"Simulating the Effect of Controlling for Post-Treatment Variables on Treatment Estimates"

Shannon Kay

APSTA 2012 Causal Inference

Motivation

While it seems obvious that including variables affected by an experimental treatment when analyzing results might bias the estimate of a treatment, the reality is not as straightforward. One of the main motivations for researching the effects of post-treatment variables comes from my experience working as a Research Assistant at the Marron Institute. Oftentimes, our practitioners are interested in what they consider the true effect of the treatment—the difference in outcomes for those in the treatment group as adjusted for dosage or another mediating factor, rather than simply looking at an intent to treat analysis of those who were assigned to the treatment condition. In some trials, there may be a valid method to complete this analysis with appropriate statistical rigor. More often, though, unless this possibility is considered during the trial's design, intent to treat is the only possible method to make any causal attributions. The problem, however, arises when we are asked to subset the data based on a post-treatment variable, or include a non-randomized dosage measure in the analysis.

When treatment groups are randomized, we assume ignorability—that once randomized, the treatment and control groups are equal on all measured and unmeasured pre-treatment characteristics. The addition of post-treatment variables breaks the ignorability assumption by controlling for a variable that was not taken into consideration during randomization. In other words, "Conditioning on post-treatment variables eliminates the advantages of randomization because we are now comparing dissimilar groups" (Montgomery, Nyhan, & Torres, 2016). To understand how the inclusion of post-treatment variables influences treatment effect estimates, I imagine a hypothetical situation where 1000 NYU students enrolled in Stats 101 are randomly encouraged to attend an optional lab to supplement their statistics class. Random assignment generates treatment and control groups that are evenly distributed with regard to gender and age. Students who are not encouraged may choose to attend the lab but must find the information independently. There are 10 lab sessions. Students' final grades for Stats 101 are recorded at the end of the semester, as well as the number of lab sessions they attended. Attendance thus becomes our post-treatment variable.

To understand how controlling for attendance affects the estimate of the treatment effect, I will use an instrumental variable (IV) approach to estimate the Complier Average Causal Effect (CACE) and two propensity score methods to estimate the Average Treatment Effect on the Treated (ATT). Since the instrumental variable method requires an instrument as well as treatment, which the propensity score estimators require only treatment status, I will simulate data using two data-generating processes. All methods will compare three post-treatment attendance variables of different strength correlation to the instrument for instrumental variables, and to treatment for propensity scores. I choose to investigate these estimands because they are most comparable to the interests of our practitioners.

Assumptions

Instrumental Variables

The first assumption behind instrumental variables is *ignorability*, which means that the instrument is randomly assigned. Random assignment allows us to assume that the groups generated are equal on all observed covariates, and that the observed covariates are the only confounding covariates. With this established, we can reasonably attribute any differences in outcomes to the treatment. In the case of instrumental variables, assignment to encouragement, the instrument, is randomized.

In instrumental variables, there are four possible subject types. The never-takers will not participate in the treatment regardless of whether they are offered the treatment. The compliers will participate if they are offered the treatment and will not participate if they are not offered the treatment. Always-takers will participate in the treatment regardless of being offered the treatment (i.e., they will seek out and obtain the treatment on their own). Defiers will do the opposite of their treatment assignment; if they are offered the treatment, they will refuse it, but if they are not offered the treatment, they will seek it out and obtain it. Our second assumption, *monotonicity*, assumes that there are no defiers.

From there, we assume *exclusion restriction*. Within our defined subject types, always-takers and never-takers will always be treated or untreated, regardless of instrument assignment (encouragement)—the always takers will always find a way to receive the treatment, and never-takers will never accept treatment. By exclusion treatment, if your treatment condition remains the same under either instrument assignment, you will have the same outcome under either

instrument assignment. This implies that always-takers and never-takers will have the same final grade in Stats 101 whether or not they are encouraged to attend the optional lab sessions.

An important assumption for instrumental variables is a *non-zero correlation between the instrument and the treatment*. In this case, encouragement must correlate

Lastly, we assume *SUTVA,* or the *stable unit treatment value assumption*. Colloquially, the treatment assignment of one subject will not affect the outcome of another subject. In terms of encouragement to go to an optional, supplementary lab, this means assuming that those who are encouraged to go do not change the treatment assignment of another student by bringing friends who were not encouraged to go to the lab sessions, and that …….

Propensity Scores

Propensity score models also assume ignorability of the treatment assignment (rather than instrument assignment) and SUTVA in the same manner that these assumptions are made for instrumental variables. The additional assumptions for propensity score methods are *balance* and *overlap*. Propensity scores assume that the method use appropriately specified, and that ultimately balance is achieved. Balance can easily be tested, unlike many other assumptions. Overlap assumes that the treated and control groups have enough commonalities to make a fair comparison rather than extrapolating beyond the range of the data. Like balance, overlap can also be tested.

Designs & Estimators

Estimator 1: Two-Stage Least Squares Regression

In Two-Stage Least Squares (TSLS), we estimate the CACE by first regressing whether or not students attended supplementary lab sessions (treatment) on whether or not they were encouraged to attend the labs (instrument). The coefficient for on the instrument provides an estimate of the percentage of compliers in our sample. We then regress final Stats 101 grades on the treatment along with any other relevant predictors. The resulting coefficient on treatment is the estimated Complier Average Causal Effect. While TSLS is known to provide incorrect estimates of standard errors because it does not account for correlation between the two regression equations, it is a simple and flexible way to utilize instrumental variables. I originally intended to use the ivreg function, which corrects the standard error estimates, but the function returned an error message that there were more regressors than instruments. Some research

suggests that this error was related to the data-generation process of the covariates (see simulation setup).

Estimator 2: Propensity Score Weighting using MatchIt

MatchIt with nearest-neighbor matching selects the best control comparison from largest to smallest for each individual within the treatment group one at a time, reweighting the controls to include in the final dataset as it proceeds through the data. Though the Matchit function supports a variety of distance metrics, I chose nearest neighbor for its simplicity. Nearest neighbor additionally makes sense for such a small set of covariates, as opposed to Mahalanobis distance, which is multivariate and arguably more useful when reducing dimensionality on a larger set of predictors. The weights are determined using all specified, exogenous pre-treatment variables. Once weighted, the weights are used in a linear regression model, and the resulting coefficient for treatment is the Average Treatment Effect on the Treated (ATT).

Estimator 3: Propensity Score Weighting using Boosted Logistic Regression Trees (PS)

For logistic regression trees, I also set the model to estimate the ATT. Boosted logistic regression trees calculate propensity score weights through a decision-tree model, rather than a distance metric. While the default number of trees is 10,000, this was not computationally feasible for 1000 iterations of the simulation and was reduced to 500 trees.

Comparison

Though CACE and ATT are different estimands, they both estimate areas of interest for our practitioners. CACE examines how the treatment affects outcomes for those randomly encouraged to participate in the treatment who adhere to the treatment, while ATT estimates the average treatment effect on those individuals who received the treatment, which does not consider the various subject types as in instrumental variables (i.e., that those who received the treatment could be either compliers or always-takers, where the first might demonstrate a change in outcome due to the treatment whereas the second is assumed to not demonstrate a change in outcome from treatment assignment).

Testing several different methods to estimate treatment effects will illustrate which are more or less robust with regard to the potential bias induced by controlling for post-treatment variables. Although much of the literature advises against using these variables to subset data or as controls in a model, intuition would suggest that some methods are less susceptible, and as

such their use in these models, while still ill-advised, may be less problematic. Since in practice important predictors are often measured post-treatment, there is benefit in determining ways that retain the improvements from incorporating these measures while minimizing the potential bias induced by recording them post-treatment.

Simulation Set-Up

Generating Covariates

In this simulation, I aim to explore treatment estimation methods' reliance on the ignorability assumption by controlling for post-treatment variables in my models. For simplicity, the age and gender covariates included in the model are already balanced across groups, to allow for isolation of violating the ignorability assumption via the inclusion of the post-treatment variable. Age is generated as a random normal distribution with a mean of 20 and a standard deviation of .5 to approximate the age distribution of undergrad students.
If other covariates in the model had significant differences between groups, they could potentially bias the estimates. Precluding other possible bias ensures that any bias in the estimates of treatment effects is due to the post-treatment variables.

Generating Potential Outcomes

Potential outcomes for TSLS are generated separately for always-takers, never-takers, and compliers. Always-takers and never-takers have the same potential outcome when not randomly encouraged to partake in the supplementary lab sessions (Y0) as when encouraged (Y1). The final Stats 101 grades for always-takers come from a normal distribution with a mean of 87 and a standard deviation of 3.5, while the grades for never-takers are simulated from a normal distribution with a mean of 75 and a standard deviation of 7. This assumes that always-takers generally take advantage of all available resources and as such tend to perform better on average, as opposed to never-takers who do not take advantage of available resources and as such perform worse on average, but may a mix of students who naturally do well on their own and students who do not typically do as well and could benefit from additional academic support. Contrary to the always-takers and the never-takers, compliers' outcomes vary with their instrument assignment, and thus their treatment conditions. I assume that compliers do moderately well in the Stats 101 class under the baseline condition where they are not encouraged to attend supplementary labs and generate their unencouraged test scores from a

normal distribution with a mean of 80 and a standard deviation of 3.5. Since compliers are known to benefit from participation in the treatment, we retain the same standard deviation but increase the final grades from 80 to 89. This assigns a true CACE of 9 points from participating in the supplementary lab.

Potential outcomes for propensity scores are only generated as grade if not treated (Y0) or grade if treated (Y1). Y0 is generated from a normal distribution with a mean of 80 and a standard deviation of 3.5, and Y1 is generated from a normal distribution with a mean of 89 and stardard deviation of 3.5.

Generating Post-Treatment Variables

I generate three post-treatment attendance variables that are correlated with compliance in the case of instrumental variables, or with treatment in the case of propensity score matching. Attendance is a proportion of the possible number of lab sessions. If the attendance variable is 1, the student attended all 10 sessions, if attendance is 0, the student did not attend any lab sessions. I vary the degree of correlation by using a random beta distribution where alpha is held at 2 while beta is manipulated.

For instrumental variables, all three attend variables are always 1, or 100% attendance, for the always-takers and always 0, or never attended. Compliers who are not randomly assigned encouragement also always receive 0's for attendance. The manipulation is contained to compliers who are randomly assigned to encouragement. In Attend 1, beta is equal to 15; this assigns the majority of the compliers an attendance rate under .3, or 3 labs, and potentially makes the compliers look more like the never-takers. In Attend 2, beta is equal to 2, which assigns attendance normally from 0 to 1, and places most compliers' attendance in between that of the always-takers and the never-takers. In Attend 3, beta is equal to .5, which places the majority of compliers at an attendance rate closer to that of the always-takers. This allows me test the possibility that compliers may not completely comply. Though we understand that by nature compliers tend to adhere to their assignments, it does not necessarily follow that they do so to the fullest possible extent. Breaking ignorability by correlating the post-treatment attendance variable to compliance in this manner illustrates the different degrees to which compliers satisfy their instrument assignment.

For propensity scores, manipulation of post-treatment variables is relegated to those randomly assigned to the treatment. All non-treated subjects receive a 0 for attendance. The

distribution for the three attendance variables is manipulated in the same manner as in the IV data generating progress, but for all who are assigned the treatment condition. The data were balanced with respect to the age and gender covariates prior to modeling, and checked after matching/weighting, though balance was not included each draw of the simulation.

Z (instrument assignment in IV estimates and treatment assignment in propensity score methods) is generated from random binomial distribution, and an observed data frame is constructed based on the unit's assignment.

Modeling

For each method, I simulate a baseline estimate that does not control for a post-treatment variable, followed by one model per post-treatment variable. This generates 4 estimates per method. I expect the baseline model for each method to be unbiased and will use this as a comparison for how including post-treatment variables may influence the estimate. I attempted to run a few different models using the ivreg function; ivreg, though, requires the model to include some endogenous variables, which was not compatible with the method use to generate the data and isolate violation of the ignorability assumption.

Simulation Results

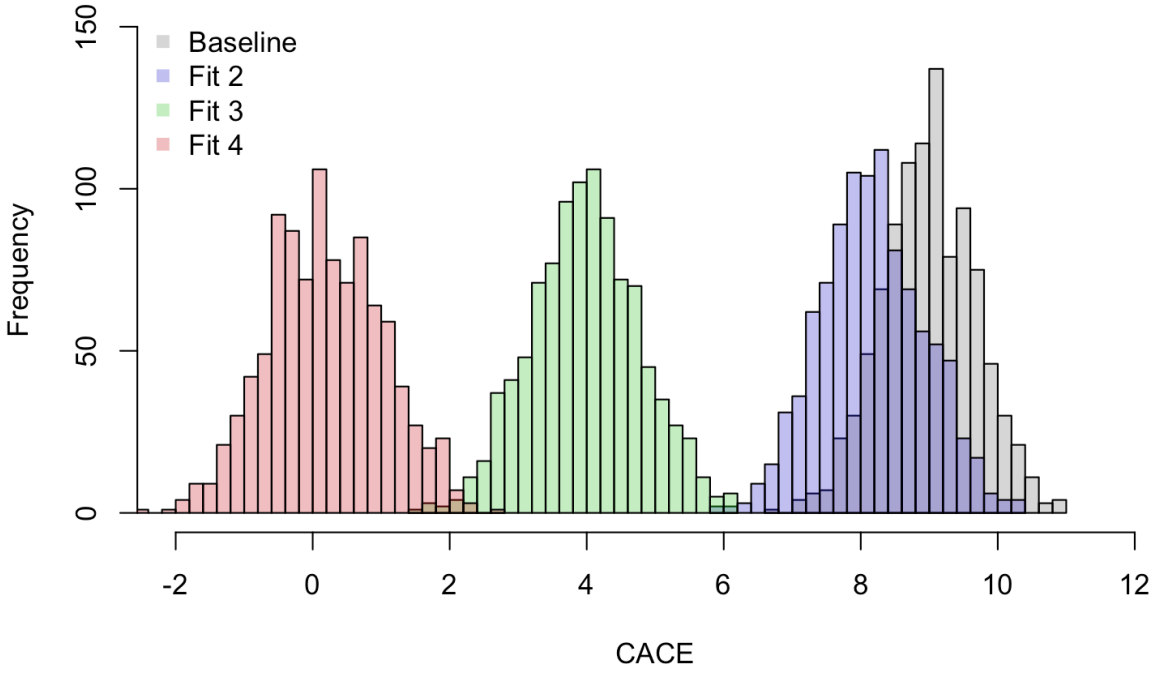
Table 1. Estimates by Model—Mean(Sd.)

Model	TSLS	MatchIt	Logistic Regression Trees
Estimand	**CACE**	**ATT**	**ATT**
Baseline	8.99 (0.67)	8.97 (0.34)	9.01 (0.28)
Fit 2	8.16 (0.76)	8.89 (3.55)	9.01 (0.43)
Fit 3	3.98 (0.80)	8.88 (3.59)	9.01 (0.52)
Fit 4	0.19 (0.84)	9.04 (3.51)	9.01 (0.52)

Generally speaking, the baseline models all captured the true treatment effect—CACE for TSLS, and ATT for MatchIt and Logistic Regression Trees. Table 1 provides the mean and standard deviation of the distribution of estimates for each model. In TSLS regression, we see that the post-treatment variables of different strength correlation to compliance considerably bias

the estimate of the CACE. Most likely, this is because varying the correlation to compliance made the compliers more similar to either the never-takers or the always-takers. For example, in Fit 2, the attendance variable is designed to generate attendance rates mostly between 0 and .3. This diminishes the CACE estimate slightly because some of the compliers appear more like the never-takers. In Fit 3, the attendance variable is designed to generate attendance rates from 0-1; in this case, some compliers look more like never-takers, and some look more like always-takers. In Fit 4, the compliers' attendance rates are almost identical to the always-takers. As illustrated in Figure 1 below, the distribution of Fit 2 contains the true CACE, while Fits 3 and 4 do not encompass the true CACE at all.

Figure 1: Bias In TSLS



Both propensity score methods are centered around the true CACE. However, they have a much larger standard deviation, and thus a much wider distribution, than the baseline estimate. While the loss in the TSLS regression was with respect to accuracy, the loss here occurs in precision. When we look at balance before and after propensity score matching, it becomes clear that neither the MatchIt nor the Boosted Logistic Regression Tree methods are able to achieve balance for any of the attendance variables. This makes sense; it was occasionally possible for to match treatment observations to controls when the beta distribution assigned a treatment a 0 for

attendance, but for the majority of cases there was no way to match treatment and controls for comparison. Examination of Figures 3 and 4 illustrates that MatchIt and Boosted Logistic Regression Trees both have baseline estimations of the ATT that center around 9 points and range from 8 to 10. Boosted Logistic Regression has greater variation than on all fits that include a post-treatment variable, with a range of roughly 7 to 11, but performs better than MatchIt, which has a range wider than 0 to 18.

Figure 2: Bias In MatchIt Estimates

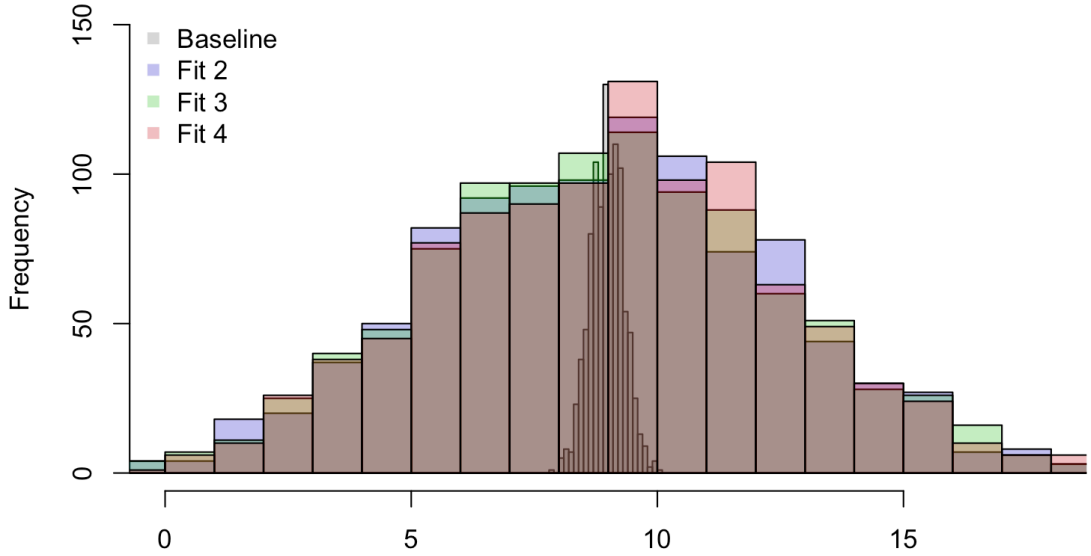
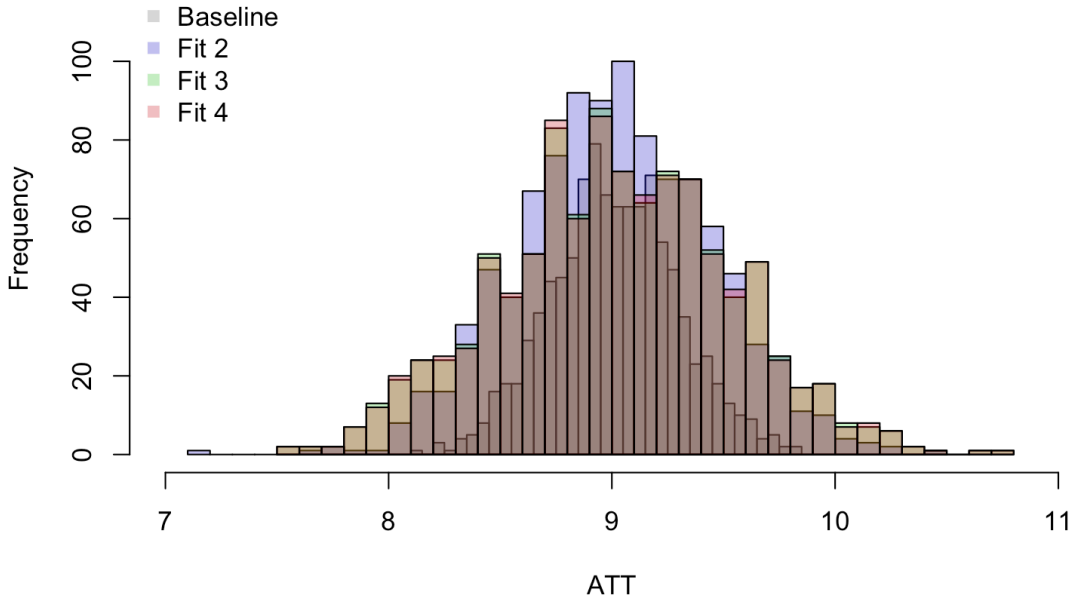


Figure 3: Bias In Logistic Regression Tree Propensity Score Estimates



Discussion

Ideally, I would have simulated the effects of including post-treatment variables with one perfectly correlated variable, one variable with a correlation of about .5, and one completely uncorrelated variable. However, regressions produce NA coefficients when highly correlated variables are included, and generating a completely uncorrelated post-treatment variable would break other assumptions, such as exclusion restriction. Out of the three estimation methods tested, propensity score matching using Boosted Logistic Regression Trees performed better than Two-Stage Least Squares Regression and MatchIt in terms of resisting the potential bias induced by controlling for post-treatment variables. Given these three choices, I would choose the Boosted Logistic Regression Trees to estimate ATT should including a post-treatment variable was unavoidable, or considered to induce less bias than the bias resulting from not including them.

It was surprising to see that the simulations only revealed bias in the CACE estimates using instrumental variables, while the greater problem with propensity scores estimates was the amount of variation in the estimates. I initially expected to see biased treatment estimates from the models that included an attendance variable. These simulations helped me think critically about the consequences of choosing control for post-treatment variables on a deeper level, which will allow me to give better explanations of the problem when it next presents itself.

References

Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL http://www.jstatsoft.org/v42/i08/

Greg Ridgeway, Dan McCaffrey, Andrew Morral, Beth Ann Griffin and Lane Burgette (2017). twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.5. https://CRAN.R-project.org/package=twang

Montgomery JM, Nyhan B, Torres M. How conditioning on post-treatment variables can ruin your experiment and what to do about it. Annual meeting of the Midwest Political Science Association, Chicago, IL; 2016.

Rosenbaum, P. R. (1984). The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A (General)*, *147*(5), 656. doi: 10.2307/2981697