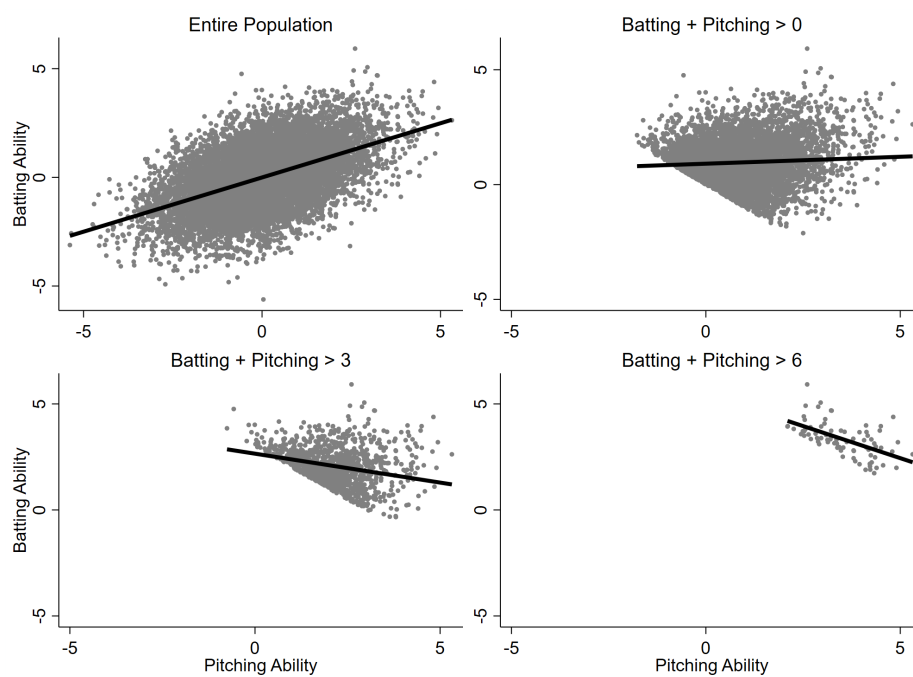


Figure 16.4: Increased selection can turn a positive correlation into a negative correlation.

pitchers must be genuinely good hitters, bad hitters must be genuinely good pitchers, and some players can make the team by being solid at both. By eliminating even more players who are only ok at both, this level of selectivity induces an even stronger positive correlation between batting and pitching ability.

Finally, the panel on the bottom right might represent the contemporary major leagues (at least the National league, where pitches have to bat). Here selectivity is very high, since only an elite few can make it. So bad pitchers must be truly great hitters, bad hitters must be truly great pitchers, and for players to make it with a combination of skills, they must be terrific at both. At the highest levels, then, we expect to see a strong negative correlation between pitching and batting ability, even without any difference in time spent practicing or body type.

Figure 16.4 illustrates a fairly general phenomenon. Correlations in selected samples are often quite different from those in the broader population. This is important because we often only have data on selected samples. But we may want to make predictions and inferences about the broader population.

If you're a baseball scout, you have rich data on major league players available to you. You might try to use the data to predict who will become a star player so you can search for those features among high school and college athletes. But if you look for correlations of great performance among major league players, you'll draw misleading inferences. For example, you might find that slow runners are really good power hitters. So should you go find the slowest runners and recruit them to play professional baseball? Surely not! The reason that slow runners are good power hitters in your selected sample is the same as the reason good pitchers are bad hitters. The only way a slow runner can make it to the major leagues is if they're a great power hitter.

Will Things Change Once You Act?

In 1696 King William III of England needed money. Kings, of course, always need money. But this need was particularly pressing. Up until the 1660s, Britain produced coins made of hammered silver. These coins had a serious problem—people shaved or clipped the valuable silver around the edges. As a result, the value of the coins in silver was less than their face value. This widespread practice of coin clipping threatened to undermine the credibility of English currency.

To address the problem, the Crown undertook the great recoinage of 1696, offering to buy back clipped coins in exchange for new, machined coins that could not be clipped. (As a fun aside, the new coins couldn't be clipped because they had milled edge, a feature which persists today, even though our coins are not made of precious metals. Milled edges were an innovation created by Isaac Newton in his role as warden of the Royal Mint at the time of the great recoinage.) But buying back clipped coins for proper coins was expensive. The Crown essentially had to make up the difference between the face value and the value of the silver. So the Crown needed to raise revenue. But how to do so?

The Crown wanted a progressive tax, i.e., one that placed a greater financial burden on the wealthy than on the poor. One natural way to achieve this is through a progressive income tax. But the English were opposed to income taxes because assessing income involved an invasion of personal privacy. So the Crown needed to find some progressive tax that was more politically palatable. The solution they landed on was the duty on lights and windows, better known as the window tax.

The window tax had the virtue that it could be assessed from outside of a home, thereby limiting any invasion of privacy. In the earliest version of the tax, the Crown established a two shilling base fee for all homes. In addition, homes that had between 10-20 windows paid an extra 4-6 shillings and homes that had over 20 windows paid an extra 8-10 shillings.

Windows in rooms used for work didn't count. The exact thresholds and fees changed over time (such taxes lasted for well over a century), but you get the basic idea.

The argument for the window tax was an obvious correlation (of course, they wouldn't have put it in those terms). On average, people whose houses had more windows were wealthier. Thus, by taxing windows, the Crown could raise revenues in ways that put more of the burden on the rich and less on the poor, which was the mission.

But the story doesn't end there. The English, like many others, don't really like to pay taxes. And so they responded in the obvious way. In the short-run, windows were boarded up or bricked over, to reduce the taxes owed. And, over the long run, architecture changed. Larger homes began to include fewer windows and more rooms that could be presented as work rooms. And so, with the passage of time, both the revenue from and progressivity of the window tax declined.

This story illustrates a third key issue for thinking about measuring your mission. Sometimes, there is a true relationship in the world which would help you achieve your goal. But once you actually use that relationship to try to do so, the relationship itself disappears or changes, so that it is no longer so helpful. In this case, the Crown's mission was to raise revenue in a progressive way. To do so, it needed to identify and tax wealthier people without invading their privacy. It noticed a correlation between windows and wealth, which seemed like just what it needed to achieve its mission. But, using that correlation in service of its mission led homeowners to change their behavior such that the correlation no longer held (or, at least, held much less strongly), undermining the mission. Thus, in considering some change in behavior or policy in response to some piece of evidence, one must always ask whether the relationship uncovered by the evidence will persist once you change your behavior or policy.

Let's look at a couple more contemporary examples of the same phenomenon.

The shift in baseball

We know we've already spent a good bit of time on baseball in this chapter. But, if you'll indulge us, we would like to do one more example. It is a pretty good one for illustrating the idea.

There was a time when defenders in baseball stood in their spot and waited to see if the ball would come their way. To be sure, fielders would adjust their position a bit depending on whether a left-handed or right-handed batter was up. But, for the most part, defensive strategy wasn't too complicated.

That time came to an end with the rise of big data in professional sports. Now teams have detailed "spray charts" for every batter. These charts provide data on how frequently each batter hits to various parts of the field, whether they hit the ball on the ground or in the air, the angle at which they make contact with the ball, and so on. Using this kind of data, teams can make well-informed forecasts about exactly where a given batter is likely to hit the ball. And armed with such information, teams have started adjusting their defensive set-ups aggressively, batter-by-batter.

The most famous version of this change in defensive strategy is called the *shift*. Examining spray charts, teams discovered that when batters (especially power hitters) hit the ball on the ground, it is almost never to the so-called "opposite field" (for right-handed batters, this is to their right and for left-handed batters, this is to their left). Rather, if they are going to hit the ball on the ground, they "pull" the ball (i.e., right-handed batters hit it to their left and left-handed batters hit it to their right). The shift is the obvious response to this correlation—when facing a right-handed batter, shift the infield way over to the batter's left and when facing a left-handed batter shift the infield way over the batter's right. The benefit of such a move is that it makes it much less likely that a ground ball that is pulled will sneak through a hole in the infield for a base hit. The cost is that it leaves a

big hole in the infield in the opposite field. But since batters find it very hard to hit ground balls to the opposite field, this cost is small.

A few teams started shifting pretty aggressively in the late 2000s. In 2010, the Tampa Bay Rays—led by manager Joe Maddon, an early advocate for evidence-based defense—accounted for 10% of all shifts, although they were just one of thirty teams. Maddon consulted spray chart and strategically placed his infielders in locations that were optimal for the particular pattern of ground balls associated with each batter. The Rays and other early adopters had a lot of success. That is, there was a positive correlation between using the shift and giving up fewer runs.

Observing this correlation, all teams started implementing the shift. In 2011, there were only about 2,000 shifts used in total across all major league baseball games. By 2014, that number had grown to 13,000. And in 2016, it reached over 28,000.

But something else happened too. At first, the correlation that inspired this surge in shifting held. Teams that shifted allowed fewer runs. But batters noticed that the shift was hurting them. And they developed a strategy to avoid hitting so many ground balls into the shift. Now, as we've said, for many major league hitters, they couldn't really do this effectively by hitting ground balls to the giant hole in the opposite field. So, instead, they started focusing on hitting the ball in the air. This had two advantages. First, by hitting the ball to the outfield, they avoided the shifted infield. Second, major league baseball players can hit the ball to the opposite field in the air. So, to some extent, they were able to exploit the hole the shift left in the infield, just not by hitting the ball on the ground.

As things stand today, major league teams still use the shift a lot. But, because hitters adapted, the correlation between shifting and runs allowed that drove teams to use the shift in the first place does not hold nearly as strongly as it used to. Setting defensive strategy in response to the correlation led to changes in behavior that undid that correlation. It is perhaps worth noting that Joe Maddon—the early innovator who, as manager of the Rays,

helped make the shift so popular—remains a believer in evidence-based defense. As of this writing, he is the manager of the Chicago Cubs, where he employs the shift less than any other manager in major league baseball.

The war on drugs

Before leaving the topic of whether things will change once you act, it is worth pausing to reflect on the overlap between this question and our earlier discussion of the problem of partial measures. That overlap comes from the fact that strategic adaptation can create both phenomena.

Remember what we are concerned about in the case of partial measures. Suppose you have only a partial measure of your mission (like hijackings). You take an action and things appear to improve on that measure. But there might have been strategic adaptation such that getting better on that one dimension of your mission implied getting worse on some other dimension. Hence, improvements on your partial measure may not mean improvements on your overall mission.

Strategic adaptation is again at the root when we worry about whether things will change once you act on some piece of evidence. There is some relationship in the world. You act on that relationship. People adapt to your actions. And, thus, the relationship disappears.

Many examples can fit into both categories, depending on how you think about them. Let us give you one last example to illustrate the point, this time about America's so-called war on drugs.

As we write, most of the illegal drugs in the United States enter the country through Mexico, a country which has been ravaged by a decade long drug war. But this was not always the case. In the 1970s and early 1980s, very few drugs reached the United

States through Mexico. The transshipment route of choice was from Colombia through the Caribbean and into Florida.

In 1980, the United States' government launched a major offensive against the Colombian drug cartels. The Drug Enforcement Administration, Coast Guard, and other agencies deployed thousands of personnel and considerable naval and air power to shut down the Caribbean transshipment route. By the mid-1980s, the Colombian drug cartels were actively pulling out of the Caribbean and the flow of drugs into Florida had plummeted.

But that is not the whole story. The reduction in drugs flowing through the Caribbean and into Florida in the 1980s does not reflect a reduction in drugs flowing to the United States during that period. Indeed, drugs continued to flow to the United States at increasing rates, as evidenced by the fact that the price of cocaine plummeted four-fold during the course of the 1980's (from a wholesale price of \$60,000 per kilogram in 1980 to \$15,000 in 1989) despite soaring demand.

So what did happen? The Colombian cartels abandoned the Caribbean and Florida in favor of Mexico. In 1989, one-third of all cocaine in the United States entered through Mexico. Just three years later, that number had increased to one-half. Today, ninety percent of cocaine sold in the United States is smuggled up from Mexico.

This adaptation by the drug organizations has had devastating effects on Mexico. Throughout the 1990s, the Mexican drug trafficking organizations became larger and more powerful. They shifted from being middlemen for the Colombians to having their own suppliers and distribution networks. The drug trade became larger and more important—by the mid-1990s the Mexican drug trade was worth roughly 20 billion dollars, dwarfing Mexico's largest legal commodity, oil, with a value of about \$7.5 billion. As this expansion occurred, Mexican drug organizations became more fragmented and more violent. In 2010, the Mexican drug war claimed over 1,000 lives per month. The Mexican government struggled to exert basic control over parts of the country.

One can fruitfully think about this story from both the *partial measures* and the *will things change* perspectives.

From the *partial measures* perspective, here's how you'd tell it. We had a mission of stopping drug flows into the United States. We noticed that almost all the drugs came up through the Caribbean. So we collected data—drugs flowing through the Caribbean—that was only a partial measure of our overall counter-narcotics mission. Then we took actions that made things improve according to our partial measure. But to conclude that the policy was a success would be a mistake. Because of strategic adaptation, getting better on our partial measure (drugs coming through the Caribbean) goes along with getting worse on other dimensions of the problem (drugs coming through Mexico). Here, the story illustrates the importance of not over-interpreting improvements on a partial measure of your mission.

From the *will things change* perspective, we tell the story slightly differently. There was a real correlation in the world—drugs were much more likely to enter the United States through the Caribbean than anywhere else. The government decided to act on the basis of this relationship by targeting its interdiction efforts in the Caribbean and Florida. Drug traffickers adapted their behavior in response to this action, moving transshipment to Mexico. And so, as a consequence of the government's own actions, the correlation that formed the basis of the government's actions ceased to exist.

Both of these perspective are right. Which is more useful depends on the particular question you are trying to answer and the context in which you are trying to answer it.

Key Words

- **Internal validity:** An estimate is internally valid if it is a credible estimate of the estimand (e.g., the estimator is unbiased).
- **External validity:** An estimator is externally valid if there is good reason to think

the relationship will hold in a context other than the one from which the data is drawn.

- **Selected sample:** A sample of data that wasn't drawn at random from the population of interest, but rather was selected to be studied because it possessed some particular set of characteristics.

Exercises

Over the past several decades, high stakes testing has become an increasingly important part of American education policy. The idea of high stakes testing is to create consequences for students, teachers, or schools associated with performance on standardized tests. The hope is that this will improve educational achievement by creating incentives for better performance.

1. Standardized tests scores are, at best, a partial measure of educational achievement. Given this, give an example of why some policy intervention that leads to an improvement in test scores might nonetheless not be leading to an overall improvement in educational outcomes.
2. In a required math sequence at the United States Air Force Academy, students take the same exams but are randomly assigned to different sections taught by different instructors. Scott Carrell and James West show that students assigned to an instructor with better teaching evaluations perform better on the course's exams. But they also show that being assigned to a popular instructor *decreases* students' scores in subsequent math classes. What might explain this puzzling pattern? How does it relate to the problem of failing to measure your mission?
3. The way high stakes testing is implemented in primary and secondary education is

typically based on thresholds. A student passes the test if they get a score above some minimal cutoff. And a school is deemed to be meeting standards if the number of students passing the test is above some other minimal threshold.

- (a) Think of students in three categories: those who will pass the test no matter what, those who will pass the test if and only if they get teacher attention, and those who will not pass the test no matter what. Which students does high stakes testing of this form incentivize teachers to focus on?
- (b) Derek Neal and Diane Whitmore Schanzenbach study the implementation of high stakes testing in the Chicago public schools that we discussed in Exercise 1 of Chapter 13.

But, unlike in that question, the average effect of high-stakes testing isn't quite what they wanted to know about. They wanted to know whether high-stakes testing affected different kids differently.

To get at this, Neal and Schanzenbach used a difference-in-difference-in-differences design. They start by using the third grade tests to divide students into 10 groups (deciles). They then perform the difference-in-differences analysis separately for each of these deciles. This allows them to learn about the difference in the difference-in-differences estimate of the causal effect of high-stakes testing across kids in the different deciles. Their findings are reflected in Figure 16.5.

Is this evidence consistent with your answer to part (a) of this question? Explain.

- (c) In light of this, is a simple difference-in-differences design that uses percentage of students passing the standardized tests a good way to evaluate whether high stakes testing achieves its mission? Why or why not?

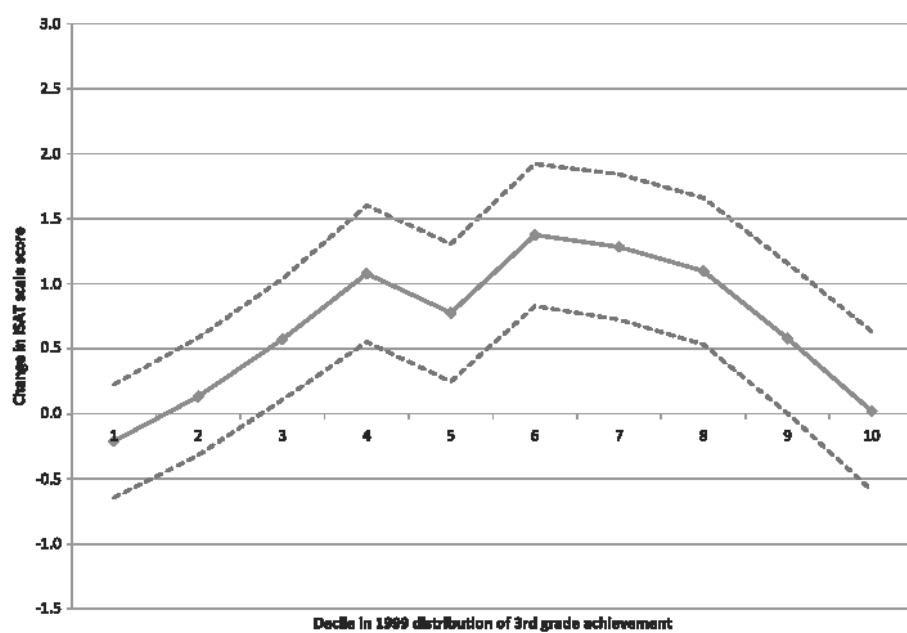


Figure 16.5: Evidence of effects of high stakes testing by decile. Source: Neal and Schanzenbach)

Readings

To read more about hijackings and metal detectors, see

Walter Enders and Todd Sandler. 1993. "The Effectiveness of Anti-Terrorism Policies: A Vector-Autoregression-Intervention Analysis." *American Political Science Review* 87(4):829-844.

The study of math exams and instructor quality in the Air Force Academy is

Scott E. Carrell and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3):409-432.

You can learn more about intermediate outcomes in medical research in

Thomas Fleming. 1994. "Surrogate Markets in Aids and Cancer Trials." *Statistics in Medicine* 13:1423-1435.

Thomas R. Fleming and David L. DeMets. 1996. "Surrogate End Points in Clinical Trials: Are We Being Misled?" *Annals of Internal Medicine* 125:605-613.

The study of temperature fluctuations and economic growth is

Marshall Burke, Solomon M. Hsiang, and Edward Miguel. 2015. "Global Non-Linear Effect of Temperature on Economic Production." *Nature* 527(7577): 235–239.

For a comparison of the Tamil Nadu and Bangladesh Integrated Nutrition Projects, see

Howard White and Edoardo Masset. 2007. "Assessing Interventions to Improve Child Nutrition: A TheoryBased Impact Evaluation of the Bangladesh Integrated Nutrition Project." *Journal of International Development* 19(5):627-652.

The historical Major League Baseball statistics come from the Baseball Databank at seanlahman.com. The high school baseball statistics are from GameChanger at gc.com.

If you are interested in the history of adaptation to the window tax in England, have a look at:

Andrew E. Glantz. 2008. "A Tax on Light and Air: Impact of the Window Duty on Tax Administration and Architecture, 1696-1851." *Penn History Review* 15(2).

Wallace E. Oates and Robert M. Schwab. 2015. "The Window Tax: A Case Study in Excess Burden." *Journal of Economic Perspectives* 29(1): 163-180.

On this history of the shift in baseball, you can read:

Travis Sawchik. "We've Reached Peak Shift." blogs.fangraphs.com/weve-reached-peak-shift/

The statistics on drug flows to the United States come from two reports

United Nations Office on Drugs and Crime. 2010. "The Globalization of Crime: A Transnational Organized Crime Threat Assessment." Chapter 4. http://www.unodc.org/documents/data-and-analysis/tocta/TOCTA_Report_2010_low_res.png

Office of National Drug Control Policy. October 2001. "The Price of Illicit Drugs: 1981 through the Second Quarter of 2000." https://www.whitehouse.gov/sites/default/files/ondcp/policy-and-research/bullet_5.png

The discussion of high-stakes testing and teaching to students on the bubble in the exercises is most directly based on:

Derek Neal and Diane Whitmore Schanzenbach. 2010. "Left behind by design: Proficiency counts and test-based accountability." *Review of Economics and Statistics* 92(2):263-283.

It also draws on:

Bengt Holmstrom and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*.

Chapter 17

The Perils and Promise of Quantitativication

TBA