All Else Equal in Theory and Data (Big or Small)[*]

Scott Ashworth[†] Christopher Berry[‡] Ethan Bueno de Mesquita[§]

August 1, 2014

The forms of explanation that dominate political science research in the formal theory and causal inference traditions are closely connected. Specifically, each makes essential use of different, but related, kinds of all-else-equal claims. The emergence of "big data" has already begun to alter the landscape of empirical social science by making new sources of information (and, thus, new phenomena) amenable to quantitative analysis. But neither the centrality of all-else-equal explanations, nor the challenges associated with providing them, will be altered in the least by big data.

1 All Else Equal (on Average)

The formal theory and causal inference research agendas in political science share some deep connections. Both approaches strive to learn about mechanisms that play a role in explaining social phenomena. And each relies on all-else-equal comparisons to explore such mechanisms. To better see the connections, we start by laying out our view of how each approach works on its own terms. We then return to explicate the bridge between them.

Formal, game theoretic models are simplifications, meant to represent some small piece of a larger phenomenon. The goal of such models is typically to offer qualitative insight into the mechanisms of some social process or pattern of behavior. This goal dictates simple models, for at least two reasons.

First, when we say that a model offers insight into mechanisms, we mean that it makes clear which assumptions generate some particular outcome in the model. If some of those assumptions and outcomes, and the relationships between them, represent analogous features

[*] We have benefited from conversations with Anthony Fowler, Pablo Montagnes, Rocio Titunik, Dustin Tingley, and Stephan Wolton.

[†] Harris School of Public Policy Studies, University of Chicago, email: sashwort@uchicago.edu

[‡] Harris School of Public Policy Studies, University of Chicago, email: crberry@uchicago.edu

[§] Harris School of Public Policy Studies, University of Chicago, email: bdm@uchicago.edu

of the empirical political environment, then the model is useful for explanation. Without knowing which features of a model drive a result, one cannot know whether the proposed explanation is robust, or whether it is the product of very special assumptions that are too distant from reality to constitute a satisfying explanation. This kind of clarity is typically only achievable in a simple model. Perhaps the most important reason is that it is easier to "see through" models without too many moving parts. But another reason is that abstraction and simplification also highlight robust categories of well-understood incentive problems—signaling, multi-task, externalities, bargaining, commitment problems, coordination problems, and so on—that describe a large number of social situations and support explanation by analogy.

Second, when applied game theorists talk about explanation, they typically want more than a model in which agents behave in a way consistent with behavior in the world. Rather, they seek an understanding of why the agents in their models behave as they do in terms of preferences, beliefs, information, and constraints. Thus, another important value of simple models: the hermeneutic goals of game theoretic analysis cannot be achieved with more complex models.

So the fundamental explanatory goals of formal theory require simple models. But this simplicity constrains the kind of explanatory statements one can make based on a formal model. Formal models in the social sciences are typically not attempts to write down a complete description of the mechanisms underlying some phenomenon. And since a theorist deliberately omits most of what is important about the world from any particular model, explanatory statements based on such a model are of necessity (implicitly or explicitly) all-else-equal claims. For instance, a theoretical model's comparative static prediction about the relationship between two variables explicitly holds everything else in the model constant. And it implicitly holds constant everything else that is omitted from the model but might alter the relationship in the world.

Let's turn now to the goals of empirical research in the causal inference tradition. Obviously, the goal of such work is to estimate causal effects. What are these? A general definition comes from the potential outcomes framework (Holland, 1986): the effect of x on y, for a particular unit, is the difference in the values of y for different assignments of x. Since this effect is defined at the level of an individual unit, it is an all-else-equal relationship.

A fundamental challenge for empirical researchers is that it is impossible in principle to observe the same unit under two different treatment statuses at the same time. Hence, in a deep sense, it is impossible to observe a causal effect directly. In practice, empirical researchers try to hold all else orthogonal, which is to say, all else equal on average.[1] The gold standard for doing so is experimental randomization. Often, practical or ethical considerations render this impossible, and empiricists interested in causal inference must find research designs that allow them to credibly claim to have held all else equal, on average, in observational data.

It's clear now why there is a tight connection between the formal theory and causal inference research traditions. Formal models, of necessity, generate explanations in the form of all-else equal claims. Moreover, because formal models leave out so much that is important at the individual level, these claims are best thought of as all-else-equal claims about the central tendency of behavior in collections of individuals with preferences, beliefs, information, and constraints similar to those in the model. The causal inference tradition is focused on research designs that make credible the claim to have held all-else-equal on average across such groups. Thus, the two traditions are naturally complementary.

This view of the role of theory and its connection to empirical analysis suggests rethinking some of the ways political scientists typically talk about this relationship. We want to highlight two.

First, political scientists often speak of statistical models as though the goal were to estimate the entire process in the world that generates some outcome (sometimes referred to casually as the "data generating process"). It's natural, from this point of view, to think that the goal of empirically useful theory is explicating such processes in their entirety. Combining these ideas leads to the view that the variables included in the statistical model and the formal model should be the same—i.e., that your formal model tells you what should be on the right-hand side of your regression.

Our view is quite different. Formal models generate qualitative hypotheses about particular all-else-equal relationships. The goal of empirical work motivated by such theory is to credibly assess those claims about all-else-equal relationships. Neither approach is about describing the entire process that gives rise to the outcome. Hence, the standard for whether some variable should be included in a regression is neither whether it was included in the motivating formal model nor is it whether that variable is part of the process that generates the outcome. Rather, the standard is whether omitting the variable would confound estimation of the relevant all-else-equal relationship.

Second, viewing all-else-equal claims as a bridge between theoretical and empirical research points to a way that the approaches developed by causal inference scholars are

[1] Of course, "all" should not be taken literally. For instance, it is a mistake, when estimating causal effects, to hold post-treatment variables equal.

actually of broader applicability than the name suggests. After all, the all-else-equal claims that come from formal models need not be causal (though they often are). Empirically assessing a non-causal, but still all-else-equal, theoretical claim requires the same rigorous attention to confounding that is required for estimating causal effects. As such, the research designs developed for causal inference are often the right tool for the job, even when the job is not causal inference.

We can illustrate both points in an examples about the performance of men vs. women in Congress (this example is based on Anzia and Berry, 2011). Consider a simple setting in which an electorate has to choose between two candidates, each drawn randomly from the same talent pool. The voters prefer more talented candidates, but also discriminate against women. In such a setting, if faced with a male and a female candidate, the voters will elect the female candidate only if she is sufficiently more talented than the male candidate to overcome the voters' bias against women. (It is straightforward to formalize this model and the results discussed below. We don't to conserve space.)

This theoretical model makes a causal prediction—increased voter bias causes an increase in the average quality of elected women and a decrease in the average quality of elected men. But the theoretical model also makes an interesting non-causal prediction—if talent translates into performance, then, all else equal, elected women will perform better than elected men, on average. This prediction is non-causal because the theoretical model does not assert that turning an elected official from a man into a woman improves performance, but it is nonetheless an all-else-equal claim.[2]

Empirically assessing this non-causal prediction requires just as much attention to issues of confounding as would estimating a causal relationship. Imagine a researcher taking this prediction to data from the U.S. House, comparing how well male and female representatives perform at procuring federal spending for their districts. Suppose she found that women perform better on average. She would be concerned that this difference might reflect attributes of the districts—e.g., that they are poorer, have larger African American populations, and are more urban—that both are associated with electing women and also lead those districts to attract more federal spending. As such, to credibly claim to have provided evidence for the theoretical model's explanation, she would need a research design that credibly held all-else-equal, even though the prediction being investigated is not causal.

Moreover, these concerns about specific confounders did not come from the theoretical

[2] This lack of a causal interpretation is distinct from the old issue of the potential conceptual incoherence of talking about causal effects of gender. We could have described the model and results in terms of voter discrimination about any feature of a candidate, including ones that are in principle experimentally manipulable. Gender happens to be the subject of Anzia and Berry (2011).

model. That model is not (and was not intended to be) a complete description of the process by which federal funds are allocated. It is a simplification meant to highlight the impact of voter discrimination. As such, it would have been an absurd distraction for the theoretical model to include district race or urbanness. These are in fact correlated with federal spending and with legislator gender, but were not the subject of the theoretical analysis. That the theoretical model does not itself suggest district characteristics as confounders, however, does not justify the empirical researcher in omitting them. The theoretical model offers an all-else-equal prediction. The job of the empirical researcher, if she wants to assess this prediction, is to find a research design that takes this all-else-equal condition seriously by addressing all possible confounders, whether they appear in the theoretical model or not.[3]

2 Datafication

Before turning to what role there is for big data in social science, we should start by saying what we mean. The term, “big data”, emerged in computer science to describe problems associated with data sets too large to be read into the memory of a single computer (Cox and Ellsworth, 1997). Big data in this sense remains an active area of research in computer science (National Research Council, 2013). But this is not what we are concerned with here. In political science, the term has come to refer, loosely, to a collection of techniques associated with data mining and machine learning.[4] (See, for example, the papers collected in the *Political Analysis* virtual issue on Big Data in Political Science.)

One productive way that the tools of machine learning have been put to good use in the social sciences is by automating various aspects of data collection, a process we call *datafication*. The most straightforward version of datafication is collecting data from very many, widely distributed locations, as in web scraping. In addition to straightforward data collection, machine learning techniques can also be used to convert non-traditional sources of information—e.g., text or images—into data that can be used for statistical analysis. For instance, Grimmer (2010) uses unsupervised learning techniques to code the substantive content of Congressional press releases in terms of the policy agendas they prioritize. It is impractical to do such data collection by hand, so it is only with the advent of machine learning techniques and powerful computers that the these information sources become

[3] Anzia and Berry (2011) use a difference-in-differences design.

[4] This includes methods such as clustering, classification and regression trees, and penalized regression techniques such as the LASSO. Hastie, Tibshirani and Friedman (2009) is a canonical introduction to these methods.

accessible to quantitative researchers.

It is critical to distinguish the use of machine learning for datafication, and the use of machine learning for inference-oriented data analysis. To see the distinction we have in mind, consider Gentzkow and Shapiro's (2010) analysis of the extent to which media slant is driven by consumer demand versus the ideology of newspaper owners. They use machine learning techniques for text categorization to classify the partisanship of newspapers. In particular, they measure media slant by comparing the words used in newspaper editorials to the words used in speeches by Congressional Republicans and Democrats. Once this datafication is done, they proceed to tackle their substantive question with a combination of theory and traditional causal inference approaches. They first show that observed media slant is inconsistent with the predictions of a theoretical model in which variation in slant is driven only by demand-side factors. Given this, they turn to more direct attempts to decompose the effects of consumer demand and owner ideology: they compare the media slant of newspapers with the same owner located in markets with different demand-side characteristics and they instrument for consumer political attitudes using a measure of religiosity.

In our view, Gentzkow and Shapiro's (2010) study exemplifies the promise of big data for the social sciences. Datafication unlocked new information for quantitative analysis, but those data were analyzed in light of a research design exploiting plausibly exogenous sources of variation in consumer political attitudes. It would have been impossible to provide a credible answer to the research question—all-else-equal, what are the separate effects of consumer and publisher ideology on media slant?—without such a research design. A naive regression of newspaper slant on consumer ideology would have done nothing to deal with the obvious problems of unobservable confounders—for instance, that conservative newspaper owners might simply be more likely to locate in conservative markets. This is because the naive regression depends on selection on observables as its research design—an incredible assumption in this setting. It follows that any method that relied on selection on observables for identification (e.g., matching) would not have yielded credible estimates of the causal effect. Indeed, the same is also true of machine learning techniques that assume selection on observables, including the most sophisticated machine learning techniques for estimating treatment effects, such as the post-double-LASSO estimator of Belloni, Chernozhukov and Hansen (2014).

The more breathless accounts of "big data" might lead one to suppose that this task of finding credible research designs can also be automated. But this supposition can't be right. Identifying a source of exogenous variation requires a creative insight about the phenomenon under study. Arguments for the credibility of such a source of exogenous variation rest on substantive knowledge of that phenomenon. No existing computational technique can come remotely close to doing either of these. For the time being, good empirical research will continue to rely on actual people doing actual thinking about social phenomena.

3 Familiar Mistakes at Scale

Not only do machine learning techniques not help with the development of research design, they in fact can be harmful for inference, even in the presence of a research design.

To see the issue, imagine one has a perfect research design—experimental randomization of a treatment. Here are two things that could go wrong. First, suppose there is only one outcome, but the researcher observes many covariates regarding the subjects. In principle, the researcher could search across all those covariates and all their possible interactions to find a sub-population on which the treatment effect was statistically significant. Of course, given enough independently varying covariates, such a sub-population is guaranteed to exist, even if there is no real treatment effect.

Second, suppose there are many possible outcomes that could in principle have been affected by the treatment. Again, the researcher could search across all these outcomes to find a statistically significant treatment effect. Given that statistically significant treatment effects will be found spuriously five percent of the time, the researcher who conducted 100 such tests should expect to find five significant effects, even when there are no real treatment effects. When only the five significant results are reported, or just some one of them, the reader has no basis for judging the probability that the result in question arose by chance, even when that probability may be much closer to 1 than to 0.05.

Of course, the problems of multiple hypothesis testing did not originate with machine learning. Other than time constraints, nothing stops a researcher from running all the regressions we just described by hand. But because machine learning automates and accelerates the process, we fear that it facilitates the industrial scale production of spurious results through multiple hypothesis testing. When coupled with the well known problems of publication bias (Gerber and Malhotra, 2008), industrial scale multiple hypothesis testing may seriously retard and distort the accumulation of knowledge.

One approach to the problem of multiple hypothesis testing is technical—for instance, by attempting to adjust p-values or control the false discovery rate. Political scientists have much to learn from the work being done in computer science and statistics on this front. However, we think a narrow technical focus will miss a larger problem. As we indicated above, the goal of theoretically motivated empirical research is to estimate a particular all-else-equal relationship. One ought to have a research design for the purpose of answering a specific research question. With this clear and guiding purpose, the analyst has no reason to be searching across many specifications (or sub-populations or outcomes) looking for results. Indeed, to the extent that the analyst finds herself searching across specifications, it should be for the purpose of probing the robustness of a theoretically interesting finding. Loosely speaking, the analyst should be searching to make sure there aren't reasonable specifications that make the theoretically motivated result disappear, rather than searching for the one specification that produces a statistically significant "finding". In this sense, we see the nascent preregistration movement, which urges researches to publicly state their research design and model specification before analyzing their data, and the associated willingness to publish null results as pre-requisites for big data or machine learning to be truly useful contributors to inference-oriented empirical research in political science.

4 Forecasting, Invariance, and Policy

A more radical view is that the rise of big data approaches will diminish the importance of explanation of the sort we discussed in Section 1. Such explanations are not only inherently interesting, but have also been important because they facilitate mechanism-based forecasting of the effects of policy interventions. Machine learning techniques facilitate purely statistical forecasting on a larger scale and with greater accuracy than could previously be conceived. Perhaps these improvements in purely statistical forecasting will allow big data to supplant mechanism-based forecasting.

What do we mean when we distinguish forecasting by purely statistical means from mechanism-based forecasting? At a conceptual level, one can see the distinction just by thinking about causal effects (as a special case of a mechanism) and ordinary least squares regressions. Suppose there are two variables, y and x, and that the true causal effect of a change in x on y is given by β. Further, suppose there is some other variable, z, which is correlated with both y and x. If a researcher regresses y on x, but can't observe z, then the omitted variables bias formula tells us that the regression coefficient, γ, is equal to β plus a correction term that depends on the correlation between x and z and partial correlation between y and z conditional on x. Purely statistical forecasting predicts y from x using γ. The causal effect (β) is of no particular interest. A mechanism-based forecast's predictions attempt to use information about β.

When do we want to take a purely statistical versus mechanism-based approach to forecasting? The advantage of the purely statistical approach is that it makes optimal use of all the information. In the example, it is precisely because γ exploits correlations with z that it is statistically optimal—the coefficient of the best linear predictor of y given x. So, if one is interested in simply forecasting y based on the x's that are already out there in the world, then one typically wants to estimate γ.

But there are also sometimes costs to this purely statistical approach. Any forecast relies on some type of extrapolation. Concerns about forecasts often reflect skepticism regarding invariance through the extrapolation. Hume's problem of induction arises from concerns about invariance across time. External validity problems arise from concerns about invariance of treatment effects across sites. The possibility of non-monotonic dose-response relationships raises concerns about invariance of incremental treatment effects across treatment intensities. Lack of overlap raises concerns about invariance of treatment effects across covariate values.

The cost of purely statistical forecasting is that it relies on a very strong invariance assumption that is typically false for policy interventions. Suppose one is interested in forecasting the effects on y of a policy intervention that changes x. Having changed the relationship between x and z means that it no longer makes sense to think that the observed values of x are informative about z in the way they were prior to the intervention. Forecasting as if the relationship between x and z were invariant is, thus, a mistake.

To make this a bit more concrete, consider a simple example. Suppose we observe data on vote share and challenger campaign spending. It is plausible that high quality challengers have an easier time raising money and do better in elections for reasons independent of that money as well. The coefficient in a regression of vote share on challenger spending (γ) reflects information about candidate quality as well as the true causal effect of spending (β). As such, it is a biased estimate of the causal effect. Nonetheless, if we simply want to forecast election outcomes based on observed campaign spending, we want to use this regression coefficient because all of the information it contains is useful for the forecast. The situation is very different if we want to evaluate the consequences of public financing of campaigns for electoral competitiveness. A strong version of such a reform (for instance, one that simply gives all candidates the same funds) will eliminate the relationship between spending and quality. Hence, it would be a mistake to make a purely statistical forecast—i.e., one that assumes that relationship between challenger spending and election outcomes will continue to be given by γ after the reform.

Of course, the mechanism-based forecasting approach also relies on invariance assumptions to make a forecast of the effects of this reform. Assume we have a research design that identifies β. Now, if we want to simply forecast the portion of the effect of the reform that operates through the direct causal effect of spending, we need to assume that voter responses to challenger spending are invariant to the source of challenger funds. If we want to go further and forecast the total effect of the reform, we need also to assume invariance of all non-spending characteristics of challengers. While both of these assumptions are arguably quite strong, they are weaker than needed for the purely statistical forecasting approach. In particular, if either of these assumptions does not hold, then (other than by a miracle) γ will not remain invariant. But the reverse is not the case.

The deep challenge of forecasting the effect of policy interventions is precisely that even invariance assumptions like these latter two might well not hold. For example, invariance of voter response would not hold if voters draw different inferences about candidates when they observe spending raised from donors rather than given from the public coffers (Prat, 2002; Ashworth, 2006; Houser and Stratmann, 2008). Similarly, invariance of non-spending characteristics of challengers would not hold if the availability of public financing encourages a different group of candidates to seek office (Tillmann, 2014). Each of these is an instance of a more general kind of invariance failure— strategies of actors may not be invariant across policy regimes. (In some economic contexts, this point is called "the Lucas critique", after the forceful articulation of the point in Lucas (1976).)

Typically, political scientists are called on to forecast the effects of policy interventions in settings—e.g., campaign finance regulation, civil service reform, counterinsurgency strategies— in which strategic adjustment is of the essence. In such settings, we cannot escape the Lucas critique. The data (whether big or small) are always drawn from the world that exists in the absence of the policy intervention. But an accurate forecast of the effect of the intervention requires thinking counterfactually about situations for which there may in fact be no data (even in principle) to reveal how strategic adjustment will occur. In such settings, the only hope is a combination of theoretical accounts of mechanisms and data that are informative about the magnitudes of the effects associated with those mechanisms which, together, might give the researcher some leverage over the key counterfactuals.

References

Anzia, Sarah F. and Christopher R. Berry. 2011. “The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?” *American Journal of Political Science* 55(3):478–493.

Ashworth, Scott. 2006. “Campaign Finance and Voter Welfare with Entrenched Incumbents.” *American Political Science Review* 100(1):55–68.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. “Inference on Treatment Effects after Selection among High-Dimensional Controls.” *The Review of Economic Studies* 81(2):608–650.

Cox, Michael and David Ellsworth. 1997. Application-controlled Demand Paging for Out-of-core Visualization. In *Proceedings of the 8th Conference on Visualization '97*. VIS '97 Los Alamitos, CA, USA: IEEE Computer Society Press pp. 235–ff.
URL: *http://dl.acm.org/citation.cfm?id=266989.267068*

Gentzkow, Matthew and Jesse M Shapiro. 2010. “What drives media slant? Evidence from US daily newspapers.” *Econometrica* 78(1):35–71.

Gerber, Alan and Neil Malhotra. 2008. “Do Statistical Reporting Standards Affect What is Published? Publication Bias in Two Leading Political Science Journals.” *Quarterly Journal of Political Science* 3(3):313–326.

Grimmer, Justin. 2010. “A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.” *Political Analysis* 18(1):1–35.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second ed. Springer-Verlag.

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

Houser, Daniel and Thomas Stratmann. 2008. “Selling favors in the lab: experiments on campaign finance reform.” *Public Choice* 136(1-2):215–239.

Lucas, Robert E, Jr. 1976. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*. Vol. 1 Elsevier pp. 19–46.

National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, D.C.: The National Academies Press.

Prat, Andrea. 2002. "Campaign Advertising and Voter Welfare." *Review of Economic Studies* 69(4):997–1017.

Tillmann, Philipp. 2014. "Entry into Electoral Races and the Quality of Representation." University of Chicago typescript.