Misspecification and the Propensity Score: The Possibility of Overadjustment *

Kevin A. Clarke †
Brenton Kenkel §
Miguel R. Rueda ‡

Abstract

The popularity of propensity score matching has given rise to a robust, sometimes informal, debate concerning the number of pre-treatment variables that should be included in the propensity score. The standard practice when estimating a treatment effect is to include all available pre-treatment variables, and we demonstrate that this approach is not always optimal when the goal is bias reduction. We characterize the conditions under which including an additional relevant variable in the propensity score increases the bias on the effect of interest across a variety of different implementations of the propensity score methodology. Moreover, we find that balance tests and sensitivity analysis provide limited protection against overadjustment.

July 12, 2011

*We thank Jake Bowers, John Jackson, and Michael Peress for helpful comments and discussion. A previous version of this paper was given at the 27th Annual Summer Meeting of the Society for Political Methodology. We thank the participants for their comments. Errors remain our own.

†Corresponding author. Associate Professor, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: kevin.clarke@rochester.edu.

§Graduate Student, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: bkenkel@mail.rochester.edu.

‡Graduate Student, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: mrueda@mail.rochester.edu.

1 Introduction

A recent debate in the pages of *Statistics in Medicine* featured Shrier [1], Pearl [2], and Sjölander [3] responding to a paper by Rubin [4]. A chief concern of the participants was the question of whether conditioning on a new covariate can only decrease confounding bias. The three responses note an instance when the use of propensity score methods would increase the bias on estimates of the average treatment effect (ATE) above that of an unadjusted comparison between treated and untreated observations [2, 1415]. Rubin [5] responds by arguing that the example amounts to a mere mathematical curiosity.[1]

This debate included plenty of entertaining rhetoric, but was short on systematic evidence. The goal of this paper is to provide that evidence by assessing these claims and their implications. First, we characterize the conditions under which controlling for an additional variable in the propensity score increases the bias on the estimate of the ATE. We find that these circumstances are not unusual. Second, we characterize these conditions across a variety of different implementations of the propensity score methodology in order to assess which, if any, of these implementations helps mitigate the problem. Our results show that differences across techniques are minor. Finally, we assess the ability of post-estimation tools such as balance tests and sensitivity analysis to alert us to the problem of overconditioning. The results are not encouraging.

Our findings lend little credence to the claim that a researcher should condition on all available pretreatment covariates. Which variables should be included in a data analysis depends, as we demonstrate, on a number of factors that vary from situation to situation. In choosing covariates, researchers need to rely on theory, judgment, and common sense. The paper ends with a discussion of how our results can be helpful to applied researchers.

[1] A similar debate began between Judea Pearl and Andrew Gelman on the latter's weblog (http://stat.columbia.edu/∼gelman/blog/). This exchange attracted the notice of other prominent Bayesian statisticians such as Philip Dawid.

2 The debate and previous research

The exchange in *Statistics in Medicine* begins with Shrier's [2008] letter regarding Rubin's [2007] paper on the design of observational studies. Specifically, Shrier [1] raises the issue of selection bias caused by controlling for a covariate that is the common effect of two independent variables. He is interested in "M-structures," where a treatment X causes an outcome Y, an unmeasured covariate, U_1, causes both X and a measured covariate Z, and a second unmeasured covariate, U_2, causes both the measured covariate Z and Y (see Figure 1). In this situation, if a researcher controls for the measured covariate Z, a spurious dependence between X and Y is created that would bias the estimate of the causal effect of X on Y [6, 1].

[Figure 1 about here.]

Pearl [2, 1415] expands on Shrier's [2008] point and argues that the use of propensity score techniques increases the bias on the estimate average treatment effect whenever

> ...treatment is strongly ignorable to begin with and becomes non-ignorable at some levels of e_i. In other words, although treated and untreated units are balanced in each stratum of e_i, the balance only holds relative to the covariates measured; unobserved confounders may be highly unbalanced in each stratum of e_i, capable of producing significant bias.

That is, the propensity score balances treated and untreated observations relative only to observed covariates. A new association is introduced between treatment and outcome by conditioning on a variable that is not causally related to either, but is an indicator of unobserved factors that are not balanced. This new association may increase or decrease the bias. Pearl [2, 1416] concludes that the effectiveness of propensity score techniques depends "critically on the choice of covariates, and that choice cannot be left to guesswork."

Rubin's [2009, 1421] response (and to a lesser extent Gelman's blog response[2]) to these claims is that not controlling for an observed covariate is

[2] http://www.stat.columbia.edu/c̃ook/movabletype/archives/2009/07/disputes_about.html

bad practical advice "in all but the most unusual circumstances." Furthermore, he argues that even if one were to condition on Z in Figure 1, the result would be inefficient, but not biased. In the end, he argues that not conditioning on an observed covariate because of fears of increasing bias "is neither Bayesian nor scientifically sound but rather it is distinctly frequentist and nonscientific ad hocery" [1421].

We take no position on the Bayesian or un-Bayesian nature of conditioning. Our interest lies in assessing the risk of increasing bias through conditioning, and whether the choice of propensity score techniques makes a difference. Simulation is our method of choice, and consequently we address the history of Monte Carlo studies used to explore misspecification of the propensity score. Numerous articles have been published on the subject, and we touch only on those studies that are particularly relevant to the discussion. Drake [7], for example, finds that there is little difference between propensity score methods and prognostic (regression) models with regard to omitted confounders. The biases for both techniques are "large and of the same magnitude" [1231]. Additionally, she finds that misspecifications of the propensity score in terms of functional form have much smaller biases than similar misspecifications of the response model.

The simulation design of Augurzky and Schmidt [8] includes a set of variables, $\mathbf{Z}$, that strongly influence exposure to treatment, but do not or only weakly determine the outcome. Also included is a set of variables, $\mathbf{Y}$, that influence the outcome, but are irrelevant to exposure. Their results indicate that including $\mathbf{Z}$ and $\mathbf{Y}$ in the propensity score has two effects. One, the inclusion of these variables balances $\mathbf{Z}$ at the expense of the variables most relevant to exposure and outcome, and two, unnecessary effort is used to remove small imbalances in $\mathbf{Y}$ [26]. They find that leaving $\mathbf{Z}$ and $\mathbf{Y}$ out of the propensity score equation produces average treatment effect estimates that are often better in terms of root mean squared error than including all the covariates. The authors recommend including only highly significant variables in the propensity score equation.

Brookhard et al. [9] present two simulations that pick up on some of the same design features as Augurzky and Schmidt [8]. In their first simulation, they include three different types of covariates: one related to both the outcome and exposure X_1; one related to the outcome, but not the exposure X_2; and one related to the exposure, but not the outcome X_3. They find that

the model that best predicts exposure does not yield the optimal propensity score model in terms of MSE; the optimal model included X_1 and X_2, but not X_3 [6]. Thus, one should include in the selection equation variables that are thought to be related to the outcome, whether or not they are related to the exposure. They also found that adding variables to the propensity score model that are unrelated to the outcome but related to the exposure increases the variance of an estimated exposure effect without decreasing its bias [7].

In their second simulation, Brookhart *et al.* (2006) look at the addition of a covariate to a propensity score model when varying the strength of the covariate-outcome and covariate-exposure relations [3]. They found that in small studies situations exist where it would be better, in terms of MSE, to exclude a true confounder from the propensity score model [7].

Millimet and Tchernis [10] report a simulation that focuses on, essentially, the functional form of the propensity score. In particular, they look at the exclusion of relevant higher-order terms and the inclusion of irrelevant higher-order terms. Their results suggest that overfitting the propensity score model results in greater efficiency, and in the other cases, overfitting does no worse than the correctly specified model. They conclude that the penalty for overfitting is minimal. The experimental evidence, therefore, is mixed.

The Monte Carlo experiments described in the next section are closely related to the experiments performed in Clarke [11] and Clarke [12]. In those papers, Clarke revisits the omitted variable bias result familiar to most applied researchers. The standard omitted variable result compares a linear regression that includes all relevant variables to a specification where some are omitted, finding that the latter is biased. Clarke compares a specification with multiple omitted variables to one where some, but not all, of these variables are included. He finds that, in some circumstances, the inclusion of additional control variables increases the bias on a coefficient of interest. Clarke [12] extends the result to generalized linear models.

3 Design of the Experiments

The design of our Monte Carlo experiments draws on the experiments performed in the research on propensity score estimation described above. The

difference between our experiments and the experiments previously performed lies with a variable Z that always remains unobserved. The experiments are designed to investigate the sensitivity of estimates of the average treatment effect (ATE) to a newly included variable W given that Z is never observed. Note that our design is innovative in two additional ways: we examine multiple estimators, and it is the first paper to examine the possibility of conditioning induced biased set up explicitly as potential outcomes (as opposed to graphical models).

The elements that are common to the first two experiments are listed below:

- a single variable $X \sim N(1, 1)$ that is observed;
- a single variable $Z \sim N(2, 1)$ that is always unobserved;
- a single variable $W \sim N(1, 1)$ that is added to the analysis;
- the correlation between Z and W is set at 0.25;
- the canonical correlation between $\mathbf{X}$ and $\mathbf{Z} = \begin{bmatrix} Z & W \end{bmatrix}$ varies (details below);
- the number of observations generated in each iteration is $N = 1000$.

Experiment 1: linear specification

In the first experiment, the equations describing the data-generating process for the potential outcomes and the log odds of treatment assignment are linear in all covariates.[3] For ease of exposition, observation subscripts are omitted in the equations below.

[3]We consider multiplicative and quadratic terms in experiment 2.

Outcome in treated state
$Y_1 = \beta_{10} + \beta_{11}X + \beta_{12}Z + \beta_{13}W + \epsilon_1$
Outcome in untreated state
$Y_0 = \beta_{00} + \beta_{01}X + \beta_{02}Z + \beta_{13}W + \epsilon_0$
Latent index function
$T^* = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 W + u$
Treatment indicator
$T = I(T^* > 0)$, where $I(\cdot)$ is the indicator function

The moving parts of the experiment include the coefficient on W in the propensity score, which varies between -0.5 and 0.5, $\gamma_3 \in \{-0.5, -0.25, 0, 0.25, 0.5\}$ and the coefficient on W in the outcome equations, which varies between -2 and 2, $\beta_{13} \in \{-2, 0, 2\}$. The canonical correlation between X and $\mathbf{Z}$ varies from low ($cc_{XZ} = 0.2$) to medium (0.5) to high (0.8). The error terms, ϵ_0 and ϵ_1, are normally distributed with mean 0 and variance 1 and are correlated at 0.25.[4] The error term u in the propensity score equation has a logistic distribution. γ_0 is chosen so that 25% of the observations are in the treatment group. The remaining coefficients are set at reasonable values.[5] We performed 500 replications of the experiment per parameter combination, and for each of those combination, we confirmed that the estimators we examine recover the treatment effect without bias when the correct propensity score specification is used.

An easy way to understand the various data generating processes (DGPs) used in the experiment is to look at the directed acyclic graphs (DAGs) in Figures 7-9 (see the Appendix). In each graph, an arrow indicates variables that affect one another; dashed arrows indicate unobserved variables. In Figure 7, W affects both the treatment T and the outcome Y. In Figure 8, W is related only to the treatment, which corresponds to the case where $\beta_{13} = 0$. In Figure 9, W is related only to the outcome, which corresponds to the case where $\gamma_3 = 0$.

The key to this experiment is the difference in the estimated ATE between the two misspecified models depicted in Figures 10-11. In the first

[4] We also ran the experiment with ϵ_0 and ϵ_1 uncorrelated; the results were essentially identical.

[5] $\gamma_1 = 0.5$, $\gamma_2 = 1$, $\beta_{00} = 1$, $\beta_{01} = 1$, $\beta_{02} = 1$, $\beta_{10} = 2.5$, $\beta_{11} = 1.5$, and $\beta_{12} = 0.5$. Extensive robustness checks were performed; see the Results section.

misspecified model (Figure 10), both Z and W are unobserved. In the second misspecified model (Figure 11), W is included in the propensity score while Z remains unobserved. The question is whether the bias on the ATE in the second misspecified model is ever greater than the bias on the ATE in the first. If so, then there are cases where controlling for all observed covariates is incorrect advice.

Experiment 2: nonlinear specification

In the second experiment, we introduce multiplicative and quadratic terms into the outcome equations.

Outcome in treated state
$Y_1 = \beta_{10} + \beta_{11}X + \beta_{12}X^2 + \beta_{13}Z + \beta_{14}Z^2 + \beta_{15}XZ + \beta_{16}W + \epsilon_1$
Outcome in untreated state
$Y_0 = \beta_{00} + \beta_{01}X + \beta_{02}X^2 + \beta_{03}Z + \beta_{04}Z^2 + \beta_{05}XZ + \beta_{16}W + \epsilon_0$
Latent index function
$T^* = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 W + u$
Treatment indicator
$T = I(T^* > 0)$, where $I(\cdot)$ is the indicator function

As before, the moving parts of the experiment are the coefficient on W in the propensity score, $\gamma_3 \in \{-0.5, -0.25, 0, 0.25, 0.5\}$; the coefficient on W in the outcome equations, $\beta_{16} \in \{-2, 0, 2\}$; and the canonical correlation between $\mathbf{X}$ and $\mathbf{Z}$, $cc_{XZ} \in \{0.2, 0.5, 0.8\}$. The error distributions are the same as in experiment 1, and the non-moving parameters are set at reasonable values. [6] We again performed 500 replications of the experiment per parameter combination.

Experiment 3: real data

In the first two experiments, all of the variables are drawn from a normal distribution—an ideal case for common matching methods [13], but one that

[6] $\gamma_1 = 0.5$, $\gamma_2 = 1$, $\beta_{00} = 1$, $\beta_{01} = 1$, $\beta_{02} = -2$, $\beta_{03} = 1$, $\beta_{04} = 0.2$, $\beta_{05} = 0.25$, $\beta_{10} = 2.5$, $\beta_{11} = 1.5$, $\beta_{12} = -1$, $\beta_{13} = 0.5$, $\beta_{14} = 0.4$, and $\beta_{15} = 0.5$.

hardly resembles a typical data set. To assess the relative performance of the propensity score estimators in a more realistic setting, we ran a third experiment with LaLonde's [1986] widely used data set.[7] As in the prior experiments, we simulate the treatment and outcome, as well as the new variable being added to the propensity score equation, but the other included and omitted variables are now taken from real data. The variables we use are *Age* (in years), *Education* (in years), ln(1+*Earnings 1974*) (in logged USD), and ln(1+*Earnings 1975*) (in logged USD). The last of these is always omitted from the propensity score specification, while the other three are always included. In each trial, we generate a variable W (mean 1, variance 1) that is correlated with the omitted variable, ln(1+*Earnings 1975*), at the level $\rho \in \{-0.75, -0.25, 0, 0.25, 0.75\}$. The data generating process follows.

Outcome in treated state
$Y_1 = 2 + 0.1\,Age + 0.15\,Educ + 0.4\,\ln(1 + Earn74) + 0.4\,\ln(1 + Earn75) + \beta W + \epsilon_1$
Outcome in untreated state
$Y_0 = 1 + 0.05\,Age + 0.1\,Educ + 0.2\,\ln(1 + Earn74) + 0.2\,\ln(1 + Earn75) + \beta W + \epsilon_0$
Latent index function
$T^* = \gamma_0 + 0.03\,Age + 0.14\,Educ + 0.12\,\ln(1 + Earn74) + 0.12\,\ln(1 + Earn75) + \gamma_1 W + u$
Treatment indicator
$T = I(T^* > 0)$, where $I(\cdot)$ is the indicator function

The other moving parts are once again the coefficient of W in the true propensity score, $\gamma_1 \in \{-0.3, -0.15, 0, 0.15, 0.3\}$, and its coefficient in the outcome equation for treated units, $\beta \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The intercept in the propensity score equation, γ_0, is chosen so that 25% of observations are treated. The error terms in the outcome equations, ϵ_0 and ϵ_1, have a standard deviation of 3 and are correlated at 0.25. As in the prior experiments, our interest focuses on whether including W in the propensity score specification decreases the bias and mean squared error of the estimated treatment effect.

[7] Following Dehejia and Wahba (1999), we use the subset of male participants for which 1974 earnings are available, giving us 445 observations.

Techniques

The five different techniques used for estimating the ATE in our experiments are described briefly below. The matching techniques were performed using the Matching package in R [15]. We coded the remaining techniques in R.

- Nearest-neighbor matching (with replacement): In this method, all the units are ordered randomly, and the first treated unit is matched with the control unit having the nearest propensity score. The first treated unit is then removed from the data set while its matched control unit is kept to be used in future matches. The process is repeated for the second treated unit and so on. The causal effect is estimated by averaging the outcome differences between the matched treatment and control groups. Following Rosenbaum and Rubin [16, 36], we use the linear predictor in place of the estimated probability to avoid compression of the propensity scores near 0 and 1.

- Caliper matching (with replacement): In this method, all the units are ordered randomly, and for the first treated unit, the control units with propensity scores (again the linear predictor) within a specified distance of the treated unit are gathered (0.25 standard deviations of the linear predictor), and the treated unit is matched with the closest control unit within the group in terms of Mahalanobis distance. The first treated unit is then removed from the data set while its matched control unit is kept to be used in future matches. The process is repeated for the second treated unit and so on. Again, the causal effect is estimated by averaging the outcome differences between the matched treatment and control units.

- Blocked matching: In this method, all the observations are divided into blocks, or strata, based on the value of the propensity score, which should be approximately constant within the strata. (We used deciles.) The difference of means between the treated and the control is estimated within each block, and the estimated causal effect is the weighted mean of these differences.

- Weighting: Wooldridge [17, 616] provides a consistent estimator of the ATE based on simple weighting that is identical to the Horvitz-

Thompson estimator [18],[8]

$$\begin{aligned}\widehat{ATE} &= N^{-1}\sum_{i=1}^{N}\left\{\frac{[T_i-\hat{e}(\mathbf{X}_i)]Y_i}{\hat{e}(\mathbf{X}_i)[1-\hat{e}(\mathbf{X}_i)]}\right\} \\ &= N^{-1}\sum_{i=1}^{N}\left\{\frac{T_iY_i}{\hat{e}(\mathbf{X}_i)}-\frac{(1-T_i)Y_i}{1-\hat{e}(\mathbf{X}_i)}\right\}.\end{aligned}$$

- Covariance adjustment: In this method, the ATE is estimated from a regression of the response variable on a constant, a variable denoting treatment assignment, the estimated propensity score, and a multiplicative term comprising the treatment variable and deviations about the sample mean of the estimated propensity score. Specifically, Rosenbaum and Rubin [19, 46] demonstrate, assuming that $E[Y_0|e(\mathbf{X})]$ and $E[Y_1|e(\mathbf{X})]$ are linear in $e(\mathbf{X})$, that the estimated coefficient on the multiplicative term, $\hat{\beta}_1$ in the equation below, is a consistent estimator of the ATE,

  $$Y_i = \beta_0 + \beta_1 T_i + \beta_2\hat{e}(\mathbf{X}_i) + \beta_3 T_i\left[\hat{e}(\mathbf{X}_i) - \hat{\mu}_{e(\mathbf{X})}\right] + \epsilon_i,$$

  where $\hat{\mu}_{e(\mathbf{X})}$ is the sample average of the estimated propensity score, $\hat{e}(\mathbf{X})$.

In each Monte Carlo iteration, we obtain two different estimates of the propensity score and run each of the five techniques with both estimates. The first is a specification using all available covariates (recall that we assume Z is unobserved):

$$\hat{e}_1(\cdot) = \Lambda\left(\hat{\gamma}_0^1 + \hat{\gamma}_1^1 X_i + \hat{\gamma}_2^1 W_i\right),$$

where $\Lambda(\cdot)$ is the logistic CDF and the coefficients $\hat{\gamma}^1$ are estimated via maximum likelihood. The second is a specification where W is omitted,

$$\hat{e}_0(\cdot) = \Lambda\left(\hat{\gamma}_0^0 + \hat{\gamma}_1^0 X_i\right).$$

[8] $\mathbf{X}$ here is the subset of observed covariates chosen to condition on.

We denote the ATE estimated from the first and second specifications as $\hat{\tau}_1$ and $\hat{\tau}_0$ respectively. Our interest is in whether there are parameter combinations under which $\hat{\tau}_1$ is more biased than $\hat{\tau}_0$, meaning it is better not to condition on all observed covariates.

4 Results

Experiment 1: linear specification

[Figure 2 about here.]

We first examine the results of the experiment in which all covariates enter the outcome equations linearly.[9] Figure 2 summarizes the main findings. Each point in a subplot represents the difference in absolute bias on the estimated ATE between the model that included W in the propensity score and the model that did not, as a function of γ_3 (W's coefficient in the true propensity score equation). Each subplot represents one combination of the other varying parameters: cc_{XZ}, the canonical correlation between $\mathbf{X}$ and $\mathbf{Z}$, and β_{13}, the coefficient on W in the outcome equation. Each line within a subplot represents the quantity of interest estimated with a specific method. Positive values indicate that the absolute value of the bias is greater when W is included in the estimated propensity score equation than when it is not. For example, the subplot in the top left corner presents the difference in absolute bias when $\beta_{13} = -2$ and $cc_{XZ} = 0.2$. In this example, intermediate values of γ_3 result in the bias increasing when W is included in the estimated propensity score, regardless of which method is used to estimate the ATE. (We ran a similar set of experiments in which the estimand was the average treatment effect on the treated (ATT), and obtained substantively identical results.)

Two observations are apparent. First, the absolute bias on the estimated ATE increases when W is included in the estimated propensity score equation for several parameter combinations. That is, it is *not* always optimal to condition on all available pre-treatment variables. Specifically, we find that when β_{13} is positive but γ_3 is equal to -0.5, it is often worse to include

[9] A replication file is available upon request.

W in the propensity score. The same is true if W's effect on the outcome is negative but its effect on the probability of treatment is weakly positive. This pattern is stronger when the canonical correlation between X and Z is low. Notice also that for all values of the canonical correlation, whenever there is no effect of W on the outcome, we see that there is not much difference between the bias when the propensity score model includes the variable and when it does not. If anything, when γ_3 is negative and $\beta_{13} = 0$, it is slightly worse to include the variable.

[Figure 3 about here.]

The second observation is that the five estimation procedures generally agree regarding whether including W in the propensity score equation worsens the bias on the ATE. Figure 3 shows the bias when W is excluded from the estimated propensity score and when it is not for a low canonical correlation between X and Z. When W is not included, all five methods return nearly identical estimates but the variation between methods is greater when W is included.[10] It appears, that weighting and nearest neighbor matching perform slightly better at reducing the bias when W is included; however, as we will see later, this finding is not robust to changes implemented in other experiments.

[Figure 4 about here.]

We can also see how the inclusion vs. exclusion of W from the propensity score equation affects the root mean squared error of the ATE estimates. Figure 4 presents plots for the difference in the root mean squared error (RMSE) of these experiments. The plots keep the same basic structure of the ones in Figure 2. We find similar results. When the coefficient of W in the treatment and outcome equations have opposite signs, the treatment effect is weak, and the canonical correlation is low, it is worse to include W. In addition, if the outcome effect is null and the treatment effect is strong and negative, including W slightly increases the RMSE.

If the magnitude of the increment in bias or RMSE were negligible for the conditions that we have characterized, we could be confident that adding all

[10]This is true accounting for the fact that the figure has different y scales for both specifications of the propensity score.

available pre-treatment variables would not seriously affect the conclusions derived by the results of our empirical analysis. Unfortunately, this is not always the case. If we consider the parameter combination of $\beta_{13} = 2$, $\gamma_3 = -0.5$ and $cc_{XZ} = 0.2$, the bias when W is included is approximately 0.53, compared to 0.05 when it is excluded. Including W increases by more than 10 times the bias for this parameter combination. Moreover, the increment in the bias represents 16% of the total ATE (which is 3 for this parameter combination).

[Table 1 about here.]

To confirm that the relationships between the sign and magnitude of the coefficients in the propensity and outcome equations are indeed driving the results, we ran a number of additional Monte Carlo experiments. We ran the baseline model with each of the following changes (separately):

- Coefficient on Z in the propensity score equation reversed to -1;
- Coefficients on Z in the outcome equations, Y_0 and Y_1, reversed to -1 and -0.5 respectively;
- Correlation between W and Z reversed to -0.25.

In the first two cases, the results were nearly a mirror image of those shown in Figure 2: adding the new variable W increased the bias mainly in cases where it had the same effect on the treatment assignment and outcome.[11] In the third case, with the correlation between Z and W reversed, the results were substantively the same as in the original experiment. Our findings are summarized in Table 1, which describes the conditions or patterns of signs of effects under which adding a variable W to the propensity score equation increases the bias in the ATE when another variable Z is unavailable.

The cases where including W in the estimated propensity score equation increases the bias are those where balancing on W and balancing on Z have countervailing effects (patterns 2 and 3 in Table 1), and the effect of Z's

[11] In particular, these were $(\gamma_3, \beta_{13}) = (-0.5, -2)$ and $(0.5, 2)$, whereas in the original model the cases where including W increased bias were $(0.5, -2)$ and $(-0.5, 2)$.

confounding is stronger than that of W's. In this experiment, Z has a positive effect on the probability of treatment and a positive direct effect on the outcome, meaning the ATE would be overstated if only Z were omitted. If W has a negative effect on treatment assignment but a positive direct effect on the outcome, then W's omission causes the ATE to be understated. In this situation, balancing on Z causes the estimated ATE to decrease, and balancing on W causes an increase. Suppose the confounding effect of Z is larger than that of W, so the estimated ATE is too high when both are omitted. Including W in the propensity score specification would further increase the estimated ATE, exacerbating the bias due to the omission of Z.[12] This is precisely what we observe in Figure 2 when $\beta_{13} = 2$ and $\gamma_3 = -0.5$.

The patterns that we have consistently found in which adding a relevant pre-treatment variable increases the bias of the ATE could easily occur in empirical applications. Imagine an observational study on whether methamphetamine use T increases an individual's risk of heart disease Y. Suppose data are available on whether each individual is white W, but not on their household income Z. Compared to racial minorities, whites are more likely to be methamphetamine users and less likely to have heart disease, so the ATE would be underestimated if only W were left out. Conversely, high-income individuals are both less likely to use methamphetamines and to have heart disease, so the omission of Z causes the ATE to be overestimated. This corresponds to pattern 2 in Table 1: If the confounding effect of income is stronger than that of race, controlling for race when income data are unavailable will likely increase the bias.

Experiment 2: nonlinear specification

[Figure 5 about here.]

Figure 5 presents the difference in absolute biases as a function of W's coefficient in the true propensity score when we add interactions and quadratic terms to the true outcome equation as specified in the previous section.[13]

[12] The only exception is if W and Z are so strongly positively correlated that balancing on W substantially improves balance on Z.

[13] We present results only for a canonical correlation of 0.2, cc_{XZ}=0.2. The results for the other correlations are nearly identical.

We again find that including W is worse in terms of bias when the signs of β_{13} and γ_3 are different, but this time the result holds regardless of the value of the canonical correlation. The second pattern found in the linear outcome equation case is still present: when γ_3 is negative and $\beta_{16} = 0$, including W increases the ATE's bias. The most noticeable differences between these results and those from the first experiment are that including W in the propensity score equation increases the bias in more cases, and that the magnitude of these increases is higher than those from before.

As in the previous experiment, the differences across estimation methods are slight. Figure 5 shows that there is no case where including W in the propensity score equation increases the bias when one method is used but decreases it under other methods. Finally, the results for the RMSE in the nonlinear case are almost identical to those for the difference in absolute bias, and thus are omitted.

Experiment 3: real data

[Figure 6 about here.]

Results from the final experiment, in which the independent variables are taken from LaLonde's data rather than being simulated, are summarized in Figure 6. These results are similar to those of the first two experiments, though the pattern identified in Table 1 is somewhat weaker. For example, take the fourth column of Figure 6, where $\beta = 0.25$ (i.e., the new variable has a positive effect on the outcome). When the correlation between W and the omitted variable ($\ln(1+Earn75)$) is 0.25, including W in the estimated propensity score equation increases the bias when W's true effect on receiving the treatment is negative. Since the omitted variable has a positive effect on both assignment and outcome, this corresponds to the second row of Table 1. There are some anomalies when the magnitude of the correlation between W and Z is high. For example, when $\beta = -0.25$, inclusion of W in the propensity score specification when γ_1 is positive exacerbates bias for all ρ except 0.75. In this case, conditioning on W increases balance on Z enough to offset any potential exacerbation of the bias.

The differences between estimation methods are still relatively small in this experiment, but they are more noticeable than in the first two exper-

iments. We find that including W in the propensity score equation is less likely to increase bias under caliper matching than under other methods. Covariance adjustment (regression on the propensity score) also seems to perform well with the Lalonde data, which was not the case in the fully simulated experiments. When using covariance adjustment to estimate the ATE, including W in the estimated propensity score equation increases the magnitude of bias in 40 cases, out of 125 total combinations of the parameters (β, γ, and ρ). That figure increases to 58 cases under nearest-neighbor matching; in between are caliper matching (46), weighting (53), and blocking (55).

Balance tests

The previous experiments identified situations where conditioning on all available pre-treatment variables could lead to an increase in bias on the estimated ATE (or ATT). The results show that when considering whether to include a pre-treatment variable in the propensity score equation a researcher should take into account the potential effect of unobservables on treatment and outcomes, as well as their relation with the variable in question. At this point it is natural to ask whether a post-matching balance test on treated and untreated units could be thought of as an alternative way to identify whether a covariate should be included in the propensity score equation. If balance on matched units improves once the variable is included in the propensity score, the inclusion might be justified. Improving balance, however, does not necessarily mean reducing bias. Adopting this practice could lead to worse estimation results. Using the Lalonde data, we show that it is not uncommon to find situations where a post-matching test indicates that a candidate variable should be included, when in fact its inclusion worsens the bias. What this result suggests is that balance tests are not substitutes for careful consideration of the potential effects of unobservables when choosing a specification.

For the following exercises, we use Lalonde's subset of treated units and the CPS (Current Population Survey) control individuals from the well known study of the impact of the NSW labor training program on post-intervention income [14].[14] This data set has been used to evaluate the per-

[14]The data set is included in the R package of Random Recursive Partitioning [20]. It

formance of treatment effect estimators on observational data by comparing them with an experimental benchmark. We are interested in finding pairs of variables from this data set that would play the role of Z and W in our previous experiments and that simultaneously satisfy three conditions: 1) They have one of the countervailing effects identified in Table 1, 2) balance seems to improve with W's inclusion in the propensity score according to a post-matching balance test, and 3) the estimated ATT is more biased when W is included. Identifying such cases gives evidence of the risks of relying exclusively on balance tests, and the potential benefit of using the identified countervailing patterns in justifying the inclusion of additional variables on the propensity score. Given that we are using observational data, we do not know the true effects of W and Z on the outcome and treatment. We generate estimates of these effects by relying on the propensity score specification that in Dehejia and Wahba [21] gave the authors the closest ATT estimate to the experimental benchmark when using the CPS control individuals.[15]

To further clarify the exercise, consider the pair of variables $re75$ (real earnings in 1975) and *nodegree* (indicator variable of not possessing a degree) and suppose that the first one takes the role of Z (the unobserved variable) and the second one is the candidate variable to be included W. This would represent a situation where a researcher interested in finding the effect of the training program on income is considering to include the indicator of no degree in the propensity score equation when past earnings are not available. Following the Dehejia and Wahba's specification that gives the best estimate of the ATT with this data, it is found that the effect of $re75$ and *nodegree* on income are both positive and that the effect of $re75$ on treatment assignment is negative while this same effect for *nodegree* is positive. This example follows pattern 3 in Table 1. For this specific case, and using as a benchmark the experimental ATT, we calculate that the bias after including *nodegree* (while leaving out $re75$) increases by 150.71, which is approximately 8.5% of the experimental ATT.[16]

---

has a total of 16177 observations with no missings on the variables from the original study. For a description of the data set see Dehejia and Wahba [21, 1054]

[15] The specification includes the following variables: *age, age*^{2}*, education, education*^{2}*, no degree, married, black, hispanic, real earnings in 1974 and 1975 (re74 and re75), indicators of zero earnings in 1974 and 1975 (u74 and u75), education* $\times$ $re74$ *and age*^{3}.

[16] The ATT was calculated using caliper matching using the same estimation set up used in the Monte Carlo experiments.

We also compared the results of a balance test after matching when *nodegree* was not included in the treatment equation with the results of the same test when it was. We found that after including it, the p-value of a t-test between the treated and untreated matched units increased for 6 variables (out of a total of 11 covariates) compared to the p-value of the same test when it was left out of the propensity score. This result suggests that in more than half of the rest of the covariates, balance seems to have improved after matching when *nodegree* was included. In this example, a researcher expecting a positive relation between past earnings and income, could infer with the help of Table 1 that the inclusion of this variable could bring an increase in bias if there is a positive relation between *nodegree* and income once other variables are controled for in the outcome equation.

The case of $re75$ and *nodegree* is not the only one. Table 2 shows 14 other cases where a countervailing effect is present and including W improves balance for a given number of covariates, but increases bias. These results are not definitive as we have relied on a particular specification from Dehejia and Wahba to find estimates of the true effects on income and participation on the program. However, given that the bias increased in the situations where we expected it to happen, based on our Monte Carlo findings, we can be more confident that the results are not unique to this data set or this particular specification.

[Table 2 about here.]

Sensitivity analysis

We ran additional simulations to determine whether researchers can use sensitivity analysis to determine when a particular covariate's inclusion would increase bias in the estimated ATE. We use Rosenbaum's (2002) method for calculating bounds on the estimate when the log odds of treatment assignment differ by up to Γ due to unobserved confounding.[17] Using the same parameters as in the simulations described above (with both the fully simulated data and the Lalonde data), we calculated the average lowest level of Γ at which the bounds on the treatment effect contained 0, under both inclusion and exclusion of W from the propensity score equation. In all cases,

[17] To calculate the bounds, we used the R package `rbounds`.[22]

this level is almost entirely determined by the magnitude of the estimated ATE. There appears to be no particular relationship between the level of Γ and whether including W increases bias, except insofar as the bias is toward or away from $\hat{\tau} = 0$. Therefore, this kind of sensitivity analysis is not useful for determining whether it is worse to include an observed pre-treatment variable because of its interaction with an unobserved confounder.

5 Discussion

This paper investigates claims made by both sides in recent debates regarding conditioning and matching using propensity scores. The results of our experiments suggest that conditioning on all available pre-treatment variables is not always optimal. In every case, the researcher must consider the effects of unobserved pre-treatment variables and their relationships with observed pre-treatment variables. Whether conditioning on an additional observed pre-treatment variable increases or decreases the bias on the ATE depends on these relationships. Specifically, in the linear case, we show that when the newly included covariate has a positive effect on the outcome and a negative effect on the propensity (and when there is an unobserved covariate whose effects on the outcome and treatment have the same sign), it is often worse to include the covariate. This basic pattern also holds in nonlinear specifications and in simulations using real data.

We have yet to address how researchers can best make use of our findings. Our results suggest that researchers cannot rely on advice such as condition on all pre-treatment covariates or on balance and sensitivity tests. Some progress can be made if we consider the two kinds of unobserved covariates that plague empirical analyses. To paraphrase Donald Rumsfeld [23], there are known unknowns and unknown unknowns. That is, there are covariates, perhaps suggested by theory, that cannot be measured or perhaps measurement is infeasible. These are the known unknown covariates. A researcher can hypothesize about the relationships of such a covariate with previously included variables and any variables that are candidates for inclusion. Our results provide some guidance in such a situation. If the candidate covariate and the unobserved covariate have countervailing effects, a case can be made for leaving the candidate covariate unadjusted.

On the other hand, there exist, in Rumsfeldian terms, unknown unknown covariates. These are variables that have not been suggested by theory and have not crossed the mind of the researcher in question (or anyone else). In such a case, no theorizing can take place, and our results demonstrate that including a new covariate in a propensity score equation may increase or decrease the bias on the estimated ATE. Sensitivity analysis that explicitly takes unobserved covariates into account, e.g. Rosenbaum [24], seems to be of little use. The only surefire response a researcher has to the problem discussed in this paper is to be modest in the claims she makes based on her results. Scientific progress is rarely the result of a single study, and empirical generalizations are accepted only after many repeated demonstrations across varying spatial and temporal domains.

Appendix

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

References

[1] Ian Shrier. Letter to the editor. *Statistics in Medicine*, 27(14):2740–2741, June 2008.

[2] Judea Pearl. Remarks on the method of propensity score. *Statistics in Medicine*, 28(9):1415–1416, April 2009.

[3] Arvid Sjölander. Propensity scores and m-structures. *Statistics in Medicine*, 28(9):1416–1420, April 2009.

[4] Donald B. Rubin. The design *versus* the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):78–97, January 2007.

[5] Donald B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, April 2009.

[6] Judea Pearl. Myth, confusion, and science in causal analysis. Technical Report R-348, University of California, Los Angeles, CA, 2009.

[7] Christiana Drake. Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics*, 49(4):1231–1236, December 1993.

[8] Boris Augurzky and Christoph M. Schmidt. The propensity score: A means to an end. Technical Report 271, The Institute for the Study of Labor, Bonn, Germany, March 2001.

[9] M. Alan Brookhard, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12): 1149–1156, June 2006.

[10] Daniel L. Millimet and Rusty Tchernis. On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business and Economic Statistics*, 27(3):397–415, July 2009.

[11] Kevin A. Clarke. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352, September 2005.

[12] Kevin A. Clarke. Return of the phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 26 (1):46–66, February 2009.

[13] Donald B. Rubin and Neal Thomas. Affinely invariant matching methods with ellipsoideal distributions. *Annals of Statistics*, 20(2):1079–1093, June 1992.

[14] Robert LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4): 604–620, September 1986.

[15] Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, Forthcoming.

[16] Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, February 1985.

[17] Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA, 2002.

[18] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, Decemeber 1952.

[19] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effect. *Biometrika*, 70(1):41–55, April 1983.

[20] S.M. Iacus. *rrp: Random Recursive Partitioning*, 2007. URL `http://cran.r-project.org/web/packages/rrp/`. R package version 0.7.

[21] R. Dehejia and S. Wahba. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

[22] Luke J. Keele. *rbounds: Perform Rosenbaum bounds sensitivity tests for matched data.*, 2011. URL `http://CRAN.R-project.org/package=rbounds`. R package version 0.7.

[23] Errol Morris. The anosognosic's dilemma: Something's wrong but you'll never know what it is (part 1). *The New York Times*, June 20 2010.

[24] Paul R. Rosenbaum. *Observational studies*. Springer, New York, 2002.

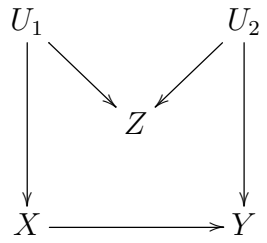


Figure 1: A causal directed acyclic graph of an M-structure.

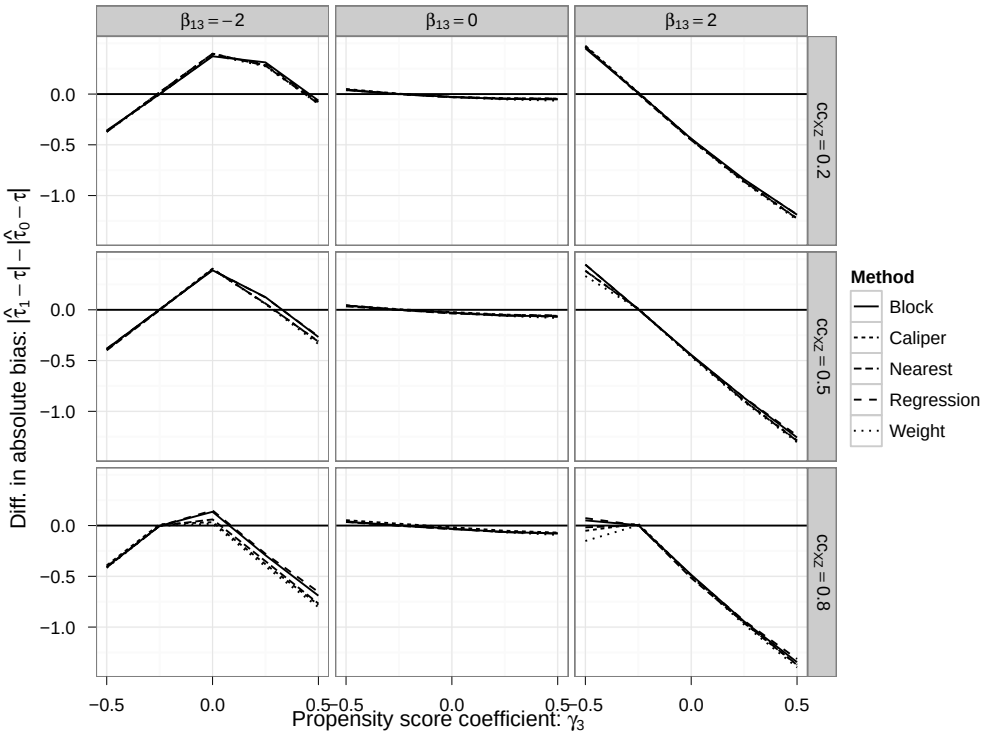
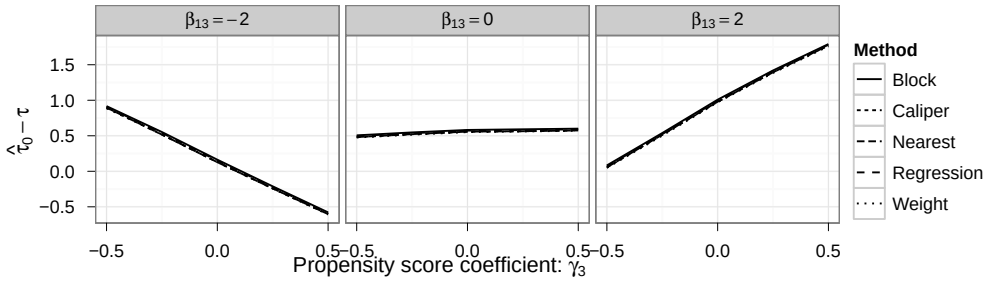
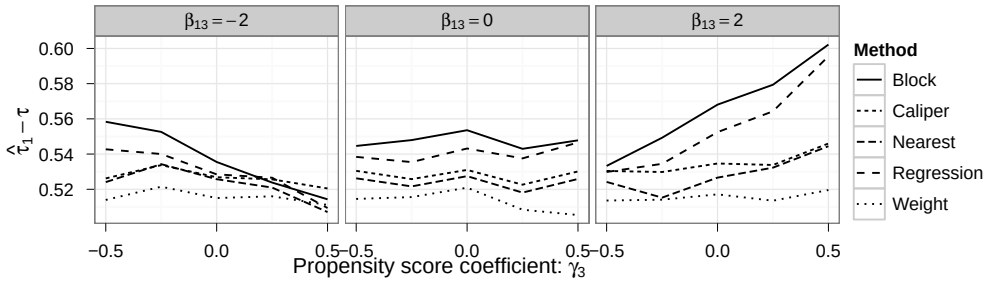


Figure 2: Difference in absolute bias with linear outcome equations.



(a) Bias when W is excluded.



(b) Bias when W is included.

Figure 3: Biases in the experiment with linear outcome equations (the canonical correlation is held at 0.2, $cc_{XZ} = 0.2$).

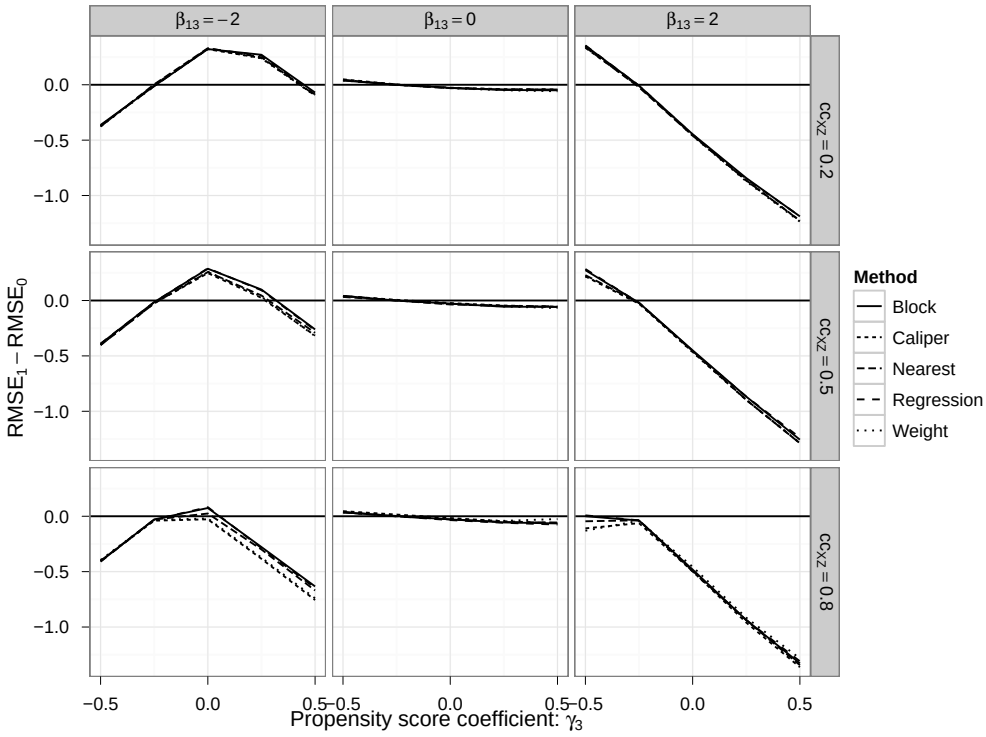


Figure 4: Difference in root mean squared error with linear outcome equations.

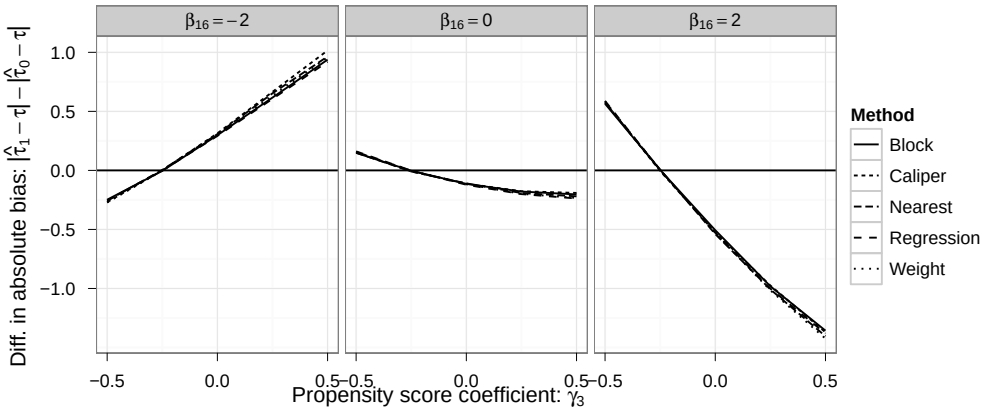


Figure 5: Difference in absolute bias with nonlinear outcome equations.

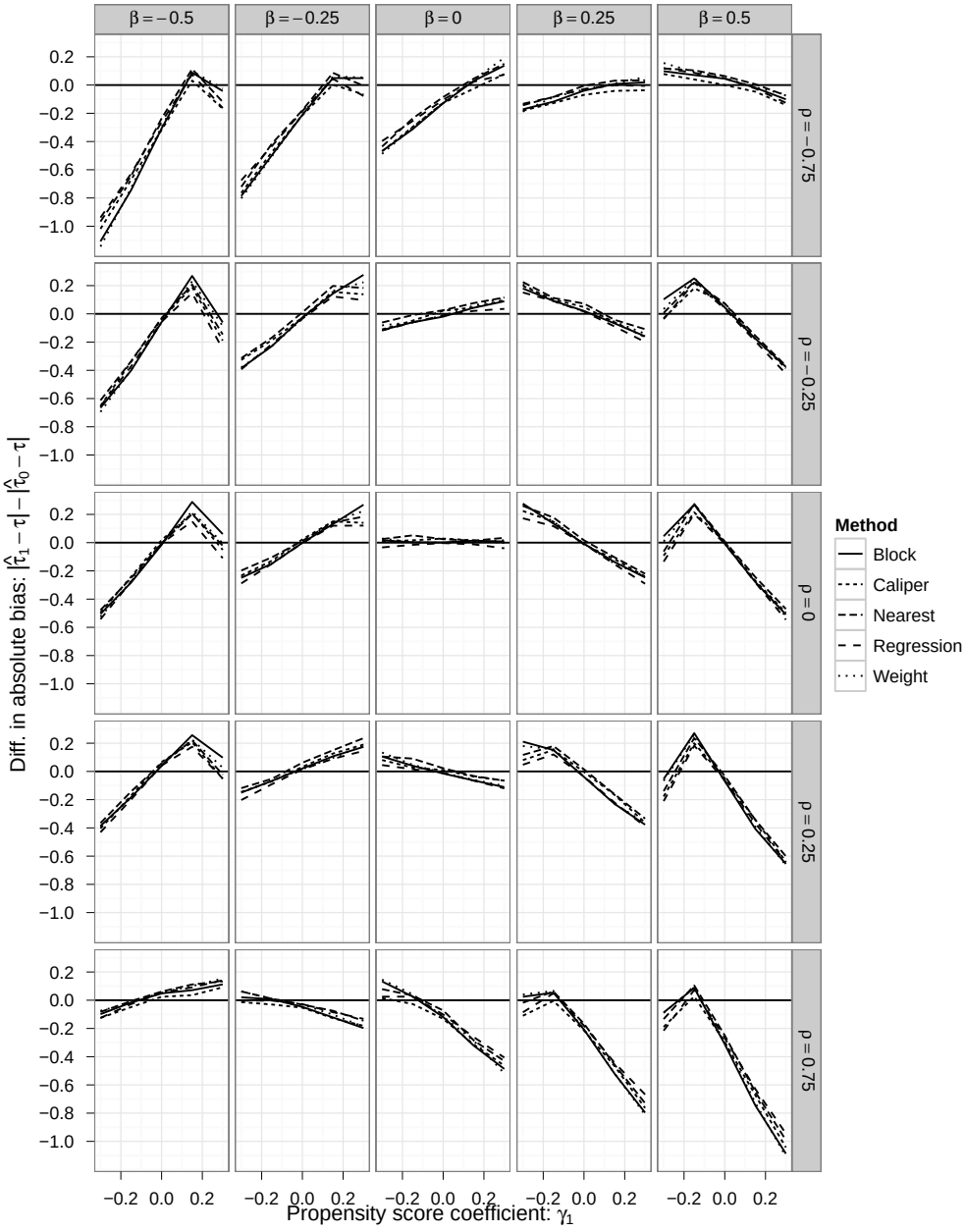


Figure 6: Difference in absolute bias with regressors taken from the LaLonde (1986) data.

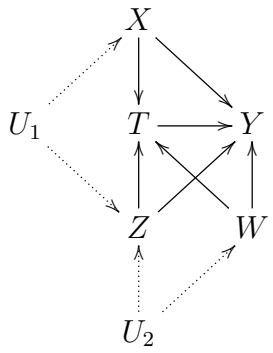


Figure 7: DGP 1 — W is related to both treatment and outcome.

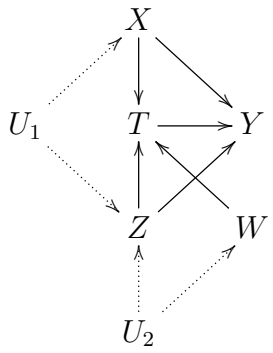


Figure 8: DGP 2 — W is related only to the treatment.

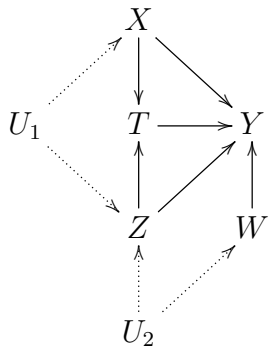


Figure 9: DGP 3 — W is related to the outcome.

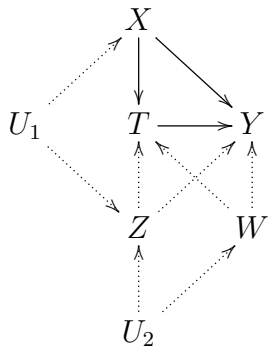


Figure 10: Misspecified model 1 — Z and W are unobserved.

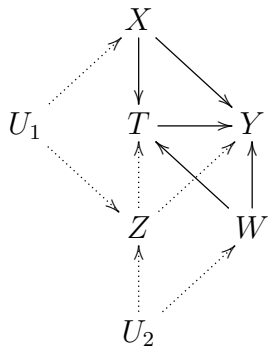


Figure 11: Misspecified model 2 — Z is unobserved, and W is assumed related to treatment and outcome.

Pattern	$\gamma_2 \cdot \beta_{12}$	$\gamma_3 \cdot \beta_{13}$	Including W increases bias?
1	+	+	no
2	+	-	yes
3	-	+	yes
4	-	-	no

Table 1: When does including a new variable W in the propensity score equation increase the bias of the estimated average treatment effect? (γ_2 and β_{12} are the coefficients on Z in the propensity and outcome equations, respectively. γ_3 and β_{13} are the coefficients on W in the propensity and outcome equations, respectively.)

Countervailing pattern	Z	W	Δ Bias	$\frac{\text{Vars. p-val increased}}{\text{Total}}$
2	*nodegree*	age^2	27.13	7/12
2	*education*	*age*	171.64	4/10
3	*black*	*nodegree*	112.75	8/12
3	*black*	*u*74	690.72	7/12
3	*u*75	*nodegree*	150.70	6/11
3	*age*	*education*	104.99	8/10
3	*age*	*u*74	94.37	8/10
3	age^2	*education*	104.99	8/10
3	age^2	*u*74	94.37	8/10
3	*black*	*education*	11.85	8/12
3	*married*	*nodegree*	107.37	9/12
3	*married*	*u*74	24.87	9/12
3	*re*75	*nodegree*	150.70	6/11
3	$education^3$	*education*	104.99	8/10
3	$education^3$	*u*74	94.37	8/10

Notes: The column 'Countervailing patterns' refers to the patterns defined in Table 1. $\frac{\text{Vars. p-val increased}}{\text{Total}}$ is the number of control variables that had an increase in the p-value of the t-test after matching when W was included over the number of common controls between the specifications with and without W.

Table 2: Balance test and propensity score selection of variables