Using Double-Lasso Regression for Principled Variable Selection

Oleg Urminsky

Booth School of Business, University of Chicago

Christian Hansen

Booth School of Business, University of Chicago

Victor Chernozhukov

Department of Economics and Center for Statistics, Massachusetts Institute of Technology

Please address correspondence to:
Oleg Urminsky
Booth School of Business
University of Chicago
5807 S. Woodlawn Ave Chicago, IL 60637
oleg.urminsky@chicagobooth.edu

Abstract

The decision of whether to control for covariates, and how to select which covariates to include, is ubiquitous in psychological research. Failing to control for valid covariates can yield biased parameter estimates in correlational analyses or in imperfectly randomized experiments and contributes to underpowered analyses even in effectively randomized experiments. We introduce double-lasso regression as a principle method for variable selection. The double lasso method is calibrated to not over-select potentially spurious covariates, and simulations demonstrate that using this method reduces error and increases statistical power. This method can be used to identify which covariates have sufficient empirical support for inclusion in analyses of correlations, moderation, mediation and experimental interventions, as well as to test for the effectiveness of randomization. We illustrate both the method's usefulness and how to implement it in practice by applying it to four analyses from the prior literature, using both correlational and experimental data.

Keywords: research methods, covariate, regression, variable selection, confound, omitted variable bias

Although people's behavior is shaped by many factors, psychological research typically attempts to isolate the effects of one construct of interest (or sometimes a small number of constructs). In focusing on a key predictor, it is not always clear how to best account for the possibility that other factors may also affect the outcome variable. While statistically controlling for valid covariates in correlational or experimental analyses can yield more accurate estimates and significance tests of focal effects, including unnecessary covariates can also be problematic, and can even be misused.

In this paper, we present a two-step method using lasso regression (Belloni, Chernozhukov, & Hansen 2014) as a practical solution to the problem of principled variable selection for covariates, drawing on recent advances in statistics and econometrics. We apply this method to four datasets drawn from recent literature to illustrate the usefulness of double-lasso variable selection in both correlational analyses and experimental designs.

The Covariate Selection Problem

Analyses that fail to take into account valid predictors of the dependent variable can suffer from multiple problems. In correlational analyses, omitted variables that predict the dependent variable and are correlated with the focal independent variable(s) can cause bias in estimated parameters (Darlington, 1990; Mauro, 1990). When valid covariates are excluded, the estimated coefficient of interest may either be artificially strong (when the covariate is a confound), or may be artificially weak (a suppression effect; MacKinnon, Krull, & Lockwood, 2000; Thompson, 2006).

A common way of avoiding this problem is to use experimental manipulations as independent variables. When randomization is successful, the experimental independent variable

should be uncorrelated with any omitted variables, precluding bias in estimating the effect of the independent experimental variable. However, in practice, attempted random assignment may not always yield the desired independence (Darlington, 1990; Zhou & Fishbach, 2016), and formal tests are rarely conducted to confirm randomization (Wilkinson, 1999). Even when randomization is successful, failure to statistically control for valid predictors of the dependent variable reduces the statistical power of the experiment (Darlington, 1990; Judd, McClelland, & Ryan, 2011), exacerbating the already typically low likelihood of detecting a true effect of the independent variable (Rossi, 1990).

Given these benefits of controlling for covariates, it may seem surprising that the practice is not widely promoted and is not prevalent in the literature. The common absence of covariates is likely attributable to several factors. As a practical concern, researchers may rely on rules of thumb (e.g., Green, 1991) and conclude that their sample size is too small to support including more predictors. In the extreme case, in datasets with more potential covariates than participants, including all the variables in a linear regression is not even possible.

Furthermore, it is not always clear how to go about selecting covariates and discussions of best practices in psychological research provide little guidance on this issue (Cumming, 2014; Wilkinson, 1999). Automatically controlling for a standard set of variables, such as demographics, is not recommended (Meehl, 1971). Automated methods, particularly stepwise regression, are widely recognized to perform poorly (i.e., over-fitting the data, selecting non-optimal models, inflating R^2; Freedman, 1983; Thompson, 1995; Thompson, 2006) and are rarely used.

In fact, recent work on research integrity has cautioned that the decision of whether or not to control for covariates can contribute to the problem of researcher degrees of freedom and false positives. Simmons, Nelson, and Simonsohn (2011) provide an elegant example of how selective reporting of only those analyses including controls that contribute to a significant focal result can lead to spurious findings. In fact, Simonsohn, Nelson, and Simmons (2014) use presence of covariate controls as a suspicious characteristic to distinguish between studies more or less likely to have been "p-hacked", and report evidence supporting this suspicion. Thus, researchers may feel that it is simpler and more conservative to report main effects without covariates. Even when controlling for covariates could be beneficial, the lack of established principled methods for doing so may discourage researchers from doing such analyses.

Double-Lasso Variable Selection

We propose that recently developed methods based on lasso regression (e.g. Tibshirani, 1996) provide a useful solution to these problems. We describe a "double-lasso" approach (Belloni et al., 2014) that can help researchers select variables for inclusion in analyses in a principled manner that avoids inflated Type I errors. The goal is to identify covariates for inclusion in two steps, finding those that predict the dependent variable and those that predict the independent variable. The second step is important, because exclusion of a covariate that is a modest predictor of the dependent variable but a strong predictor of the independent variable can create a substantial omitted variable bias. In experimental data, the second step also serves as a test of randomization. While we recommend using lasso regression, calibrated to avoid over-fitting, in these variable selection steps, we also discuss similarly-performing alternative methods. The variables selected in either step are then included in the regression of interest.

Without loss of generality, we focus on the case with a single focal independent variable of interest, $X_{i,}$ and we want to know how it relates to dependent variable Y_i. The focal variable X_i could be either a measured variable or an experimental condition code. In addition, we have multiple potential covariates, W_{i1} to W_{iK}. We could estimate a linear regression model, finding βs that minimize the sum of squared errors in the regression equation:

$$Y_i=\beta_0+\beta_1X_i+\beta_2W_{i1}+\ldots+\beta_{K+1}W_{iK}+\varepsilon_i$$

A lasso regression instead finds βs that minimize the sum of squared errors in the regression equation with an additional penalty term:

$$\text{Min}[\Sigma_i(Y_i-\beta_0+\beta_1X_i+\beta_2W_{i1}+\ldots+\beta_{K+1}W_{iK})^2+\lambda\ \Sigma_k|\beta_k|]$$

The penalty term results in the lasso regression shrinking the estimated regression coefficients towards zero and potentially setting coefficients on some variables exactly to zero, both of which help reduce over-fitting. The lasso, by setting some coefficients to zero, also performs variable selection. These shrinkage properties allow Lasso regression to be used even when the number of observations is small relative to the number of predictors (e.g. discussion in James, Witten, Hastie, & Tibshirani, 2013).

However, directly using lasso regression can be problematic. Those lasso-estimated coefficients that are actually non-zero are typically underestimated, and lasso may mistakenly exclude variables with non-zero coefficients, particularly variables with moderate effects. Each of these phenomena generally causes significant regularization bias that adversely affects estimation and inference about β_1. The omission of covariates with moderate but non-zero coefficients is especially problematic and results in omitted variable bias when these covariates

are relevant predictors of the focal variable. In order to overcome such biases, we recommend using the "double-lasso" variable selection procedure (Belloni, et al., 2014), which was explicitly designed to alleviate both sources of bias, as follows:

Step 1: Fit a lasso regression predicting the dependent variable, and keeping track of the variables with non-zero estimated coefficients:

$$Y_i=\alpha_0+ \alpha_1 W_{i1}+\ldots+ \alpha_K W_{iK}+\varepsilon_i$$

Step 2: Fit a lasso regression predicting the focal independent variable, keeping track of the variables with non-zero estimated coefficients:

$$X_i=\delta_0+ \delta_1 W_{i1}+\ldots+ \delta_K W_{iK}+\varepsilon_i$$

If X_i is an effectively randomized treatment, no covariates should be selected in this step.

Step 3: Fit a linear regression of the dependent variable on the focal independent variable, including the covariates (W_{ik}) selected in either of the first two steps:

$$Y_i=\beta_0+\beta_1 X_i+\Sigma_{k \in A} \beta_{k+1} W_{ik}+\varepsilon_{i,}$$

In the equation, *A* is the union of the variables estimated to have non-zero coefficients in Steps 1 and 2. This regression could also include a small set of additional covariates identified a priori as necessary. Interpret and report the coefficient estimates and significance tests on the focal variable(s) as the final results.

While implementation is fairly straightforward and extends easily to multiple focal variables by repeating Step 2 for each, the choice of the tuning parameter λ is very important for successfully avoiding over-fitting (see SOM-R). Dedicated code for this procedure is available

for use with STATA. We also provide details and examples in the SOM-U for how to apply this method in both STATA and in the R statistical package, and how to closely approximate the results using modified forward-selection in SPSS (Kozbur, 2015).

We note that other approaches are available in the statistics literature (as reviewed, e.g., in Chernozhukov, Hansen & Spindler, 2015) including approaches, such as Athey & Imbens (2015), that accommodate inferring treatment effects in settings with fully heterogeneous treatment effects where the linear model may be inappropriate (Imbens & Rubin, 2015).

Simulation Results

We ran a simple simulation to illustrate the practical benefits of double-lasso variable selection. We generated 10,000 datasets from eight known sets of parameters, varying the number of available covariates and the sample size (full details in the SOM-R). We compare the double-lasso procedure to five alternatives: regression including no covariates ("none"), including all the covariates when possible ("all"), including all covariates selected in a step-wise regression ("stepwise"), choosing covariates to maximize the chances of the independent variable being significant ("p-hacking") and, as a baseline, using the correct variables ("true").

On average (see Table 1), the double-lasso selected close to the right number of covariates (2.6 vs. 2 actual), and far fewer than stepwise (5.5) or p-hacking (4.9). The average error in estimating the coefficient of the dependent variable in the double-lasso was very close to the true baseline (RMSE=.261 vs. .246), lower than including no covariates (RMSE=.356) or all covariates (RMSE=.383) and much lower than p-hacking (RMSE=.496). The significance test for the focal independent variable was well calibrated in the double lasso, rejecting the null hypothesis 5.4% of the time, similar to "none" and "all" (5.3% and 5.6%, respectively) and much

better than "stepwise" (11.8%) or p-hacking (29.8%). Perhaps most importantly, using the double-lasso yielded substantial benefits in statistical power compared to those other methods that also did not over-reject (.579 vs. .626 for the "true" baseline, compared to .323 for "none" and .396 for "all").

We also tested two procedures that roughly approximate the double-lasso. Two-step multiple regression (including all potential covariates, and then re-running the regression removing non-significant covariates) provides reasonable solutions, but underperforms the double-lasso and is infeasible in settings with more available covariates than observations. Double-forward regression (using forward regression to do both steps, with modified p-value cutoffs, see the SOM-R) yields results quite similar to the double lasso.

Table 1.

Comparison of double-lasso with alternative analysis methods, averaging across simulated datasets.

Method	Variables Selected	Bias	RMSE	Size	Power
True model (baseline)	2	0.001	0.246	0.050	0.626
No covariates	0	0.001	0.356	0.053	0.323
All covariates	40	0.001	0.383	0.056	0.396
P-hacking	4.9	0.279	0.496	0.298	0.911
Stepwise regression	5.5	-0.092	0.268	0.118	0.511
Two-step multiple regression	3.8	0.003	0.274	0.066	0.596
Double-forward-regression	1.6	-0.007	0.261	0.053	0.587
Double-Lasso	2.6	-0.015	0.261	0.054	0.579

Notably, the double-lasso performs well even in demanding situations. For example, with a sample size of 60, it is impossible to include 120 covariates in a standard regression, and researchers in this situation might therefore not include any covariates. However, double-lasso variable selection identifies covariates and improves on the no-covariate model in this situation, with both lower error (RMSE=.276 vs. .390) and higher statistical power (.422 vs. .261). Next,

we illustrate the proposed variable selection method by using it to analyze four datasets discussed in the literature.

Analysis 1: Correlational Analysis of Parents' Life Satisfaction

Nelson, Kushlev, English, Dunn, and Lyubormirsky (2013) concluded, based on analyses of three datasets, that parents report relatively higher levels of life satisfaction, happiness, and meaning in life than do nonparents, contrary to some prior research. Their analyses were based on mean comparisons (t-tests) and correlations, without including any covariates as statistical controls. These conclusions were criticized by Bhargava, Kassam, and Loewenstein (2014), who re-analyzed the data and found that the satisfaction and happiness were not higher for parents, controlling for demographics.

In Study 1, Bhargava et al. (2014) reported a significant positive relationship between parental status and life satisfaction in the World Values Survey data without any controls (β=0.224, p<.001), similar to Nelson et al. (2013). This suggests that parents are more satisfied with their lives than non-parents, on average. However, they also find that the relationship is instead negative when controlling for marital status, age and gender (β=-0.144, p=.04), and is non-significant when controlling for income as well (β=-0.065, p=.34).

Thus, whether the relationship between life satisfaction and parenthood was significantly positive, non-significant or significantly negative depended on whether and which covariates were included. One important consideration is whether the potential covariates should be considered controls, on theoretical and logical grounds (e.g, as opposed to being analyzed as potential mediators). Assuming that the variables are valid potential controls, it is not clear which covariates should be included and what conclusion should be drawn. In fact, the lack of

difference found by Bhargava et al. (2014) could potentially even be spurious, a result of multiple testing of many potential covariates. The double-lasso addresses this concern by identifying whether there is sufficient empirical justification for including the published covariates and potentially identifying other covariates.

To be comprehensive and test the ability of the method to handle many potential covariates, we began with 9 demographic variables, created dummy codes from the categorical variables and computed powers of the variables and interactions, yielding a total of 524 potential covariates. Using the procedure described above, we then identified a subset of 17 covariates for which there was sufficient empirical support to be included in the final test. These were covariates that were either strong predictors of life satisfaction or of being a parent.

In particular, our analysis using double-lasso variable selection confirms that there is sufficient evidence to include the variables identified by Bhargava et al. (2014) (age, gender, marital status and income) as covariates, as well as several other variables and interactions (Table 2). In the resulting model, we find a significant negative relationship between parental status and life satisfaction (β=-0.196, p=.006), controlling for the identified covariates. Note that it is important for the results to conduct both variable selection steps. If we only use covariates identified as predictors of the dependent variable (life satisfaction) and leave out those identified as predictors of being a parent, we instead find a weaker negative relationship between parental status and life satisfaction (β=-0.127, p=.051).

While this analysis tells us which potential covariates have empirical support for inclusion, it cannot determine which variables make logical sense or are theoretically justified to include. As an example, one potential concern with this analysis is that parenthood might

causally impact some of the variables, particularly income and employment status. When we re-run the regression excluding these variables as potential covariates, we still find a significant negative relationship between parental status and satisfaction (β=-.221, p=.002), controlling only for marital status, age, gender and interactions of those covariates (see SOM-U).

Table 2.

Regression of parenthood on life satisfaction, with double-lasso selected covariates.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	6.750	0.128	52.57	.000	6.498	7.001
Parent	-0.196	0.071	-2.75	.006	-0.336	-0.056
Main effect covariates:						
Married (including living together as married)	0.513	0.157	3.27	.001	0.206	0.821
Income (3 point scale)	0.582	0.119	4.90	.000	0.349	0.815
Age	0.912	0.235	3.88	.000	0.451	1.373
Age=18	0.300	0.213	1.41	.159	-0.117	0.717
Age=19	0.129	0.213	0.61	.544	-0.288	0.547
Age=20	0.521	0.191	2.73	.006	0.147	0.896
Age=21	0.175	0.175	1.00	.319	-0.169	0.518
Age=22	0.545	0.169	3.23	.001	0.214	0.876
Age=23	0.187	0.171	1.09	.274	-0.148	0.523
Gender (Male)	-0.142	0.063	-2.24	.025	-0.267	-0.018
Employment: Housewife	0.066	0.123	0.54	.590	-0.175	0.308
Chief wage earner	0.144	0.070	2.07	.038	0.008	0.281
Interaction covariates:						
Married x Age	0.143	0.352	0.41	.685	-0.547	0.833
Married x Age to fourth power	0.491	0.608	0.81	.419	-0.701	1.684
Married x Income rating (3 point)	-0.106	0.181	-0.59	.557	-0.461	0.248
Married x Income rating (11 point)	0.303	0.221	1.37	.170	-0.130	0.737
Employment: Student x Male	0.344	0.269	1.28	.202	-0.184	0.871

Nelson et al. (2013) and Bhargava et al. (2014) also debated whether demographics moderated the effect of parenthood on life satisfaction, with the latter paper arguing that evidence of such moderation was weak. In separate analyses, we find that marital status (β=.362,

p=.015) and age (β_{age}=3.281, p=.009; $\beta_{age\text{-}squared}$=-3.107, p=.045) moderates the effect of being a parent, controlling for selected covariates, but gender does not.

Overall, our findings parallel those of Bhargava et al. (2014). The higher self-reported life satisfaction of parents may be explained primarily by the differing demographic characteristics of parents vs. nonparents, disguising lower life-satisfaction among parents, all else equal. Controlling for covariates, we also find a significant relationship between parenthood and lower happiness (β=-.050, p=.034). In contrast, we find a significant positive relationship between parenthood and more thoughts about meaning in life (β=.094, p=.002, see SOM-U).

Analysis 2: Mediation Analysis of Conservative Happiness

Next, we look at how double-lasso regression can be used to inform variable selection when conducting a mediation analysis. Napier and Jost (2008) report that conservatives demonstrate higher levels of subjective well-being than liberals do. Using large secondary data sets, they identify rationalization of inequality as a mediator of this difference, in accordance with system-justification theory. However, a potential concern is that the mediation result may be spurious, if the mediator is merely a proxy for other factors that relate to political orientation and subjective well-being. While the paper reports mediations including selected demographic controls, we can use lasso regression to test whether inclusion of the covariates is supported by the data.

We reanalyzed the 1,192 participants in the 2000 American National Election Survey who had completed the measures analyzed in Study 1 of Napier and Jost (2008).[1] We confirmed

[1] Our sample size differs slightly from the original paper. An additional analysis without excluding missing demographic variables is presented in the SOM.

that conservatives were happier (β=.282, t(1190)=4.70, p<.001), and the difference was partially mediated by the rationalization of inequality scale (indirect β=.062, bootstrap CI=[.0004,.1246], p=.049, Preacher & Hayes, 2004).

The dataset also includes a large number of demographic variables that could be potentially confounding the mediation result. We chose 35 demographic variables, including all the variables used in the paper, and created dummy codes for all categorical variables. We conducted a double-lasso analysis to test the resulting 141 demographic variables for inclusion, and identified four covariates (church attendance and dummy codes for employment, being married and African-American ethnicity). Controlling for these covariates, the effect of political orientation on happiness is marginally significant (Model 1, Table 3). While the coefficient of political orientation does decrease when rationalization of inequality is included (Model 2, Table 3), the mediation is not significant (indirect β=.045, bootstrap CI=[-.011,.102], p=.116). These results suggest that the proposed mediation in the original study is sensitive to the inclusion of a more complete set of covariates, which are identifiable using the double-lasso procedure.

Table 3

Regressions of political orientation on happiness.

	Model 1				Model 2			
Variable	β	*SE*	*t*	*P*	β	*SE*	*T*	*p*
Primary variables:								
Constant	2.048	.183	11.18	.000	1.803	.225	8.03	.000
Political Orientation	0.125	.067	1.87	.061	0.078	.073	1.06	.287
Mediator:								
Rationalization of inequality					0.090	.053	1.69	.091
Covariates:								
Church attendance	0.101	.030	3.34	.000	0.102	.030	3.36	.001
Married	0.235	.074	3.16	.002	0.225	.074	3.03	.003
Not unemployed or disabled	0.671	.162	4.13	.000	0.686	.161	4.25	.000
Black	-0.270	.132	2.05	.041	-0.233	.132	1.77	.077

Analysis 3: Spurious Experiment on Chronological Rejuvenation

While controlling for covariates can be helpful, as in the prior analyses, Simmons et al. (2011) have shown that opportunistically controlling for covariates can contribute to spurious findings. The potential for controls to distort the primary findings is a particular concern in experimental studies, where successful randomization addresses issues of confounding. We re-analyzed the data from Study 2 of their paper to test whether using the double-lasso would reduce the likelihood of including spurious covariates.

In Study 2, Simmons et al. (2011) present an intentionally spurious finding, reporting that a randomized experimental intervention (having people listen to "When I'm Sixty Four" by the Beatles vs. a control song) had a significant effect on the participant's age, controlling for father's age (M=20.1 vs. 21.5 years old, β = -521.85, $t(17)=2.22$, $p=.040$). The covariate was also significant in the regression (β = 98.34, $t(17)=3.86$, $p=.001$).

We identified nine potential covariates in the dataset. Dummy-coding a categorical variable yielded 10 variables. The double-lasso takes into account not only the multiple comparisons, but also the small sample size (N=20), setting a higher bar for covariates to be included. As a result, the double-lasso analysis revealed insufficient empirical support to include any of the potential covariates in the regression. The resulting single-variable regression accurately revealed no significant effect of the experimental manipulation (β = -305.30, $t(18)=1.00$, $p=.329$). In this case, simply using a principled variable-selection method eliminated the spurious finding.

Analysis 4: Suggested Defaults Experiment in Donation

In the last analysis, we demonstrate that the double-lasso method can be used to identify valid covariates in a randomized experiment, testing whether randomization was successful, and increasing the statistical power to detect a result. Goswami and Urminsky (2016) conducted a field experiment on the effect of recommended amounts in donation appeals. Each appeal letter included three donation amounts (low, medium and high) that were based on the recipient's most recent donation amount. In the four focal conditions, appeal letters were randomly assigned to include a recommendation to give the low amount, medium amount, high amount or to not include any recommended amount.

One of the hypotheses tested was a "scale-back" effect, in which donors would anchor on the recommended amount and give less when the low amount was recommended, compared to the control condition with no recommendation. People who chose to donate gave less in the low-recommendation condition (vs. control), but the difference was not significant (M=$162 vs. $283, $t(46)=1.38$, $p=.175$). Ten potential covariates were available, to which we added non-linear transformations and interactions, for a total of 196 potential covariates. There was a strong a priori rationale for including one of the potential covariates, the prior amount donated, since the choice options in the appeal letter actually differed depending on the most recent donation.

We used a double-lasso regression to identify which covariates had empirical support for inclusion. First, none of the covariates had a significant relationship with experimental condition. While this might be expected since the conditions were randomly generated in the stimuli, it can be problematic to simply assume effective random assignment in the data collected (Darlington,

1990). Thus, this step provides a valuable tool for validating the effectiveness of randomization, which is rarely tested in psychological research.

Based on the relationship to amount donated, three covariates were identified (most recent donation amount and two interactions) in the lasso step predicting donation amount. As noted above, no covariates were selected in the step predicting experimental condition, revealing no evidence of failed randomization based on the covariates. A linear regression including the identified covariates (as well as orthogonal experimental conditions), confirmed that there was a significant effect of the low recommendation condition on log donation amount relative to control (β= -.365, t(67)=3.03, p=.004, Table 4). This analysis demonstrates how double-lasso regression can be used to identify valid covariates, increasing the power of experimental tests.

Table 4.

Regression of randomized suggestion level on donation amount among donors, with double-lasso selected covariates.

Variable	β	SE	*t*	*p*	Low CI	High CI
Primary variables:						
Constant	1.064	0.303	3.51	.001	0.459	1.670
Low Amount Recommended	-0.365	0.121	-3.03	.004	-0.605	-0.124
Medium Amount Recommended	-0.130	0.1365	-0.95	.344	-0.402	0.142
High Amount Recommended	0.025	0.1465	0.17	.865	-0.267	0.316
Other randomized manipulations:						
Five earmarking options shown	-0.202	0.104	-1.94	.057	-0.409	0.006
Previous donation shown	0.011	0.098	0.11	.911	-0.185	0.207
Selected covariates:						
Log of prior year donation	0.695	0.099	7	.000	0.497	0.8937
Log of prior year donation x Log of lifetime giving	0.016	0.008	2.06	.043	0.0005	0.032
Consecutive years of giving squared x Prospect status	0.0002	0.0006	0.34	.739	-0.0009	0.001

Concluding Remarks

The simulation and re-analyses demonstrate the potential benefits of using a principled variable selection method, such as the double-lasso, for better identifying which covariates to include and not include in analyses across a range of situations. It is important to emphasize that the analytic method presented here cannot determine either the role that selected variables should play, or how their effects on the relationship of interest should be interpreted. A confound, a manipulation check and a mediator may all have similar statistical relationships in the data (MacKinnon, Krull, & Lockwood, 2000; Zhao, Lynch, & Chen, 2010), and these distinctions should typically be made on theoretical grounds.

However, either including all covariates or ignoring covariates entirely, either because of the conceptual difficulty of identifying the theoretical role of the variable or because of the potential for covariates to be used improperly (i.e., in p-hacking), is no solution. Failing to control for valid covariates can yield biased parameter estimates in correlational analyses or in imperfectly randomized experiments and contributes to underpowered analyses even in effectively randomized experiments. As demonstrated in the analyses, double lasso variable selection can be useful as a principled method to identify covariates in analyses of correlations, moderation, mediation and experimental interventions, as well as to test for the effectiveness of randomization. While variable selection methods are no substitute for thinking about what the variables mean, the approach presented here can provide an empirical basis for determining which variables to think hard about.

Author contribution:

The first author analyzed the data and wrote the first draft of the paper. The second author assisted in the data analysis and conducted the simulations. All three authors collaborated on the final draft of the paper.

References

Athey, S. & Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. NBER working paper.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81(2), 608-650.

Bhargava, S., Kassam, K. S., & Loewenstein, G. (2014). A reassessment of the defense of parenthood. *Psychological Science*, *25*(1), 299-302.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7, 649-688.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, 25(1), 7-29

Darlington, R. B. (1990). *Regression and linear models* (Chapters 4 and 8). New York: McGraw-Hill.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, *37*(2), 152-155.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 499-510.

Goswami, I., & Urminsky, O. (2016) When should the ask be a nudge? The effect of default amounts on charitable donations, *forthcoming, Journal of Marketing Research.*

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

Kozbur, D. (2015). Testing-based forward model selection. Working paper arXiv:1512.02666.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173-181.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, *108*(2), 314.

Meehl, P. E. (1971). High school yearbooks: a reply to Schwarz. *Journal of Abnormal Psychology*, *77* (2), 143-148

Napier, J. L., & Jost, J. T. (2008). Why are conservatives happier than liberals? *Psychological Science*, *19*(6), 565-572.

Nelson, S. K., Kushlev, K., English, T., Dunn, E. W., & Lyubomirsky, S. (2013). In defense of parenthood children are associated with more joy than misery. *Psychological Science*, *24*(1), 3-10.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717-731.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*,*58*(5), 646.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, 143, 534–547.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. *Educational and Psychological Measurement* 55, no. 4 (1995): 525-534.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. (Chapters 8 and 9) Guilford Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B,* 58(1), 267-288.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197-206.

Zhou, H. & Fishbach, A. (2016) Selection Bias in Online Panels: How Unattended Participant Attrition Lead to Surprising (yet False) Research Conclusions. Working paper.

Supplemental Online Materials - Reviewed

Implementation of the double-lasso procedure

A key issue in implementation is how to set the tuning parameter λ. By construction, a sufficiently low value would result in all covariates being included, while a very high value would result in no covariates being included. Based on prior work (e.g. Belloni, Chen, Chernozhukov and Hansen (2012) and Belloni, Chernozhukov and Hansen (2014)), we suggest using

$$\lambda = 2.2\sigma_R\sqrt{N}\Phi^{-1}\left(1-\frac{\alpha}{2K\ln(N)}\right).$$

In this expression, N is the sample size, K is the number of potential covariates being tested, Φ^{-1} is the standard normal distribution inverse CDF, and σ_R is the standard deviation of the residuals. In practice, the standard deviation of the residuals needs to be estimated. Estimation may proceed by first fitting a simple model, such as a model with just an intercept, and using the residuals to form an initial guess for the standard deviation. This initial guess may then be used to form λ for use in the lasso regression. One can then use the residuals from the lasso regression to update the guess for the standard deviation and repeat for a small number of iterations. Details are provided in Belloni, Chen, Chernozhukov and Hansen (2012) who also provide an appropriate generalization for heteroskedastic data. The basic procedure can also be modified to obtain estimates of average treatment effects (Belloni, Chernozhukov and Hansen 2014) and more general treatment effects (Belloni, Chernozhukov, Fernández-Val and Hansen 2015) suitable for general heterogeneous treatment effect settings as well.

This choice of λ ensures that, in the limit (i.e. as N gets very large), a model with good statistical properties and approximately the right number of covariates is chosen. In our analyses, we set α=.10. Heuristically, we note that if one were considering a Bonferroni correction for testing K hypotheses using two-sided t-tests maintaining an error rate of α, one would use a cutoff of $\Phi^{-1}\left(1-\frac{\alpha}{2K}\right)$, which is closely related to the choice above. The additional terms, especially the ln(N), can be viewed as additional factors that aid in making sure the procedure screens out all variables that are not highly relevant. In general, the method will require "more evidence" for inclusion of a covariate as the sample size increases and as the number of potential predictors increases, all else equal.

The next issue is how to easily implement the procedure. A macro is available for use with STATA, and the *glmnet* procedure in R can also be used (with some minor modifications, including using λ/2N as the regularization parameter). While the double lasso is not available in SPSS, a "double forward regression" approach can be used in SPSS (and most other statistical software programs) that closely approximates the double-lasso, by setting the *p*-value for entry to .1/[ln(N)*K] . This choice roughly corresponds to the analysis in Kozbur (2015) but does not include additional adjustments that are needed in the theory but are difficult to implement in SPSS. Without these additional corrections, formal validity of the procedure is theoretically questionable except under restrictive conditions. However, in our simulations and data analyses, the "double forward regression" performs well and yields similar results to the double lasso. Analyses of the examples using all three programs, including executable scripts, can be found in the unreviewed appendix.

Simulations

For our simulations, we generate data from a model representing a randomized trial given by

$$Y_i=\beta_0+\beta_1 X_i+\beta_2 W_{i1}+\ldots+\beta_{K+1} W_{iK}+\varepsilon_i.$$

The treatment variable X_i is set equal to one for half of the observations and 0 for the remaining observations, and we set $\beta_1 = .5$. We generate controls $W_{i1}, \ldots, W_{iK}$ by drawing initial variables $Z_{i1}, \ldots, Z_{iK}$ from a multivariate normal distribution with the mean and variance of each component equal to 0 and one, respectively, and correlation between Z_{ik} and Z_{il} equal to $.7^{|k-l|}$. We then set $W_{ik} = Z_{ik}$ for $k \leq K/2$ and $W_{ik} = 1(Z_{ik} > 0)$ for $k > K/2$ where $1(\cdot)$ denotes the indicator function that returns one when the expression inside the parentheses is true and 0 otherwise. We then set the coefficient on W_{i1} and the coefficient on $W_{i,K/2+1}$ equal to one. We consider n = 60 with K = 15, 30, 45, 60, and 120 and n = 100 with K = 25, 50, 75, 100, and 200. The error terms, ε_i, are drawn as iid mean 0 normal random variables.

We consider a homoskedastic case where the variance of ε_i is equal to one, and a heteroskedastic case where $\text{Var}(\varepsilon_i) = [.15\exp\{\beta_0+\beta_1 X_i+\beta_2 W_{i1}+\ldots+\beta_{K+1} W_{iK}\}]^2$. We report bias and RMSE for estimating β_1, rejection frequency for 5% level t-tests of the null hypothesis that β_1 is equal to the true value of .5 (size), and power of 5% level t-tests against the alternative that β_1 is equal to 0. All results are based on 10,000 simulation replications. The full set of simulation results are available in the unreviewed appendix.

References

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models for optimal instruments with an application to eminent domain. *Econometrica*, 80(6), 2369-2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81(2), 608-650.

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2015), Program evaluation with high-dimensional data. Forthcoming *Econometrica*.

Kozbur, D. (2015). Testing-based forward model selection. Working paper arXiv:1512.02666.

Supplemental Online Materials - Unreviewed

Appendix 1: Detailed Simulation Results

Column Definitions:
True: actual properties of simulated data
No Covariates: none included
All Included: included all covariates (when K<N)
p-hacking: Forward stepwise procedure that adds the variable that maximally decreases the *p*-value of the focal coefficient at each step until either no inclusion will improve the *p*-value or a maximum of five variables have been added.
Stepwise (FS): forward-selection (i.e. stepwise regression without the removal step) using $p<.05$
Two-step regression: Included covariates that had p<.05 in an initial multiple regression
Double Forward Regression: Approximation of double-lasso using forward regression
Double Lasso: Procedure described in the paper

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
				1. N = 60, K = 15, homoskedastic				
N Selected:	2.000	0.000	15.000	4.290	2.471	2.426	1.674	2.738
Bias:	0.001	0.001	0.003	0.255	-0.024	0.001	-0.015	-0.015
Std. Dev:	0.262	0.389	0.298	0.405	0.258	0.277	0.266	0.264
RMSE:	0.262	0.389	0.298	0.479	0.260	0.277	0.267	0.264
Size:	0.056	0.053	0.054	0.208	0.057	0.067	0.056	0.055
Power (H0):	0.479	0.252	0.398	0.757	0.455	0.482	0.447	0.454
Power Size Adj:	0.346	0.161	0.272	0.285	0.315	0.313	0.313	0.323
				2. N = 60, K = 30, homoskedastic				
N Selected:	2.000	0.000	30.000	4.955	3.136	3.145	1.590	2.566
Bias:	-0.004	-0.005	-0.005	0.326	-0.054	-0.003	-0.019	-0.026
Std. Dev:	0.264	0.390	0.380	0.473	0.259	0.295	0.271	0.269
RMSE:	0.264	0.390	0.380	0.574	0.264	0.295	0.272	0.270
Size:	0.056	0.057	0.064	0.303	0.064	0.077	0.056	0.057
Power (H0):	0.472	0.257	0.284	0.837	0.421	0.458	0.437	0.436
Power Size Adj:	0.354	0.165	0.175	0.262	0.267	0.277	0.319	0.308
				3. N = 60, K = 45, homoskedastic				
N Selected:	2.000	0.000	45.000	4.998	3.841	4.159	1.541	2.545
Bias:	-0.001	-0.003	-0.002	0.380	-0.076	-0.001	-0.017	-0.028
Std. Dev:	0.263	0.389	0.563	0.509	0.255	0.330	0.270	0.267
RMSE:	0.263	0.389	0.563	0.635	0.266	0.330	0.271	0.268
Size:	0.053	0.056	0.072	0.370	0.075	0.076	0.055	0.057
Power (H0):	0.481	0.256	0.191	0.890	0.406	0.409	0.439	0.438
Power Size Adj:	0.371	0.160	0.088	0.240	0.227	0.225	0.319	0.315

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
4. N = 60, K = 60, homoskedastic								
N Selected:	2.000	0.000	N/A	5.000	4.594	N/A	1.512	2.581
Bias:	-0.001	0.001	N/A	0.417	-0.101	N/A	-0.020	-0.032
Std. Dev:	0.266	0.390	N/A	0.544	0.253	N/A	0.274	0.270
RMSE:	0.266	0.390	N/A	0.686	0.272	N/A	0.274	0.272
Size:	0.059	0.059	N/A	0.437	0.090	N/A	0.059	0.062
Power (H0):	0.485	0.262	N/A	0.918	0.384	N/A	0.437	0.429
Power Size Adj:	0.343	0.150	N/A	0.238	0.174	N/A	0.297	0.283
5. N = 60, K = 120, homoskedastic								
N Selected:	2.000	0.000	N/A	5.000	9.018	N/A	1.420	2.676
Bias:	0.002	0.000	N/A	0.518	-0.223	N/A	-0.016	-0.039
Std. Dev:	0.266	0.390	N/A	0.613	0.237	N/A	0.279	0.274
RMSE:	0.266	0.390	N/A	0.802	0.325	N/A	0.279	0.276
Size:	0.058	0.058	N/A	0.577	0.262	N/A	0.060	0.064
Power (H0):	0.489	0.261	N/A	0.980	0.271	N/A	0.437	0.422
Power Size Adj:	0.342	0.158	N/A	0.239	0.006	N/A	0.286	0.266
6. N = 100, K = 25, homoskedastic								
N Selected:	2.000	0.000	25.000	4.920	2.961	3.138	1.934	2.653
Bias:	0.002	0.004	0.000	0.221	-0.022	0.001	-0.003	-0.008
Std. Dev:	0.202	0.302	0.231	0.280	0.200	0.209	0.203	0.202
RMSE:	0.202	0.302	0.231	0.356	0.202	0.209	0.203	0.202
Size:	0.052	0.054	0.050	0.187	0.054	0.061	0.052	0.052
Power (H0):	0.702	0.391	0.581	0.899	0.671	0.697	0.692	0.685
Power Size Adj:	0.573	0.272	0.447	0.508	0.521	0.538	0.569	0.548
7. N = 100, K = 50, homoskedastic								
N Selected:	2.000	0.000	50.000	5.000	4.056	4.412	1.893	2.549
Bias:	0.003	0.003	0.001	0.272	-0.044	0.003	-0.003	-0.009
Std. Dev:	0.201	0.300	0.286	0.299	0.198	0.215	0.202	0.203
RMSE:	0.201	0.300	0.286	0.404	0.203	0.215	0.202	0.203
Size:	0.051	0.052	0.053	0.257	0.063	0.066	0.051	0.053
Power (H0):	0.708	0.392	0.427	0.936	0.647	0.689	0.695	0.684
Power Size Adj:	0.601	0.280	0.298	0.483	0.494	0.518	0.581	0.558

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
8. N = 100, K = 75, homoskedastic								
N Selected:	2.000	0.000	75.000	5.000	5.238	5.896	1.869	2.529
Bias:	0.000	0.001	-0.006	0.305	-0.072	0.001	-0.007	-0.014
Std. Dev:	0.203	0.305	0.423	0.326	0.198	0.237	0.205	0.205
RMSE:	0.203	0.305	0.423	0.446	0.210	0.237	0.205	0.205
Size:	0.054	0.058	0.065	0.309	0.081	0.073	0.055	0.057
Power (H0):	0.696	0.392	0.246	0.947	0.604	0.632	0.680	0.674
Power Size Adj:	0.575	0.265	0.133	0.452	0.388	0.442	0.556	0.541
9. N = 100, K = 100, homoskedastic								
N Selected:	2.000	0.000	N/A	5.000	6.611	N/A	1.854	2.527
Bias:	0.000	0.000	N/A	0.330	-0.099	N/A	-0.007	-0.015
Std. Dev:	0.202	0.299	N/A	0.328	0.193	N/A	0.203	0.204
RMSE:	0.202	0.299	N/A	0.466	0.217	N/A	0.203	0.204
Size:	0.051	0.051	N/A	0.344	0.097	N/A	0.053	0.055
Power (H0):	0.699	0.390	N/A	0.964	0.575	N/A	0.680	0.677
Power Size Adj:	0.598	0.273	N/A	0.459	0.290	N/A	0.570	0.546
10. N = 100, K = 200, homoskedastic								
N Selected:	2.000	0.000	N/A	5.000	14.754	N/A	1.798	2.549
Bias:	0.002	0.000	N/A	0.395	-0.229	N/A	-0.006	-0.018
Std. Dev:	0.199	0.298	N/A	0.351	0.180	N/A	0.201	0.201
RMSE:	0.199	0.298	N/A	0.529	0.291	N/A	0.202	0.202
Size:	0.048	0.051	N/A	0.444	0.353	N/A	0.049	0.052
Power (H0):	0.700	0.388	N/A	0.984	0.391	N/A	0.676	0.664
Power Size Adj:	0.595	0.276	N/A	0.461	0.006	N/A	0.571	0.549
11. N = 60, K = 15, heteroskedastic								
N Selected:	2.000	0.000	15.000	4.590	2.455	2.253	1.568	2.725
Bias:	0.003	0.003	0.005	0.168	-0.020	0.008	-0.004	-0.006
Std. Dev:	0.303	0.430	0.345	0.427	0.310	0.323	0.333	0.327
RMSE:	0.303	0.430	0.345	0.459	0.311	0.323	0.333	0.327
Size:	0.049	0.052	0.046	0.165	0.055	0.058	0.054	0.053
Power (H0):	0.580	0.253	0.508	0.782	0.546	0.574	0.533	0.539
Power Size Adj:	0.509	0.166	0.438	0.475	0.448	0.470	0.438	0.452

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
12. N = 60, K = 30, heteroskedastic								
N Selected:	2.000	0.000	30.000	4.984	3.031	3.029	1.495	2.533
Bias:	-0.001	-0.005	0.002	0.201	-0.048	0.002	-0.006	-0.013
Std. Dev:	0.289	0.414	0.404	0.460	0.296	0.317	0.317	0.315
RMSE:	0.289	0.414	0.404	0.502	0.300	0.317	0.317	0.315
Size:	0.047	0.055	0.052	0.232	0.060	0.065	0.053	0.050
Power (H0):	0.605	0.261	0.406	0.866	0.535	0.580	0.551	0.549
Power Size Adj:	0.533	0.173	0.320	0.487	0.440	0.470	0.473	0.466
13. N = 60, K = 45, heteroskedastic								
N Selected:	2.000	0.000	45.000	4.999	3.660	4.376	1.430	2.443
Bias:	0.002	0.008	-0.002	0.245	-0.070	0.010	-0.004	-0.013
Std. Dev:	0.293	0.410	0.620	0.480	0.300	0.351	0.321	0.328
RMSE:	0.293	0.410	0.620	0.539	0.307	0.351	0.321	0.328
Size:	0.049	0.048	0.076	0.287	0.073	0.073	0.053	0.055
Power (H0):	0.607	0.266	0.250	0.901	0.513	0.530	0.544	0.534
Power Size Adj:	0.539	0.186	0.121	0.496	0.372	0.392	0.459	0.450
14. N = 60, K = 60, heteroskedastic								
N Selected:	2.000	0.000	N/A	4.999	4.357	N/A	1.392	2.377
Bias:	-0.003	-0.005	N/A	0.257	-0.100	N/A	-0.009	-0.019
Std. Dev:	0.287	0.411	N/A	0.504	0.292	N/A	0.322	0.329
RMSE:	0.287	0.411	N/A	0.566	0.308	N/A	0.322	0.330
Size:	0.044	0.051	N/A	0.314	0.092	N/A	0.050	0.052
Power (H0):	0.598	0.254	N/A	0.919	0.471	N/A	0.531	0.511
Power Size Adj:	0.529	0.169	N/A	0.475	0.298	N/A	0.454	0.416
15. N = 60, K = 120, heteroskedastic								
N Selected:	2.000	0.000	N/A	5.000	8.691	N/A	1.302	2.318
Bias:	-0.001	-0.007	N/A	0.323	-0.215	N/A	-0.008	-0.023
Std. Dev:	0.289	0.407	N/A	0.567	0.293	N/A	0.325	0.334
RMSE:	0.289	0.407	N/A	0.652	0.364	N/A	0.325	0.335
Size:	0.048	0.048	N/A	0.443	0.279	N/A	0.051	0.053
Power (H0):	0.599	0.254	N/A	0.963	0.340	N/A	0.525	0.494
Power Size Adj:	0.536	0.171	N/A	0.432	0.007	N/A	0.446	0.411

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
16. N = 100, K = 25, heteroskedastic								
N Selected:	2.000	0.000	25.000	4.976	2.884	2.773	1.745	2.762
Bias:	0.005	0.004	0.004	0.142	-0.018	0.006	0.003	-0.001
Std. Dev:	0.226	0.322	0.259	0.280	0.227	0.234	0.240	0.237
RMSE:	0.226	0.322	0.259	0.314	0.228	0.234	0.240	0.237
Size:	0.047	0.052	0.045	0.133	0.052	0.057	0.051	0.048
Power (H0):	0.722	0.387	0.644	0.885	0.690	0.713	0.694	0.694
Power Size Adj:	0.664	0.282	0.585	0.685	0.616	0.635	0.622	0.632
17. N = 100, K = 50, heteroskedastic								
N Selected:	2.000	0.000	50.000	5.000	3.846	3.933	1.681	2.555
Bias:	0.002	0.004	0.004	0.177	-0.042	0.003	0.000	-0.006
Std. Dev:	0.228	0.322	0.325	0.294	0.230	0.242	0.245	0.245
RMSE:	0.228	0.322	0.325	0.343	0.234	0.242	0.245	0.246
Size:	0.047	0.047	0.043	0.178	0.060	0.059	0.052	0.051
Power (H0):	0.723	0.389	0.504	0.925	0.663	0.711	0.692	0.685
Power Size Adj:	0.668	0.306	0.441	0.716	0.567	0.623	0.618	0.607
18. N = 100, K = 75, heteroskedastic								
N Selected:	2.000	0.000	75.000	5.000	4.951	5.788	1.645	2.571
Bias:	0.002	0.003	0.007	0.196	-0.069	0.007	-0.001	-0.006
Std. Dev:	0.227	0.320	0.466	0.309	0.229	0.258	0.244	0.244
RMSE:	0.227	0.320	0.466	0.366	0.240	0.258	0.244	0.244
Size:	0.046	0.053	0.056	0.211	0.072	0.060	0.051	0.051
Power (H0):	0.726	0.385	0.309	0.945	0.627	0.677	0.684	0.676
Power Size Adj:	0.669	0.274	0.200	0.707	0.502	0.578	0.614	0.604
19. N = 100, K = 100, heteroskedastic								
N Selected:	2.000	0.000	N/A	5.000	6.183	N/A	1.616	2.510
Bias:	0.003	0.005	N/A	0.213	-0.097	N/A	0.000	-0.007
Std. Dev:	0.226	0.318	N/A	0.316	0.226	N/A	0.242	0.244
RMSE:	0.226	0.318	N/A	0.380	0.246	N/A	0.242	0.244
Size:	0.046	0.053	N/A	0.238	0.095	N/A	0.049	0.050
Power (H0):	0.726	0.393	N/A	0.952	0.594	N/A	0.686	0.676
Power Size Adj:	0.671	0.276	N/A	0.696	0.406	N/A	0.611	0.599

Covariate Selection Method	True	No Covariates	All Included	p-hacking	Stepwise (FS)	Two-step regression	Double forward regression	Double Lasso
20. N = 100, K = 200, heteroskedastic								
N Selected:	2.000	0.000	N/A	5.000	13.563	N/A	1.540	2.441
Bias:	0.000	-0.001	N/A	0.246	-0.223	N/A	0.000	-0.009
Std. Dev:	0.226	0.323	N/A	0.354	0.229	N/A	0.247	0.251
RMSE:	0.226	0.323	N/A	0.431	0.320	N/A	0.247	0.251
Size:	0.047	0.054	N/A	0.315	0.334	N/A	0.051	0.052
Power (H0):	0.725	0.385	N/A	0.975	0.420	N/A	0.679	0.664
Power Size Adj:	0.667	0.260	N/A	0.685	0.007	N/A	0.606	0.590

Appendix 2: Additional Details for Analyses in STATA

Analysis 1: Correlational Analysis of Parents' Life Satisfaction

We use a slightly different sample size, because we exclude cases with missing values on the additional covariates. The results reported in Table 1 of Bhargava et al (2014) were largely the same after these exclusions.

Table A1: Replication of Table 1 in Bhargava et al (2014).

Model and predictor	Satisfaction	Happiness	Meaning
	N=5213	N=5178	N=5195
No controls			
Parenthood	β = .195, p < .001	β = .053, p = .005	β = .089, p < .001
Controls: marital status			
Parenthood	β = -.113, p = .071	β = -.040, p = .058	β = .123, p < .001
Controls: marital status & age			
Parenthood	β = -.191, p = .007	β = -.043, p = .068	β = .103, p = .001
Controls: marital status, age & gender			
Parenthood	β = -.197, p = .006	β = -.049, p = .038	β = .079, p = .009
Controls: marital status, age, gender & income			
Parenthood	β = -.118, p = .098	β = -.027, p = .253	β = .084, p = .006

In the paper, we report the results of the double-lasso analysis for life satisfaction, and we note that only using covariates identified as predictors of the dependent variable yields different results, shown in Table A2. We also re-ran the original analysis excluding income and employment variables most likely to be impacted by parenthood, and replicate our findings (Table A3).

Table A2: Regression of parenthood on life satisfaction, with DV-predictor covariates only.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	7.193	0.068	106.27	.000	7.061	7.326
Parent	-0.127	0.065	-1.95	.051	-0.255	0.000
Main effect covariates:						
Income (3 point scale)	0.398	0.105	3.78	.000	0.191	0.605
Interaction covariates:						
Married x Age	1.164	0.139	8.38	.000	0.891	1.436
Married x Income rating (3 point)	-0.005	0.177	-0.03	.975	-0.352	0.341
Married x Income rating (11 point)	0.412	0.191	2.15	.031	0.037	0.787

Table A3: Regression of parenthood on life satisfaction, with non-employment demographic covariates only.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	7.095	0.103	68.95	.000	6.893	7.296
Parent	-0.221	0.071	-3.09	.002	-0.361	-0.081
Main effect covariates:						
Married (including living together as married)	0.617	0.120	5.15	.000	0.382	0.852
Age	0.773	0.231	3.35	.001	0.321	1.226
Age=18	0.294	0.207	1.42	.156	-0.112	0.700
Age=19	0.077	0.209	0.37	.713	-0.333	0.487
Age=20	0.463	0.189	2.46	.014	0.093	0.833
Age=21	0.094	0.172	0.55	.582	-0.242	0.431
Age=22	0.462	0.167	2.77	.006	0.135	0.79
Age=23	0.125	0.174	0.72	.473	-0.216	0.466
Gender (Male)	-0.047	0.052	-0.91	.364	-0.148	0.054
Interaction covariates:						
Married x Age	0.553	0.344	1.61	.108	-0.121	1.226
Married x Age to fourth power	-0.489	0.589	-0.83	.406	-1.643	0.665

Using double-lasso selected covariates, we find that marital status and age moderates the effect of being a parent on life satisfaction, but gender does not (Table A4).

Table A4: Regression effect of parenthood on happiness, with double-lasso selected covariates.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	4.832	0.673	7.18	.000	3.512	6.152
Parent	-0.897	0.23	-3.89	.000	-1.348	-0.445
Interactions with parenthood:						
Married x Parent	0.362	0.148	2.44	.015	0.071	0.653
Age x Parent	3.281	1.265	2.59	.009	0.802	5.76
Age Squared x Parent	-3.107	1.549	-2.01	.045	-6.144	-0.069
Male x Parent	0.000	0.141	0.00	.998	-0.276	0.277

Analysis controls for 47 additional covariates (not shown)

Lastly, we conducted the same analyses for the effect of parenthood on happiness (Table A5) and on meaning in life (Table A6).

Table A5: Regression effect of parenthood on happiness, with double-lasso selected covariates.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	3.178	0.034	93.66	.000	3.111	3.244
Parent	-0.050	0.024	-2.12	.034	-0.096	-0.004
Main effect covariates:						
Married (including living together as married)	0.043	0.048	0.89	.373	-0.052	0.138
Age	0.082	0.073	1.13	.259	-0.061	0.225
Age=18	0.091	0.068	1.34	.182	-0.043	0.224
Age=19	0.076	0.067	1.12	.261	-0.056	0.208
Age=20	0.128	0.057	2.23	.026	0.016	0.240
Age=21	-0.018	0.061	-0.30	.764	-0.138	0.101
Age=22	0.101	0.060	1.67	.095	-0.018	0.220
Age=23	0.048	0.059	0.81	.418	-0.068	0.163
Gender (Male)	-0.049	0.021	-2.38	.017	-0.089	-0.009
Employment: Housewife	0.031	0.041	0.75	.454	-0.050	0.112
Chief wage earner	0.006	0.023	0.24	.807	-0.039	0.050
Interaction covariates:						
Married x Age	0.122	0.119	1.03	.304	-0.111	0.355
Married x Age to fourth power	0.200	0.214	0.94	.349	-0.219	0.620
Married x Income rating (3 point)	0.010	0.051	0.19	.846	-0.090	0.109
Married x Income rating (11 point)	0.254	0.082	3.11	.002	0.094	0.414
Employment: Student x Male	0.132	0.095	1.40	.162	-0.053	0.318

Table A6: Regression effect of parenthood on meaning in life, with double-lasso selected covariates.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	3.417	0.041	82.36	.000	3.336	3.498
Parent	0.094	0.030	3.08	.002	0.034	0.154
Main effect covariates:						
Married (including living together as married)	-0.032	0.038	-0.83	.407	-0.107	0.043
Age	0.010	0.081	0.13	.898	-0.149	0.170
Age=18	-0.116	0.080	-1.45	.147	-0.273	0.041
Age=19	-0.018	0.084	-0.22	.829	-0.183	0.147
Age=20	-0.002	0.078	-0.02	.983	-0.155	0.152
Age=21	-0.073	0.078	-0.93	.350	-0.226	0.080
Age=22	-0.172	0.086	-2.02	.044	-0.340	-0.005
Age=23	-0.077	0.078	-0.99	.322	-0.230	0.076
Gender (Male)	-0.153	0.026	-5.86	.000	-0.204	-0.102
Employment: Housewife	-0.105	0.051	-2.04	.041	-0.205	-0.004
Chief wage earner	-0.043	0.029	-1.49	.135	-0.099	0.013
Interaction covariates:						
Married x Age to fourth power	-0.080	0.224	-0.36	.721	-0.519	0.359
Married x Income rating (3 point)	-0.066	0.039	-1.71	.088	-0.142	0.010
Employment: Student x Male	0.068	0.120	0.57	.569	-0.167	0.304

Analysis 2: Mediation Analysis of Conservative Happiness

In the paper, we approximated the analyses in Napier and Jost (2008), by excluding cases with missing values on the variables used in their paper, and setting missing values for other variables to the median value (for ordinal variables) or to the mean value (for continuous variables). This results in 1192 cases (similar to the 1142 reported in their paper). As a result, when we re-run the analyses for the table in their paper, we find results that are similar but not exactly the same.

Table A7: Replicating analysis from Napier and Jost (2008).

Predictor	Step 1	Step 2	Step 3	Step 4	Step 5
Political conservativism	0.263 (.064)****	0.191 (0.064)***	0.161 (0.064)**	0.161 (0.064)**	0.114 (0.074)
Income (household)		0.011 (0.011)	0.013 (0.011)	0.013 (0.011)	0.013 (0.011)
Education		0.085 (0.025)***	0.078 (0.025)***	0.074 (0.026)***	0.076 (0.025)***
Sex		0.046 (0.073)	0.026 (0.073)	0.033 (0.075)	0.035 (0.075)
Age		0.059 (0.041)	0.045 (0.041)	0.047 (0.042)	0.038 (0.042)
Age squared		0.188 (0.041)****	0.185 (0.041)****	0.185 (0.041)****	0.182 (0.041)****
Marital status		0.361 (0.081)****	0.350 (0.081)****	0.351 (0.081)****	0.341 (0.081)****
Employment status		-0.006 (0.091)	-0.014 (0.091)	-0.011 (0.091)	-0.010 (0.091)
Church attendance			0.074 (0.03)**	0.074 (0.03)**	0.077 (0.031)**
Need for cognition				0.068 (0.119)	0.066 (0.118)
Rationalization of inequality					0.080 (0.054)

In an additional analysis, we only excluded respondents who were missing values for the dependent variable (life satisfaction), independent variable (political orientation) or the proposed mediator (rationalization of inequality scale), yielding 1364 cases. In this analysis, the evidence for mediation is a bit weaker. Conservatives were happier (β=.277, t(1362)=4.66, p<.001), and the difference was partially mediated by the rationalization of inequality scale (indirect β=.058, bootstrap CI=[.000,.016], p=.0497).

We conducted a second double-lasso analysis, and identified seven covariates, the same four as in the prior analysis (church attendance and dummy codes for employment, being married and African-American race), as well as three others (being multiracial, not attending denominational church services, and attending Protestant services). Controlling for these covariates, the effect of political orientation on happiness was marginally significant. While the coefficient of political orientation was reduced and not significant when rationalization of

inequality is included, the mediation is not significant (indirect β=.038, bootstrap CI=[-.014,.090], p=.148).

Table A8: Replicating analysis from Napier and Jost (2008), full sample.

Predictor	Step 1	Step 2	Step 3	Step 4	Step 5
Political conservativism	0.277 (.059)****	0.208 (0.059)****	0.179 (0.059)***	0.18 (0.059)***	0.137 (0.068)**
Income (household)		0.012 (0.011)	0.013 (0.01)	0.013 (0.01)	0.013 (0.01)
Education		0.079 (0.023)***	0.073 (0.023)***	0.07 (0.023)***	0.071 (0.023)***
Sex		0.059 (0.068)	0.039 (0.068)	0.045 (0.069)	0.047 (0.069)
Age		0.072 (0.038)*	0.057 (0.038)	0.059 (0.039)	0.05 (0.039)
Age squared		0.163 (0.037)****	0.159 (0.037)****	0.159 (0.037)****	0.157 (0.037)****
Marital status		0.359 (0.074)****	0.346 (0.074)****	0.347 (0.074)****	0.338 (0.074)****
Employment status		0.055 (0.086)	0.048 (0.085)	0.05 (0.086)	0.053 (0.086)
Church attendance			0.074 (0.028)***	0.074 (0.028)***	0.077 (0.028)***
Need for cognition				0.051 (0.111)	0.05 (0.11)
Rationalization of inequality					0.073 (0.051)

Table A9: Regressions of political orientation on happiness (full sample).

	Model 2a		Model 2b	
Variable	β	p	β	p
Primary variables:				
Constant	2.834	.000	2.649	.000
Political Orientation	0.107	.085	0.069	.311
Mediator:				
Rationalization of inequality			0.072	.142
Church attendance	0.072	.017	0.073	.016
Married	0.239	.000	0.231	.001
Not unemployed or disabled	-0.660	.000	-0.670	.000
Black	-0.297	.021	-0.269	.036
Multiracial	-0.122	.059	-0.095	.150
No denominational church services	-0.148	.071	-0.148	.071
Attends Protestant services	0.176	.033	0.166	.045

Appendix 3: Replication Analyses in R

Analysis 1: Correlational Analysis of Parents' Life Satisfaction

We redid our analysis using double lasso via a modification of the *glmnet* package in R, testing the effect of parenthood on life satisfaction (Table A10), happiness (Table A11) and meaning in life (Table A12).

Table A10: Regression effect of parenthood on life satisfaction, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	6.809	0.118	57.71	.000
Parent	-0.172	0.069	-2.49	.013
Main effect covariates:				
Married (including living together as married)	0.452	0.146	3.09	.002
Income rating (3 point scale)	0.493	0.114	4.34	.000
Age	0.927	0.211	4.40	.000
Age=18	0.375	0.194	1.94	.053
Age=19	0.200	0.194	1.03	.303
Age=20	0.568	0.188	3.03	.002
Age=21	0.214	0.176	1.21	.225
Age=22	0.611	0.179	3.41	.001
Age=23	0.225	0.180	1.25	.212
Age=24	0.443	0.179	2.47	.013
Gender (Male)	-0.123	0.061	-2.03	.042
Employment: Housewife	0.021	0.115	0.18	.857
Employment: Student	0.054	0.199	0.27	.785
Employment: Unemployed	-0.546	0.109	-5.02	.000
Chief wage earner	0.106	0.067	1.58	.114
Interaction covariates:				
Married x Age	0.260	0.352	0.74	.460
Married x Age to fourth power	0.227	0.652	0.35	.728
Married x Income rating (3 point)	-0.041	0.187	-0.22	.826
Married x Income rating (11 point)	0.283	0.239	1.18	.236

Table A11: Regression effect of parenthood on happiness, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	3.171	0.035	91.52	.000
Parent	-0.049	0.023	-2.10	.036
Main effect covariates:				
Married (including living together as married)	0.038	0.046	0.84	.401
Age	0.094	0.071	1.32	.187
Age=18	0.115	0.066	1.75	.080
Age=19	0.093	0.066	1.41	.160
Age=20	0.139	0.064	2.18	.029
Age=21	-0.010	0.060	-0.17	.866
Age=22	0.107	0.061	1.75	.080
Age=23	0.053	0.061	0.87	.385
Age=24	0.066	0.061	1.08	.280
Gender (Male)	-0.046	0.020	-2.29	.022
Employment: Housewife	0.031	0.039	0.79	.428
Employment: Student	-0.011	0.068	-0.16	.874
Chief wage earner	0.004	0.022	0.19	.846
Interaction covariates:				
Married x Age	0.134	0.119	1.13	.259
Married x Age to fourth power	0.166	0.222	0.75	.454
Married x Income rating (11 point)	0.268	0.047	5.67	.000

Table A12: Regression effect of parenthood on meaning in life, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	3.417	0.043	78.81	.000
Parent	0.094	0.030	3.10	.002
Main effect covariates:				
Married (including living together as married)	-0.072	0.030	-2.42	.015
Age	-0.001	0.083	-0.01	.991
Age=18	-0.137	0.085	-1.61	.107
Age=19	-0.031	0.085	-0.36	.718
Age=20	-0.019	0.082	-0.24	.812
Age=21	-0.080	0.077	-1.04	.297
Age=22	-0.173	0.079	-2.20	.028
Age=23	-0.078	0.078	-0.99	.322
Age=24	-0.039	0.078	-0.50	.616
Gender (Male)	-0.153	0.026	-5.91	.000
Employment: Housewife	-0.093	0.050	-1.87	.062
Employment: Student	0.164	0.087	1.89	.058
Chief wage earner	-0.039	0.028	-1.37	.171
Interaction covariates:				
Married x Age to fourth power	-0.013	0.203	-0.06	.949

Analysis 2: Mediation Analysis of Conservative Happiness

The double lasso using modified *glmnet* in R also revealed no significant mediation, after controlling for selected covariates. We identified five covariates, the same four as in the STATA analysis (church attendance and dummy codes for employment, being married and African-American race), as well as being retired. Controlling for these covariates, the effect of political orientation on happiness is significant. While the coefficient of political orientation is reduced and no longer significant when rationalization of inequality is included, the mediation is not significant (indirect β=.034, bootstrap CI=[-.022,.091], p=.236).

Table A13: Regressions of political orientation on happiness.

	Model 2a		Model 2b	
Variable	β	p	β	p
Primary variables:				
Constant	2.090	.000	1.891	.000
Political Orientation	0.131	.042	0.093	.176
Mediator:				
Rationalization of inequality			0.072	.110
Covariates:				
Church attendance	0.087	.001	0.089	.001
Married	0.233	.000	0.225	.001
Retired	0.279	.001	0.260	.003
Not unemployed or disabled	0.627	.000	0.641	.000
Black	-0.253	.023	-0.225	.000

Analysis 3: Spurious Experiment on Chronological Rejuvenation in R

As in the STATA analysis, the double lasso using modified *glmnet* in R revealed insufficient empirical support to include any of the potential covariates in the regression. The resulting single-variable regression revealed no significant effect of the experimental manipulation (β = -305.3, t(18)=1.00, p=.329).

Analysis 4: Experiment on Suggested Defaults in Donation in R

The double-lasso regression using modified *glmnet* in R yielded similar results to the STATA analysis. Two covariates were identified (prior donation amount and interaction between prior amount and lifetime giving) in the lasso step predicting donation amount. No covariates were selected in the step predicting experimental condition, revealing no evidence of failed randomization based on the covariates. A linear regression including the identified covariates and orthogonal experimental conditions, confirmed a significant difference between the low

recommendation condition and control condition in log donation amount (β= -.365, t(67)=3.03, p=.004, Table A14).

Table A14: Regression effect of randomized suggestion level on donation amount among donors, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	1.055	0.300	3.517	.001
Low Amount Recommended	-0.369	0.119	-3.094	.003
Medium Amount Recommended	-0.128	0.135	-0.946	.347
High Amount Recommended	0.032	0.143	0.226	.822
Other randomized manipulations:				
Five earmarking options shown	-0.192	0.100	-1.931	.058
Previous donation shown	0.006	0.096	0.058	.954
Selected covariates:				
Log of prior year donation	0.690	0.097	7.079	.000
Log of prior year donation x Log of lifetime giving	0.018	0.007	2.419	.018

Appendix 4: Replication Analysis in SPSS

Analysis 1: Correlational Analysis of Parents' Life Satisfaction

We redid our analysis in SPSS, using the double forward regression approach to approximating the double lasso, testing the effect of parenthood on life satisfaction (Table A15), happiness (Table A16) and meaning in life (Table A17).

Table A15: Regression effect of parenthood on life satisfaction, with double-lasso selected covariates.

Variable	β	SE	*t*	*p*
Primary variables:				
Constant	7.497	0.140	53.72	.000
Parent	-0.143	0.070	-2.05	.040
Main effect covariates:				
Married (including living together as married)	0.331	0.186	1.78	.075
Income rating (3 point scale)	0.424	0.123	3.45	.001
Income rating (11 point scale)	0.157	0.225	0.70	.485
Age	-4.859	1.073	-4.53	.000
Age squared	13.313	2.868	4.64	.000
Age cubed	-8.575	2.220	-3.86	.000
Gender (Male)	-0.222	0.084	-2.66	.008
Employment: Housewife	0.016	0.115	0.14	.887
Employment: Unemployed	-0.550	0.107	-5.12	.000
Interaction covariates:				
Married x Age	0.803	0.889	0.90	.366
Married x Age squared	-0.884	1.080	-0.82	.413
Married x Income rating (11 point)	0.180	0.215	0.84	.401
Married x Male	0.235	0.107	2.21	.027

Table A16: Regression effect of parenthood on happiness, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	3.342	0.050	66.18	.000
Parent	-0.036	0.024	-1.51	.130
Main effect covariates:				
Married (including living together as married)	0.054	0.063	0.86	.391
Income (8 point scale)	0.000	0.000	-5.01	.000
Income rating (11 point scale)	0.144	0.055	2.61	.009
Age	-1.057	0.363	-2.91	.004
Age squared	2.140	0.971	2.20	.028
Age cubed	-1.053	0.752	-1.40	.162
Gender (Male)	-0.069	0.028	-2.43	.015
Employment: Housewife	0.016	0.039	0.40	.688
Interaction covariates:				
Married x Age	0.531	0.301	1.77	.078
Married x Age squared	-0.469	0.366	-1.28	.200
Married x Income rating (11 point)	0.081	0.073	1.11	.267
Married x Male	0.030	0.036	0.84	.403

Table A17: Regression effect of parenthood on meaning in life, with double-lasso selected covariates.

Variable	β	SE	t	p
Primary variables:				
Constant	3.360	0.048	69.29	.000
Parent	0.081	0.031	2.66	.008
Main effect covariates:				
Married (including living together as married)	-0.170	0.052	-3.29	.001
Income rating (11 point scale)	-0.027	0.046	-0.59	.554
Age	0.840	0.442	1.90	.058
Age squared	-2.043	1.245	-1.64	.101
Age cubed	1.176	0.974	1.21	.227
Gender (Male)	-0.203	0.037	-5.55	.000
Employment: Housewife	-0.074	0.050	-1.46	.144
Interaction covariates:				
Married x Age	0.246	0.111	2.22	.027
Married x Male	0.047	0.047	1.01	.314

Analysis 2: Mediation Analysis of Conservative Happiness

The double forward regression in SPSS also revealed no significant mediation, after controlling for selected covariates. We identified six covariates, the same four as in the STATA analysis (church attendance and dummy codes for employment, being married and African-American race), as well as age squared and household union membership. Controlling for these covariates, the effect of political orientation on happiness is significant (Model 1, Table 3). While the coefficient of political orientation is reduced and no longer significant when rationalization of inequality is included, the mediation is not significant (indirect β=.034, bootstrap CI=[-.023,.091], p=.241).

Table A18: Regressions of political orientation on happiness.

	Model 1		Model 2	
Variable	β	p	β	p
Primary variables:				
Constant	1.938	.000	1.741	.000
Political Orientation	0.135	.037	0.097	.160
Mediator:				
Rationalization of inequality			0.073	.099
Covariates:				
Church attendance	0.090	.001	0.091	.001
Married	0.330	.000	0.318	.000
Age squared	0.141	.000	0.137	.000
Union household	0.004	.964	0.006	.944
Not unemployed or disabled	0.607	.000	0.621	.000
Black	-0.248	.025	-0.219	.050

Analysis 3: Spurious Experiment on Chronological Rejuvenation

As in the STATA analysis, the double forward regression in SPSS revealed insufficient empirical support to include any of the potential covariates in the regression. The resulting single-variable regression revealed no significant effect of the experimental manipulation (β = -305.3, t(18)=1.00, p=.329).

Analysis 4: Experiment on Suggested Defaults in Donation in SPSS

The double-lasso regression using double forward regression in SPSS yielded similar results to the STATA analysis. One covariate was identified (prior donation amount) in the lasso step predicting donation amount. No covariates were selected in the step predicting experimental condition, revealing no evidence of failed randomization based on the covariates. A linear regression including the identified covariate and orthogonal experimental conditions, confirmed

a significant difference between the low recommendation condition and control condition in log donation amount (β= -.338, t(66)=2.76, p=.007, Table Azz).

Table A19: Regression effect of randomized suggestion level on donation amount among donors, with double-lasso selected covariates.

Variable	β	SE	*t*	*p*
Primary variables:				
Constant	.642	.255	2.516	.014
Low Amount Recommended	-.338	.123	-2.760	.007
Medium Amount Recommended	-.144	.140	-1.029	.307
High Amount Recommended	.016	.148	.109	.913
Other randomized manipulations:				
Five earmarking options shown	-.178	.103	-1.734	.087
Previous donation shown	.013	.099	.131	.896
Selected covariates:				
Log of prior year donation	.903	.043	21.196	.000

Appendix 5: Sample Code

Analysis 1: STATA

```
** Replication of results in Kassam et al. 2013 Table 1 Column 1
** Open data file 'wvssatisfy.txt'

***********************************
*      ANALYSIS 1 -- satisfaction    *
***********************************

** REPLICATE RESULTS from table in BKL (N=5213 because of missing values)
** Results will differ slightly because of different sample size
* Row 2
reg satisfy kid, robust

* Row 3
reg satisfy ms_marry2 kid, robust

* Row 4
reg satisfy ms_marry2 kid age*, robust

* Row 5
reg satisfy ms_marry2 kid age* male, robust

* Row 6
reg satisfy ms_marry2 kid age* i.inc1 male, robust

* VARIABLE SELECTION FOR MAIN EFFECT COVARIATES
* Select variables that predict the outcome
lassoShooting satisfy male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2
inc22 inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* ,
lasiter(100) verbose(0) fdisplay(0)
local satisfySel `r(selected)'
di "`satisfySel'"

* Select variables that predict the treatment
lassoShooting kid male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2 inc22
inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* , lasiter(100)
verbose(0) fdisplay(0)
local kidSel `r(selected)'
di "`kidSel'"

* Get union of selected instruments
local satisfyDS: list satisfySel | kidSel

* Regress outcome on treatment(s) and relevant controls - baseline
reg satisfy kid `satisfyDS' , robust

* regression without X predictors
reg satisfy kid `satisfySel' , robust

* regression without income and employment covariates.
reg satisfy kid marryxage male ms_marry2 ages marryxage4 _iage_18 _iage_19
_iage_20 _iage_21 _iage_22 _iage_23 , robust
```

```
* VARIABLE SELECTION FOR PRE-SPECIFIED MODERATORS
lassoShooting ki_marry male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2
inc22 inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* ,
lasiter(100) verbose(0) fdisplay(0)
local ki_marrySel `r(selected)'
di "`ki_marrySel'"
lassoShooting ki_age male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2
inc22 inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* ,
lasiter(100) verbose(0) fdisplay(0)
local ki_ageSel `r(selected)'
di "`ki_ageSel'"
lassoShooting ki_age2 male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2
inc22 inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* ,
lasiter(100) verbose(0) fdisplay(0)
local ki_age2Sel `r(selected)'
di "`ki_age2Sel'"
lassoShooting ki_male male ms_marry2 malexmarry ages ages2 ages3 ages4 inc2
inc22 inc3 inc32 malexage* malexinc* marryx* inc2x* inc22x* _i* ,
lasiter(100) verbose(0) fdisplay(0)
local ki_maleSel `r(selected)'
di "`ki_maleSel'"

local satisfyDSm : list satisfyDS | ki_marrySel
local satisfyDSm : list satisfyDSm | ki_ageSel
local satisfyDSm : list satisfyDSm | ki_age2Sel
local satisfyDSm : list satisfyDSm | ki_maleSel

* Regress outcome on treatment(s) and relevant controls - moderation
reg satisfy kid ki_marry ki_age ki_age2 ki_male `satisfyDSm' , robust
```

Analysis 1: R

```
library(glmnet)

####################################
#      ANALYSIS 1 -- SATISFACTION   #
####################################
rm(list=ls())
study1 = read.csv("wvssatisfy.txt")  # read csv file

## REPLICATE RESULTS from table in BKL (N=5213 because of missing values)
# Row 2
fitr <- lm(satisfy ~ kid, data=study1)
summary(fitr) # show results

# Row 3
regvar <- as.matrix(subset(study1,select = c(kid, ms_marry2)))
fitr <- lm(satisfy ~ regvar, data=study1)
summary(fitr) # show results

# Row 4
regvar <- as.matrix(subset(study1,select = c(kid, ms_marry2, ages, ages2,
ages3, ages4)))
fitr <- lm(satisfy ~ regvar, data=study1)
summary(fitr) # show results

# Row 5
regvar <- as.matrix(subset(study1,select = c(kid, ms_marry2, ages, ages2,
ages3, ages4, male)))
fitr <- lm(satisfy ~ regvar, data=study1)
summary(fitr) # show results

# Row 6
reg1 <- as.matrix(subset(study1,select = c(kid, ms_marry2, ages, ages2,
ages3, ages4, male)))
reg2 <-as.matrix(subset(study1[,115:154]))
regvar <- cbind(reg1,reg2)
fitr <- lm(satisfy ~ regvar, data=study1)
summary(fitr) # show results


# VARIABLE SELECTION FOR MAIN EFFECT COVARIATES
testvar <-as.matrix(subset(study1[,3:526]))
satisfy <- as.matrix(study1$satisfy)
kid <- as.matrix(study1$kid)
kidXmarry <- as.matrix(study1$ki_marry)
kidXage <- as.matrix(study1$ki_age)
kidXage2 <- as.matrix(study1$ki_age2)
kidXmale <- as.matrix(study1$ki_male)

# STEP 1: select variables that predict outcomes
n=nrow(testvar)
p=ncol(testvar)
sda = sd(residuals(lm(satisfy ~ kid, data=study1)))
lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
```

```
k = 1
while(k < 15){
  fitY = glmnet(testvar, satisfy, lambda=lambda1)
  ba = coef(fitY, s = lambda1)
  ea = satisfy-predict(fitY,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

# STEP 2: select variables that predict treatment
n=nrow(testvar)
p=ncol(testvar)
sd1=sd(kid)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT1 = glmnet(testvar, kid, lambda=lambda1)
  ba = coef(fitT1, s = lambda1)
  ea = kid-predict(fitT1,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba


# STEP 3: linear regression with both sets of variables
use1 = union(which(abs(fitY$beta)>.0001),which(abs(fitT1$beta) > .0001))
X = cbind(kid,testvar)
use = c(1,use1+1)
fitr <- lm(satisfy ~ X[,use], data=study1)
summary(fitr) # show results

# Regression without X predictors
use1 = which(abs(fitY$beta)>.0001)
X = cbind(kid,testvar)
use = c(1,use1+1)
fitr <- lm(satisfy ~ X[,use], data=study1)
summary(fitr) # show results

# VARIABLE SELECTION FOR PRE-SPECIFIED MODERATORS
# STEP 4: select variables that predict moderators
n=nrow(testvar)
p=ncol(testvar)
sd1=sd(kidXmarry)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT2 = glmnet(testvar, kidXmarry, lambda=lambda1)
  ba = coef(fitT2, s = lambda1)
  ea = kidXmarry-predict(fitT2,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
```

```
  k = k+1
}
ba

sd1=sd(kidXage)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT3 = glmnet(testvar, kidXage, lambda=lambda1)
  ba = coef(fitT3, s = lambda1)
  ea = kidXage-predict(fitT3,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

sd1=sd(kidXage2)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT4 = glmnet(testvar, kidXage2, lambda=lambda1)
  ba = coef(fitT4, s = lambda1)
  ea = kidXmarry-predict(fitT4,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

sd1=sd(kidXmale)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT5 = glmnet(testvar, kidXmale, lambda=lambda1)
  ba = coef(fitT5, s = lambda1)
  ea = kidXmale-predict(fitT5,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba
use2 = union(which(abs(fitT2$beta)>.0001),which(abs(fitT3$beta) > .0001))
use3 = union(which(abs(fitT4$beta)>.0001),which(abs(fitT5$beta) > .0001))
use12 = union (use1, use2)
use = union (use12, use3)
X = cbind(kid,kidXmarry, kidXage, kidXage2, kidXmale, testvar)
use = c(1,2,3,4,5,use+5)
summary(use)

fitr <- lm(satisfy ~ X[,use], data=study1)
summary(fitr) # show results
```

Analysis 1: SPSS

```
* Import data file: 'Satisfy.sav'

*------------------------------------------------------------------------------
*      ANALYSIS 1 - life satisfaction    *
*------------------------------------------------------------------------------

** REPLICATE RESULTS from table in BKL (N=5213 because of missing values)
** Results will differ slightly because of different sample size

* Row 2.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid.

* Row 3.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ms_marry2.

* Row 4.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ms_marry2 kid ages ages2 ages3 ages4.

* Row 5.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ms_marry2 kid ages ages2 ages3 ages4
male.

* Row 6.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ms_marry2 kid ages ages2 ages3 ages4
male Int_Iinc1_2 to Int_Iinc1_840031.

* VARIABLE SELECTION FOR MAIN EFFECT COVARIATES.
* APPROXIMATES DOUBLE LASSO USING FORWARD REGRESSION
*------------------------------------------------------------------------------
* Select variables that predict the outcome
* p=524 and n=5213.
* cutoff = .1 / [log(n) * 2p] = .000011.

* Select variables that predict the outcome.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT satisfy
  /METHOD=FORWARD male to Int_IincXms__840031 .
*Selected: marryXinc3 marryXage2 Int_Iemp_7 inc2.

* Select variables that predict the treatment.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT kid
```

```
  /METHOD=FORWARD male to Int_IincXms__840031 .
* Selected ages ms_marry2 ages2 marryXage male maleXmarry inc3 ages3
Int_Iemp_5.

* Regress outcome on treatment(s) and relevant controls - baseline.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ms_marry2 ages ages2 ages3 inc2 inc3
male  Int_Iemp_5 Int_Iemp_7 marryXinc3 marryXage2 marryXage maleXmarry .

* VARIABLE SELECTION FOR PRE-SPECIFIED MODERATORS.
* Select variables that predict the moderators.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT ki_marry
  /METHOD=FORWARD  male to Int_IincXms__840031 .
* Selected ages ms_marry2 marryXage marryXage2 marryXage3 Int_Iemp_5
Int_IempXages_5.

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT ki_age
  /METHOD=FORWARD  male to Int_IincXms__840031 .
* Selected ages marryXage Int_Iage_93 Int_IincXagesd7 Int_IincXagec840022
Int_IincXagea3 male maleXmarry .

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT ki_age2
  /METHOD=FORWARD  male to Int_IincXms__840031 .
* Selected ages2 Int_Iage_93 marryXage Int_IincXagec840017 Int_Iage_94
Int_IincXagec840022 Int_IincXageb840017 Int_IincXagec840027
Int_IincXagec840013 Int_Iage_80 Int_Iage_78 Int_IincXage_3 .

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000011) POUT(.00002)  /NOORIGIN
  /DEPENDENT ki_male
  /METHOD=FORWARD  male to Int_IincXms__840031 .
* Selected maleXmarry maleXage maleXage2 maleXage4 Int_IchiXmale_1 .

* Regress outcome on treatment(s) and relevant controls - moderation.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT satisfy  /METHOD=ENTER kid ki_marry ki_age ki_age2 ki_male
ms_marry2 ages ages2 ages3 inc2 inc3 male  Int_Iemp_5 Int_Iemp_7 marryXinc3
marryXage2 marryXage maleXmarry marryXage3  Int_IempXages_5 Int_Iage_93
Int_IincXagesd7 Int_IincXagec840022 Int_IincXagea3 Int_IincXagec840017
Int_Iage_94  Int_IincXageb840017 Int_IincXagec840027 Int_IincXagec840013
Int_Iage_80 Int_Iage_78 Int_IincXage_3  maleXage maleXage2 maleXage4
Int_IchiXmale_1.
```

Analysis 2: STATA

```
clear
* Data file created in 'Study 2 ANES Create Variables.do'
use anes_2000Final.dta

* demographics list
global demoNJlist incomeHH educ sex Zage Zage_sq marital notemployed
global demoused income incomeHH educ sex Zage Zage_sq marital notemployed
unemployed

log using Analysis.txt , replace text

*-------------------------------------------------------------------------------
* REPLICATE TABLE 1     - using only nonmissing cases in paper (MissNJVar)
*-------------------------------------------------------------------------------

* STEP 1: main relationship
reg DV_life_sat Orientation if MissNJVar==0 [pweight=V000002a]

* STEP 2: main relationship with controls
reg DV_life_sat Orientation $demoNJlist if MissNJVar==0 [pweight=V000002a]

* STEP 3: main relationship with controls + church
reg DV_life_sat Orientation $demoNJlist church2 if MissNJVar==0
[pweight=V000002a]

* STEP 4: main relationship with controls + church + NFC
reg DV_life_sat Orientation $demoNJlist church2 a_NFC if MissNJVar==0
[pweight=V000002a]

* STEP 5: mediator test
reg DV_life_sat Orientation $demoNJlist church2 a_NFC ration_ineq if
MissNJVar==0 [pweight=V000002a]


*-------------------------------------------------------------------------------
* REPLICATE TABLE 1 - including all usable cases (MissKeyVar)
*-------------------------------------------------------------------------------
* STEP 1: main relationship
reg DV_life_sat Orientation if MissKeyVar==0 [pweight=V000002a]

* STEP 2: main relationship with controls
reg DV_life_sat Orientation $demoNJlist if MissKeyVar==0 [pweight=V000002a]

* STEP 3: main relationship with controls + church
reg DV_life_sat Orientation $demoNJlist church2 if MissKeyVar==0
[pweight=V000002a]

* STEP 4: main relationship with controls + church + NFC
reg DV_life_sat Orientation $demoNJlist church2 a_NFC if MissKeyVar==0
[pweight=V000002a]

* STEP 5: mediator test
reg DV_life_sat Orientation $demoNJlist church2 a_NFC ration_ineq if
MissKeyVar==0 [pweight=V000002a]
```

```
*------------------------------------------------------------------------------
* DOUBLE LASSO ANALYSIS A - using only nonmissing cases in paper (MissNJVar)
*------------------------------------------------------------------------------
gen tempRI=ration_ineq

 * Select variables that predict the outcome ;
lassoShooting DV_life_sat $demoused church2 d_*  ration_ineq if MissNJVar==0
, lasiter(100) verbose(0) fdisplay(0)
global yvSel `r(selected)'
di "$yvSel"
reg DV_life_sat $yvSel if (MissNJVar==0)

* Select variables that predict the treatment ;
lassoShooting Orientation $demoused church2 d_* ration_ineq if
(MissNJVar==0), lasiter(100) verbose(0) fdisplay(0)
global x1vSel `r(selected)'
di "$x1vSel"
reg Orientation $x1vSel if (MissNJVar==0)

* Get union of selected instruments ;
global vDSA2="$yvSel " + "$x1vSel "
di "$vDSA2"

replace ration_ineq=0
* Mediation Equation with selected controls excluding ration_ineq ;
reg DV_life_sat Orientation $vDSA2 ration_ineq if MissNJVar==0
[pweight=V000002a]
replace ration_ineq=tempRI

* Mediation Equation with selected controls + ration_ineq ;
reg DV_life_sat Orientation $vDSA2 ration_ineq if MissNJVar==0
[pweight=V000002a]


*------------------------------------------------------------------------------
* DOUBLE LASSO ANALYSIS B - including all usable cases (MissKeyVar)
*------------------------------------------------------------------------------

* Select variables that predict the outcome ;
lassoShooting DV_life_sat $demoused church2 d_* ration_ineq if MissKeyVar==0
, lasiter(100) verbose(0) fdisplay(0)
global yvSel `r(selected)'
di "$yvSel"
reg DV_life_sat $yvSel if (MissKeyVar==0)

* Select variables that predict the treatment ;
lassoShooting Orientation $demoused church2 d_* ration_ineq if
(MissKeyVar==0), lasiter(100) verbose(0) fdisplay(0)
global x1vSel `r(selected)'
di "$x1vSel"
reg Orientation $x1vSel if (MissKeyVar==0)

* Get union of selected instruments ;
global vDSB2="$yvSel " + "$x1vSel "
di "$vDSB2"
```

```
replace ration_ineq=0
* Mediation Equation with selected controls excluding ration_ineq ;
reg DV_life_sat Orientation $vDSB2 ration_ineq if MissKeyVar==0
[pweight=V000002a]
replace ration_ineq=tempRI

* Mediation Equation with selected controls + ration_ineq ;
reg DV_life_sat Orientation $vDSB2 ration_ineq if MissKeyVar==0
[pweight=V000002a]

log close
```

Analysis 2: R

```
library(glmnet)
study3 = read.csv("dataanes.csv")  # read csv file
fixwt <- V000002a/.987761

# REPLICATE REGRESSION ANALYSES

# STEP 1: main relationship.
fitr <- lm(DV_life_sat ~ Orientation, data=study3, weights=fixwt)
summary(fitr) # show results

# STEP 2: main relationship with controls.
regvar <- as.matrix(subset(study3,select = c(Orientation, incomeHH, educ,
sex, Zage, Zage_sq, marital, notemployed)))
fitr <- lm(DV_life_sat ~ regvar, data=study3, weights=fixwt)
summary(fitr) # show results

# STEP 3: main relationship with controls + church.
regvar <- as.matrix(subset(study3,select = c(Orientation, incomeHH, educ,
sex, Zage, Zage_sq, marital, notemployed, church2)))
fitr <- lm(DV_life_sat ~ regvar, data=study3, weights=fixwt)
summary(fitr) # show results

# STEP 4: main relationship with controls + church + NFC.
regvar <- as.matrix(subset(study3,select = c(Orientation, incomeHH, educ,
sex, Zage, Zage_sq, marital, notemployed, church2, a_NFC)))
fitr <- lm(DV_life_sat ~ regvar, data=study3, weights=fixwt)
summary(fitr) # show results

# STEP 5: mediator test.
regvar <- as.matrix(subset(study3,select = c(Orientation, incomeHH, educ,
sex, Zage, Zage_sq, marital, notemployed, church2, a_NFC, ration_ineq)))
fitr <- lm(DV_life_sat ~ regvar, data=study3, weights=fixwt)
summary(fitr) # show results

# DOUBLE LASSSO ANALYSIS -- ALL DEMOGRAPHIC COVARIATES

testvar2 <- as.matrix(study3[,4:187])


# STEP 1: select variables that predict outcomes
n=nrow(testvar2)
p=ncol(testvar2)
sda = sd(residuals(lm(DV_life_sat ~ Orientation, data=study3)))
lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitY = glmnet(testvar2, DV_life_sat, weights=fixwt, lambda=lambda1)
  ba = coef(fitY, s = lambda1)
  ea = DV_life_sat-predict(fitY,testvar2)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
```

```
ba


# STEP 2: select variables that predict treatment
sd1=sd(Orientation)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT1 = glmnet(testvar2, Orientation, weights=fixwt, lambda=lambda1)
  ba = coef(fitT1, s = lambda1)
  ea = Orientation-predict(fitT1,testvar2)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba


# STEP 3: linear regression with both sets of variables, no mediator
use = union(which(abs(fitY$beta)>.0001),which(abs(fitT1$beta) > .0001))
summary(use)
X = cbind(Orientation,testvar2)
use1 = c(1,use+1)
fitr <- lm(DV_life_sat ~ X[,use1], data=study3, weights=fixwt)
summary(fitr) # show results

# STEP 4: linear regression with both sets of variables, include mediator
X = cbind(Orientation, ration_ineq, testvar2)
use2 = c(1,2,use+2)
fitr <- lm(DV_life_sat ~ X[,use2], data=study3, weights=fixwt)
summary(fitr) # show results


# STEP 5: linear regression with both sets of variables, include mediator but
not IV
X = cbind(ration_ineq, testvar2)
use3 = c(1,use+1)
fitr <- lm(DV_life_sat ~ X[,use3], data=study3, weights=fixwt)
summary(fitr) # show results
```

Analysis 2: SPSS

```
* Import data file: 'dataanes.csv'
* weight  data.
compute fixwt=V000002a/.987761.
execute.
WEIGHT BY fixwt.

*------------------------------------------------------------------------------.
* REPLICATE TABLE 1
*------------------------------------------------------------------------------.

* STEP 1: main relationship.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation.

* STEP 2: main relationship with controls.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation incomeHH educ sex Zage
Zage_sq marital notemployed .

* STEP 3: main relationship with controls + church.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation  incomeHH educ sex Zage
Zage_sq marital notemployed church2.

* STEP 4: main relationship with controls + church + NFC.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation incomeHH educ sex Zage
Zage_sq marital notemployed  church2 a_NFC.

* STEP 5: mediator test.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation  incomeHH educ sex Zage
Zage_sq marital notemployed  church2 a_NFC ration_ineq.

*------------------------------------------------------------------------------.
* VARIABLE SELECTION ANALYSIS
*------------------------------------------------------------------------------.
* p=141 and n=1192.
* cutoff = .1 / [log(n) * 2p] = .00005.

 * Select variables that predict the outcome .
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.00005) POUT(0.00006)  /NOORIGIN
  /DEPENDENT DV_life_sat
  /METHOD=FORWARD income incomeHH educ sex Zage Zage_sq marital notemployed
unemployed church2 a_NFC d_employ1 to d_Internet.
* three variable selected: d_workedforpay1 marital Zage_sq.

* Select variables that predict the treatment.
```

```
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.00005) POUT(0.00006)  /NOORIGIN
  /DEPENDENT Orientation
  /METHOD=FORWARD income incomeHH educ sex Zage Zage_sq marital notemployed
unemployed church2 a_NFC d_employ1 to d_Internet.
* four variables selected: d_Race2 church2 d_marital2 d_UnionHH2.

* Main Equation with selected controls, excluding mediator.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation d_workedforpay1 marital
Zage_sq d_Race2 church2 d_marital2 d_UnionHH2 .

* Main Equation with selected controls, including mediator.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DV_life_sat  /METHOD=ENTER Orientation d_workedforpay1 marital
Zage_sq d_Race2 church2 d_marital2 d_UnionHH2 ration_ineq.
```

Analysis 3: STATA

```
# delimit ;
set more off;
clear ;

capture log close ;
log using SimmonsFinalStata.txt , replace text ;

* Import and code data ;
import excel using "Study 2.xls" , first ;
keep if cond != "potato" ;
replace olddays = olddays - 10 ;
xi i.quarterback ;

 * REPLICATE ANALYSES ;
 * Replicate regression with controls ;
regress aged when64 if cond != "potato" ;

 * Replicate regression with controls ;
regress aged when64 dad if cond != "potato" ;


 * VARIABLE SELECTION ;

 * Select variables that predict the outcome ;
lassoShooting aged dad mom female bird political olddays computer diner _I*
if cond != "potato" , lasiter(100) verbose(0) fdisplay(0);
local yvSel `r(selected)' ;
di "`yvSel'" ;

 * Select variables that predict the treatment ;
lassoShooting when64 dad mom female bird political olddays computer diner _I*
if cond != "potato" , lasiter(100) verbose(0) fdisplay(0);
local xvSel `r(selected)' ;
di "`xvSel'" ;

 * Get union of selected instruments ;
local vDS : list yvSel | xvSel ;

 * Equation with selected controls ;
reg aged when64 `vDS' if cond != "potato" ;

clear ;
log close ;
```

Analysis 3: R

```
library(glmnet)
study2 = read.csv("study2.csv")  # read csv file

# REPLICATE REGRESSION ANALYSES
# Linear Regression
fitr <- lm(aged ~ when64, data=study2)
summary(fitr) # show results

# Linear Regression with covariate
fitr <- lm(aged ~ when64 + dad, data=study2)
summary(fitr) # show results

summary(study2$quarterback)
quarterback.f=factor(study2$quarterback)
qm = model.matrix(~quarterback.f)
summary(qm)

# DOUBLE LASSSO ANALYSIS
xall <- as.matrix(subset(study2,select = -c(aged, root, quarterback, potato,
when64, kaimba, feelold, cond, aged365)))
summary(xall)
xall1 <- cbind(xall,qm[,2:4])
summary(xall1)
aged <- as.matrix(study2$aged)
summary(aged)
when64 <- as.matrix(study2$when64)
summary(when64)

# STEP 1: select variables that predict outcomes

n=nrow(xall1)
p=ncol(xall1)
sd1=sd(aged)
lambda1 = .5*sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)

k = 1
while(k < 15){
  fitY = glmnet(xall1, aged, lambda=lambda1)
  ba = coef(fitY, s = lambda1)
  ea = aged-predict(fitY,xall1)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

# STEP 2: select variables that predict treatment

n=nrow(xall1)
p=ncol(xall1)
sd1=sd(when64)
lambda1 = .5*sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
```

```
k = 1
while(k < 15){
  fitT = glmnet(xall1, when64, lambda=lambda1)
  ba = coef(fitT, s = lambda1)
  ea = aged-predict(fitT,xall1)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

# STEP 3: linear regression with both sets of variables
use = union(which(abs(fitY$beta)>0),which(abs(fitT$beta) > 0))

X = cbind(when64,xall1)
use = c(1,use+1)

fitr <- lm(aged ~ X[,use], data=study2)
summary(fitr) # show results
```

Analysis 3: SPSS

```
* Import data file: Study 2.xls'

* select 'when I'm 64' and control conditions.
USE ALL.
COMPUTE filter_$=(cond<>"potato").
FILTER BY filter_$.
EXECUTE.

compute QB1=0.
compute QB2=0.
compute QB3=0.
if (quarterback eq 1) QB1=1.
if (quarterback eq 2) QB2=1.
if (quarterback eq 3) QB3=1.
execute.

* replicate basic regression.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT aged
  /METHOD=ENTER when64 dad.

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT aged
  /METHOD=ENTER when64 dad mom female bird political olddays computer diner
QB1 QB2 QB3.

* VARIABLE SELECTION.
* p=11 and n=20.
* cutoff = .1 / [log(n) * 2p] = .0015.

* predict dependent variable.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.0015) POUT(.002)  /NOORIGIN
  /DEPENDENT aged
  /METHOD=FORWARD when64 dad mom female bird political olddays computer
diner.

* predict treatment.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.0015) POUT(.002)  /NOORIGIN
  /DEPENDENT when64
  /METHOD=FORWARD dad mom female bird political olddays computer diner.

* no variables selected for inclusion.

* final model.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT aged
  /METHOD=ENTER when64.
```

Analysis 4: STATA

```
# delimit ;
set more off;
clear all;
log using RunDonations.txt , replace text ;

* note: new variables were coded in 'donations data prep.do';
import delimited Study3Data.csv;

* Define variable lists;
global maniplist manip_deflvl_low manip_deflvl_med manip_deflvl_high
manip_numopt5 manip_prevremind;
display "$maniplist";
global demolist white asian ethnicity_missing gender_male prospect zage
age_missing zyrsofassos yrsassocmissing zafconsyog zboothaflifetimegiving
zboothlifetimegiving lybunt sybunt otherseg;
display "$demolist";

* Basic Regressions;
reg dvln_donamt manip_deflvl_low if (manip_deflvl_med==0 &
manip_deflvl_high== 0);
reg dvln_donamt $maniplist ;

* Variable Selection;
* Select variables that predict the outcome ;
lassoShooting dvln_donamt  $demolist sq_* ln_* int* , lasiter(100) verbose(0)
fdisplay(0) ;
global yvSel `r(selected)' ;
di "$yvSel" ;

* Select variables that predict the treatments ;
lassoShooting manip_deflvl_low $demolist sq_* ln_* int* , lasiter(100)
verbose(0) fdisplay(0) ;
global x1vSel `r(selected)' ;
di "$x1vSel" ;

lassoShooting manip_deflvl_med $demolist sq_* ln_* int* , lasiter(100)
verbose(0) fdisplay(0) ;
global x2vSel `r(selected)' ;
di "$x2vSel" ;

lassoShooting manip_deflvl_high $demolist sq_* ln_* int* , lasiter(100)
verbose(0) fdisplay(0) ;
global x3vSel `r(selected)' ;
di "$x3vSel" ;

* Get union of selected instruments ;
global vDS="$yvSel " + "$x1vSel " + "$x2vSel " + "$x3vSel ";
di "$vDS" ;

* Regression with selected controls ;
reg dvln_donamt $maniplist $vDS  ;

clear ;
log close ;
```

Analysis 4: R

```
library(glmnet)
study3 = read.csv("Study3Data.csv")  # read csv file

# REPLICATE REGRESSION ANALYSES
# Linear Regression
manip <- as.matrix(subset(study3,select = c(Manip_defLvl_Low,
Manip_defLvl_Med, Manip_defLvl_High, Manip_numOpt5, Manip_prevRemind)))
summary(manip)

fitr <- lm(DVln_DonAmt ~ Manip_defLvl_Low, data=study3)
summary(fitr) # show results

fitr <- lm(DVln_DonAmt ~ manip, data=study3)
summary(fitr) # show results


# DOUBLE LASSSO ANALYSIS
test1 <- as.matrix(subset(study3,select = c(White, Asian, Ethnicity_Missing,
Gender_Male, Prospect, Age_missing, LYBUNT, SYBUNT, OtherSeg)))
test2 <-as.matrix(subset(study3[,28:217],select= -c(int43, int63, int73)))
testvar <- cbind(test1,test2)
DVln_DonAmt <- as.matrix(study3$DVln_DonAmt)
Manip_defLvl_Low <- as.matrix(study3$Manip_defLvl_Low)
Manip_defLvl_Med <- as.matrix(study3$Manip_defLvl_Med)
Manip_defLvl_High <- as.matrix(study3$Manip_defLvl_High)
summary(DVln_DonAmt)


# STEP 1: select variables that predict outcomes
n=nrow(testvar)
p=ncol(testvar)
sda = sd(residuals(lm(DVln_DonAmt ~ manip, data=study3)))
lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)


k = 1
while(k < 15){
  fitY = glmnet(testvar, DVln_DonAmt, lambda=lambda1)
  ba = coef(fitY, s = lambda1)
  ea = DVln_DonAmt-predict(fitY,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

# STEP 2: select variables that predict treatment
n=nrow(testvar)
p=ncol(testvar)
sd1=sd(Manip_defLvl_Low)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
```

```
k = 1
while(k < 15){
  fitT1 = glmnet(testvar, Manip_defLvl_Low, lambda=lambda1)
  ba = coef(fitT1, s = lambda1)
  ea = Manip_defLvl_Low-predict(fitT1,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

n=nrow(testvar)
p=ncol(testvar)
sd1=sd(Manip_defLvl_Med)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)
k = 1
while(k < 15){
  fitT2 = glmnet(testvar, Manip_defLvl_Med, lambda=lambda1)
  ba = coef(fitT2, s = lambda1)
  ea = Manip_defLvl_Med-predict(fitT2,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

n=nrow(testvar)
p=ncol(testvar)
sd1=sd(Manip_defLvl_High)
lambda1 = sd1*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
summary(lambda1)

k = 1
while(k < 15){
  fitT3 = glmnet(testvar, Manip_defLvl_High, lambda=lambda1)
  ba = coef(fitT3, s = lambda1)
  ea = Manip_defLvl_High-predict(fitT3,testvar)
  sda = sd(ea)
  lambda1 = sda*(1.1/sqrt(n))* qnorm(1 - (.1/log(n))/(2*p))
  k = k+1
}
ba

# STEP 3: linear regression with both sets of variables
use1 = union(which(abs(fitY$beta)>.0001),which(abs(fitT1$beta) > .0001))
use2 = union(which(abs(fitT2$beta) > .0001),which(abs(fitT3$beta) > .0001))
use = union (use1, use2)
summary(use)

X = cbind(manip,testvar)
use = c(1,2,3,4,5,use+5)

#final regression model
fitr <- lm(DVln_DonAmt ~ X[,use], data=study3)
summary(fitr) # show results
```

Analysis 4: SPSS

```
* Import data file: Study 3Data.csv.
* delete variables that are constants.
delete variables YrsAssocMissing int43 int63 int73.
* note that missing age was coded to average.

* Basic effect.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DVln_DonAmt
  /METHOD=ENTER Manip_defLvl_Low Manip_defLvl_Med Manip_defLvl_High
Manip_numOpt5 Manip_prevRemind.

* VARIABLE SELECTION.
* p=196 and n=76.
* cutoff = .1 / [log(n) * 2p] = .0006.

* Select variables that predict the outcome.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000059) POUT(.00007)  /NOORIGIN
  /DEPENDENT DVln_DonAmt
  /METHOD=FORWARD White Asian Ethnicity_Missing Gender_Male Prospect ZAge
Age_missing ZYrsOfAssos  ZAFConsYoG ZBoothAFLifetimeGiving
ZBoothLifetimeGiving LYBUNT SYBUNT OtherSeg ln_Age to int98.
* variable ln_Menu2Num selected.

* Select variables that predict the treatment.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000059) POUT(.00007)  /NOORIGIN
  /DEPENDENT Manip_defLvl_Low
  /METHOD=FORWARD White Asian Ethnicity_Missing Gender_Male Prospect ZAge
Age_missing ZYrsOfAssos  ZAFConsYoG ZBoothAFLifetimeGiving
ZBoothLifetimeGiving LYBUNT SYBUNT OtherSeg ln_Age to int98.

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000059) POUT(.00007)  /NOORIGIN
  /DEPENDENT Manip_defLvl_Med
  /METHOD=FORWARD White Asian Ethnicity_Missing Gender_Male Prospect ZAge
Age_missing ZYrsOfAssos ZAFConsYoG ZBoothAFLifetimeGiving
ZBoothLifetimeGiving LYBUNT SYBUNT OtherSeg ln_Age to int98.

REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.000059) POUT(.00007)  /NOORIGIN
  /DEPENDENT Manip_defLvl_High
  /METHOD=FORWARD White Asian Ethnicity_Missing Gender_Male Prospect ZAge
Age_missing ZYrsOfAssos  ZAFConsYoG ZBoothAFLifetimeGiving
ZBoothLifetimeGiving LYBUNT SYBUNT OtherSeg ln_Age to int98.
* no variables selected.

* Final model with selected controls .
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN
  /DEPENDENT DVln_DonAmt
  /METHOD=ENTER Manip_defLvl_Low Manip_defLvl_Med Manip_defLvl_High
Manip_numOpt5 Manip_prevRemind ln_Menu2Num.
```