An Introduction to the Augmented Inverse Propensity Weighted Estimator

Adam N. Glynn* Kevin M. Quinn†

October 21, 2009

Abstract

In this paper we discuss an estimator for average treatment effects known as the augmented inverse propensity weighted (AIPW). This estimator has attractive theoretical properties and only requires practitioners to do two things they are already comfortable with: (1) specify a binary regression model for the propensity score, and (2) specify a regression model for the outcome variable. After explaining the AIPW estimator, we conduct a Monte Carlo experiment that compares the performance of the AIPW estimator to three common competitors: a regression estimator, an inverse propensity weighted (IPW) estimator, and a propensity score matching estimator. The Monte Carlo results show that the AIPW estimator is dramatically superior to the other estimators in many situations and at least as good as the other estimators across a wide range of data generating processes.

*Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

†Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. kevin_quinn@harvard.edu

1 Introduction

In this paper we discuss an estimator for average treatment effects known as the augmented inverse propensity weighted (AIPW) estimator. While the basic ideas behind the AIPW estimator were developed by biostatisticians beginning almost 15 years ago (Robins et al., 1994; Robins, 1999; Scharfstein et al., 1999) the AIPW estimator is largely unknown and unused by social scientists. This is regrettable because the AIPW estimator has very attractive theoretical properties and only requires practitioners to do two things they are already comfortable with: (1) specify a binary regression model for the propensity score, and (2) specify a regression model for the outcome variable. To demonstrate that the large-sample theory behind the AIPW estimator carries over to finite samples we conduct a Monte Carlo experiment that compares the performance of the AIPW estimator to three common competitors: a regression estimator, an inverse propensity weighted (IPW) estimator, and a propensity score matching estimator. The Monte Carlo results show that the AIPW estimator is at least as good as the other estimators across a wide range of data generating processes and in many situations it is dramatically superior to the other estimators.

This paper is organized as follows. In Section 2 we briefly review estimators of average treatment effects, dividing the discussion into those estimators that primarily utilize regression models and those that focus on a model for treatment assignment. Section 3 introduces the AIPW estimator and discusses its usage. In Section 4 we present a Monte Carlo study comparing the performance of the AIPW estimator to other standard estimators of the Average Treatment Effect. Section 5 concludes.

2 Estimators for Average Treatment Effects Based on Regression Models or Treatment Assignment Models

Throughout this paper, we will assume that units (indexed by $i = 1, ..., n$) are randomly sampled from some population or superpopulation, that treatment is binary ($X_i \in \{0 \text{ (control)}, 1 \text{ (treatment)}\}$), and we observe an outcome variable Y_i. Furthermore, we assume that potential outcomes are defined as in Rosenbaum and Rubin (1983) such that $Y_i(1)$ is the outcome that we would observe if unit i had received treatment and $Y_i(0)$ is the outcome that we would observe if unit i had received treatment. We also assume

that the stable unit treatment value assumption (SUTVA) (Angrist et al., 1996) holds such that potential outcomes ($\{Y_i(1), Y_i(0)\}$) are completely determined and the observed outcome will be equal to the potential outcome corresponding to the assigned treatment,

$$Y_i = X_i \cdot Y_i(1) + (1 - X_i) \cdot Y_i(0).$$

Furthermore, we assume that a set of observed control variables $\mathbf{Z}$ exists such that strong ignorability holds given $\mathbf{Z}$ and the propensity score ($\pi(\mathbf{Z}) = \Pr(X = 1|\mathbf{Z})$) is strictly greater than zero and less than one over the support of $\mathbf{Z}$,

$$\{Y(1), Y(0)\} \perp\!\!\!\perp X|\mathbf{Z}$$

$$0 < \pi(\mathbf{Z}) < 1.$$

Using this framework, there are a number of reasonable estimators for the average treatment effect (ATE),

$$ATE = \mathbb{E}[Y(1) - Y(0)].$$

We summarize two broad classes of these estimators in the following subsections.

2.1 Estimators for Average Treatment Effects Based on Regression Models

Much of traditional causal estimation requires the formulation of a regression model for the outcome variable Y. In other words, estimation of the conditional expectation of Y given X and $\mathbf{Z}$: $\mathbb{E}(Y|X, \mathbf{Z})$. Given the stated assumptions of this paper, it has been shown that such a model can be used to identify the ATE through the backdoor adjustment (Pearl, 1995, 2000) or the g-functional (Robins, 1986),

$$ATE = \mathbb{E}\left[\mathbb{E}(Y|X = 1, \mathbf{Z}) - \mathbb{E}(Y|X = 0, \mathbf{Z})\right]$$

where the outer expectation is taken with respect to the distribution of $\mathbf{Z}$. The empirical distribution of the conditioning set provides an easy estimate of $F_{\mathbf{Z}}$ and simplifies integration, so that the corresponding regression estimator takes the form,

$$\widehat{ATE}_{reg} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\mathbb{E}}(Y|X = 1, \mathbf{Z}_i) - \hat{\mathbb{E}}(Y|X = 0, \mathbf{Z}_i) \right\}. \quad (1)$$

where $\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)$ is the estimated conditional expectation of the outcome given $\mathbf{Z}_i$ within the treated group, and $\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)$ is defined analogously. These conditional expectation functions can be estimated using any consistent estimator. Options include ordinary least squares, generalized linear models, generalized additive models, local regression, kernel regression, etc.

The regression estimator for ATE will be perform reasonably well when the estimated conditional expectation functions are good estimates of the true regression functions. However, when the conditioning set $\mathbf{Z}$ has many dimensions, it may be difficult to estimate both regression functions over the full range of $\mathbf{Z}$. In particular, when the observed values of $\mathbf{Z}$ are not similar for the treatment and the control groups, then one of the conditional expectation functions $\mathbb{E}(Y|X=1,\mathbf{Z}_i)$ or $\mathbb{E}(Y|X=0,\mathbf{Z}_i)$ will often be poorly estimated because of the lack of data points near either $(X=0,\mathbf{Z}_i)$ or $(X=1,\mathbf{Z}_i)$. Depending on the method of estimation, the estimation over such non-overlapping ranges may massively underestimate the uncertainty in this estimator and / or result in finite sample bias (King and Zeng, 2006). Further, many versions of the regression estimator for ATE tend to be quite sensitive to small amounts of misspecification. Given these deficiencies, many researchers have opted for methods of estimation that utilize models for treatment assignment instead of regression models for the outcome.

2.2 Estimators for Average Treatment Effects Based on Treatment Models

Another broad class of estimators explicitly or implicitly utilizes a model for treatment assignment instead a regression model for the outcome. If the true model for the probability of treatment assignment were known, then this could be used to define propensity scores for every unit, and these could be used for matching or weighting estimators. Because the treatment assignment model is usually unknown, matching and weighting estimators either explicitly estimate the propensity score function model, or utilize the treatment assignment model implicitly through notions of balance. If the propensity score model is estimated, a well known weighting estimator is the inverse propensity weighted (IPW) estimator,

$$\widehat{ATE}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right\}, \tag{2}$$

where $\hat{\pi}(\mathbf{Z}_i)$ is the estimated propensity score. If the propensity scores were known, then this estimator will be unbiased for the ATE (Tsiatis, 2006). Furthermore, when the propensity scores are estimated consistently, then this estimator is consistent for the ATE. However, the IPW estimator is also known to have poor small sample properties when the propensity score gets close to zero or one for some observations. This can be seen from (2), in that division by numbers close to zero will lead to high variance in the estimator. Specifically, units that receive treatment and very low propensity scores will provide extreme contributions to the estimate. Similarly, units that receive control and very high propensity scores will provide extreme contributions to the estimate. Due to these deficiencies, weighting estimators like (2) have fallen out of favor in relation to estimators that match treatment and control units based on estimate propensity scores or that directly balance $\mathbf{Z}$ between treatment and control units. See Rubin (2006) for a book length treatment on matching, or Diamond and Sekhon (2005) and Ho et al. (2007) for recent influential papers on matching in political science.

Despite the inefficiencies of the original IPW estimator, improvements can be made. In the next section, we introduce an augmented IPW estimator that utilizes the information in the conditioning set for the prediction of the outcome variable in order to improve the small sample properties of the IPW estimator.

3 An Augmented Inverse Propensity Weighted Estimator for Average Treatment Effects

One way the IPW estimator can be improved is by fully utilizing the information in the conditioning set. The conditioning set $\mathbf{Z}$ contains information about the probability of treatment, but it also contains predictive information about the outcome variable. The AIPW estimator $\widehat{ATE}_{AIPW}$ efficiently uses this information in the following manner:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{X_i Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right] - \frac{(X_i - \hat{\pi}(\mathbf{Z}_i))}{\hat{\pi}(\mathbf{Z}_i)(1-\hat{\pi}(\mathbf{Z}_i))}\left[(1-\hat{\pi}(\mathbf{Z}_i))\hat{\mathbb{E}}(Y_i|X_i=1,\mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y_i|X_i=0,\mathbf{Z}_i)\right]\right\}, \quad (3)$$

where the first line of (3) corresponds to the IPW estimator, and the second line adjusts this estimator by a weighted average of the two regression estimators. Note that this formula does not require the *same*

adjustment set $\mathbf{Z}_i$ to be used in both the propensity score model and the outcome model. All that is required is that conditional ignorability holds given $\mathbf{Z}$. For example, a regression model that includes the full set of variables $\mathbf{Z}_i$ might constrain the regression parameter for a particular variable to be zero. This also holds for the propensity score model. This flexibility allows the researcher to, for instance, use the minimal set of adjustment variables necessary for conditional ignorability to hold in the propensity score model while including a near maximal set of adjustment variables in the outcome regression models. In Section 4 we investigate the gains/losses that are incurred by such an approach.

This adjustment term in (3) has two properties that are easily deduced from the formula. First, the adjustment term has expectation zero when the estimated propensity scores and regression models are replaced with their true counterparts (see Appendix A). Second, the adjustment term stabilizes the estimator when the propensity scores get close to zero or one. This can be seen if we examine the left hand side of Equation (3) when $X_i = 1$:

$$\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{1}{\hat{\pi}(\mathbf{Z}_i)}\left[[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)\right] = \left[\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}\right] - \hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i) \tag{4}$$

and when $X_i = 0$:

$$-\frac{Y_i}{1-\hat{\pi}(\mathbf{Z}_i)} + \frac{1}{[1-\hat{\pi}(\mathbf{Z}_i)]}\left[[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)\right] = \hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i) - \left[\frac{Y_i}{1-\hat{\pi}(\mathbf{Z}_i)} - \frac{\hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)}{1-\hat{\pi}(\mathbf{Z}_i)}\right]. \tag{5}$$

Looking at Equation (4) we see that when $\hat{\pi}(\mathbf{Z}_i)$ is close to zero $\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)}$ will get large in absolute value. However, the $\frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}$ term gets large at the same rate and the term in brackets is stabilized. When $\hat{\pi}(\mathbf{Z}_i)$ approaches one the $\frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}$ term goes to zero and the term in brackets approaches Y_i. Inspection of Equation (5) reveals similar relationships when $X_i = 0$.

$\widehat{ATE}_{AIPW}$ has a number of very attractive theoretical properties. This estimator can be shown to be asymptotically normally distributed and valid large sample standard errors can be derived through the theory of M-estimation. Lunceford and Davidian (2004) find an empirical sandwich estimator to work well

in practice. This empirical sandwich estimator of the sampling variance of $\widehat{ATE}_{AIPW}$ is $\hat{\mathbb{V}}\left(\widehat{ATE}_{AIPW}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\hat{I}_i^2$ where:

$$\hat{I}_i = \left[\frac{X_i Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right] - \frac{(X_i - \hat{\pi}(\mathbf{Z}_i))}{\hat{\pi}(\mathbf{Z}_i)(1-\hat{\pi}(\mathbf{Z}_i))}\left[(1-\hat{\pi}(\mathbf{Z}_i))\hat{\mathbb{E}}(Y_i|X_i=1,\mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y_i|X_i=0,\mathbf{Z}_i)\right] - \widehat{ATE}_{AIPW}.$$

$\widehat{ATE}_{AIPW}$ will be unbiased for ATE when both the propensity score model and the outcome models are known and consistent for ATE when the propensity score and the outcome regressions are consistently estimated.[1] When the propensity score and the regression function are modeled correctly, the AIPW achieves the semiparametric efficiency bound. Most interestingly, $\widehat{ATE}_{AIPW}$ is *doubly robust* in that it will be consistent for ATE whenever (1) the propensity score model is correctly specified *or* (2) the two outcome regression models are correctly specified (Scharfstein et al., 1999).[2] This double robustness property gives the AIPW estimator a tremendous advantage over most other estimators in that with the AIPW estimator the researcher has more hope of getting a reassonable answer in complicated real-world situations where there is uncertainty about both the treatment assignment process and the outcome model. We refer the reader to Tsiatis (2006) for a textbook treatment of the theory behind the AIPW estimator as well as related estimators.

Of course the drawback of this estimator is that one must estimate a propensity score model and two regression models (one for treatment and one for control). Nonetheless, most rearchers are already comfortable with fitting regression models for the propensity score and the outcome variable. Further, because one only needs predictions from these models, flexible routines can be used (we utilize generalized additive models (GAMs) in this paper).

Due to the good theoretical properties of the AIPW estimator, there is some hope that the estimator will perform well in small samples. In the next section we investigate bias and efficiency for the AIPW and compare this performance to other standard regression and matching estimators of ATE under a variety of conditions.

[1] See Appendix A.1 for a proof of this result.

[2] See Appendix A.2 for a proof of this result and see Ho et al. (2007) for a related but distinct estimator that has this double robustness property.

4 A Monte Carlo Study

As noted above, the theoretical results for the AIPW estimator are large-sample in nature. In order to gauge the finite sample performance of the AIPW estimator relative to the standard regression, IPW, and matching estimators we designed a Monte Carlo study.

4.1 Study Design

The basic design of the study features three levels of confounding (low, moderate, severe), two mean functions (linear, nonlinear) linking treatment status and background variables to the outcome variable, and three sample sizes (250, 500, 1000) for a total of 18 types of Monte Carlo datasets. 1000 datasets were created under each of these 18 scenarios for a total of 18,000 Monte Carlo datasets. These datasets were saved to disk and each estimator was applied to the same 18,000 datasets. The remainder of this subsection provides additional detail about how the Monte Carlo study was conducted.

4.1.1 Data Generating Processes

All of the Monte Carlo datasets feature five variables: Z_1, Z_2, Z_3, X, and Y. Z_1, Z_2, and Z_3 represent background variables, X denotes treatment status, and Y is the outcome variable. Z_1, Z_2, and Z_3 are drawn from independent standard normal distributions. New draws of these variables are obtained for each of the 18,000 datasets. With Z_1, Z_2, and Z_3 in hand, treatment status X is drawn from a Bernoulli distribution where the probabilities of $X = 1$ depend on the realized Z_1, Z_2 and the degree of confounding (low, moderate, severe). Table 1 summarizes the treatment assignment probabilities as a function of Z_1 and Z_2 under the three levels of confounding. Once Z_1, Z_2, Z_3, and X have been generated we generate the outcome variable Y. Y is assumed to follow a normal distribution with a mean that depends on Z_2, Z_3, and X and a constant variance of 1. The mean function for Y can be either linear or nonlinear in Z_2 and Z_3. Table 2 provides the mean functions for treated and control units under the linear and nonlinear scenarios.

From Tables 1 and 2 we see that treatment assignment depends on Z_1 and Z_2 while the outcome depends on Z_2, Z_3, and X. Because Z_1, Z_2, and Z_3 do not have any common causes, it follows that adjusting for just Z_2 (either in the outcome model or the treatment assignment model) is sufficient to produce a consistent

Degree of Confounding	True Treatment Assignment Probabilities
Low	$\Pr(X = 1\|\mathbf{Z}) = \Phi(0.1Z_1 + 0.1Z_2 + 0.05Z_1Z_2)$
Moderate	$\Pr(X = 1\|\mathbf{Z}) = \Phi(Z_1 + Z_2 + 0.5Z_1Z_2)$
Severe	$\Pr(X = 1\|\mathbf{Z}) = \Phi(1.5Z_1 + 1.5Z_2 + 0.75Z_1Z_2)$

Table 1: *Equations Governing Treatment Assignment in the Monte Carlo Study.* Observation-specific subscripts have been left off. $\Phi(\cdot)$ denotes the standard normal distribution function. Each unit's treatment status is assumed to be drawn independently from a Bernoulli distribution according to the probabilities above.

Outcome Mean	Outcome Equation (Control)	Outcome Equation (Treatment)
Linear	$Y = Z_2 + Z_3 + \epsilon$	$Y = 5 + 3Z_2 + Z_3 + \epsilon$
Nonlinear	$Y = Z_2 + Z_3 + \epsilon$	$Y = 5 + 3Z_2 + Z_3 + 2Z_2^2 + 2Z_3^2 + \epsilon$

Table 2: *Equations Governing the Outcome Variable in the Monte Carlo Study.* Observation-specific subscripts have been left off. It is assumed that ϵ follows a standard normal distribution and that ϵ is independent across observations.

estimate of the average treatment effect of X on Y (Pearl, 1995, 2000). In fact, given the structure of the data generating process, it is the case that one could adjust for any combination of Z_1, Z_2, and Z_3 that includes Z_2 to produce a consistent estimate of the average treatment effect—this is true of all the estimators considered in this paper. Nevertheless, as we will see in the Monte Carlo results, there will be better and worse choices of adjustment strategies in finite samples.

Figure 1 depicts the distributions of the treatment assignment probabilities among units that actually received treatment and control under the three different levels of confounding. In addition, this figure also looks at these treatment assignment probabilties conditional on both Z_1 and Z_2 (the true assignment mechanism) and conditional on Z_2 but averaged over Z_1 (the minimal assignment mechanism). Several points are worth noting here. First, under the low level of confounding the distribution of treatment assignment probabilities looks very similar across treated and control units. Thus, we would expect all the estimators to perform well on the Monte Carlo datasets that feature low confounding. Under moderate and severe confounding the distributions of treatment assignment probabilities become increasingly distinct for treated and control units. This is especially the case when the treatment assignment probabilities are conditional on both Z_1 and Z_2. Interestingly, if one calculates the treatment assignment probabilities conditional on just Z_2 (the minimal assignment mechanism) one achieves better overlap between the treated group and the control group. Thus, we would expect that estimators that make use of the propensity score and specify the

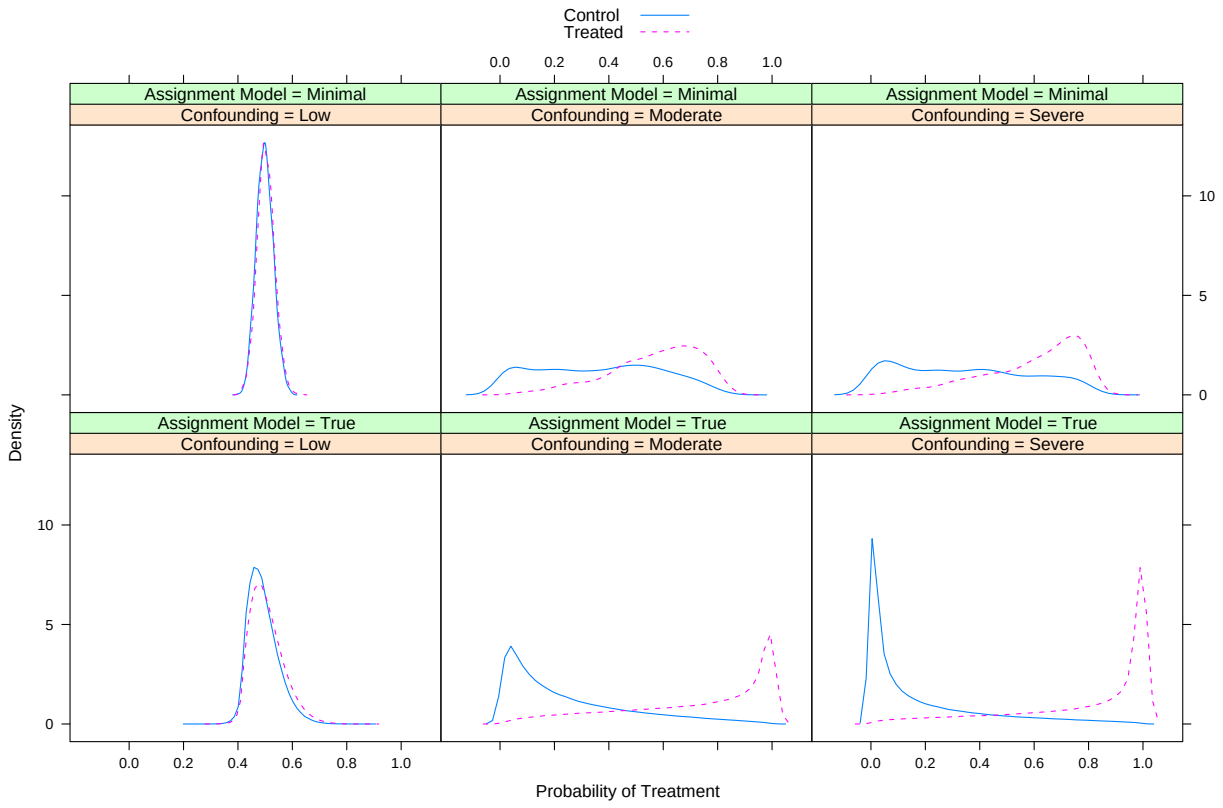

Figure 1: *Treatment Assignment Probabilities Among Treated and Control Units.* Assignment Model = True corresponds to the assignment probabilities conditional on Z_1 and Z_2. Assignment Model = Minimal corresponds to the assignment probabilities conditional on Z_2 but averaging over Z_1. Note that by using only Z_2 in the assignment model (as one would need to specify for the IPW, Matching, and AIPW estimators) one produces better overlap between the treated and control units while still alleviating confounding bias.

propensity score just as a function of Z_2 will perform better than those that specify the propensity score as a function of Z_1 and Z_2—despite the fact that the actual treatment assignment mechanism depends on Z_1.

It is also useful to get a visual depiction of the outcome variable as a function of treatment status, the single confounding variable Z_2, the degree of confounding, and the form of the mean function for the outcome variable. Figure 2 displays this information. Here we see that under low confounding there are enough treated and control units at each level of Z_2 to identify the treated and control outcome regressions across the range of Z_2. This is true for both the linear and nonlinear outcome mean functions. Extrapolation becomes necessary when we move to the datasets with either moderate or extreme confounding. For such datasets with a linear mean function for the outcome variable we expect that the extrapolation will not cause

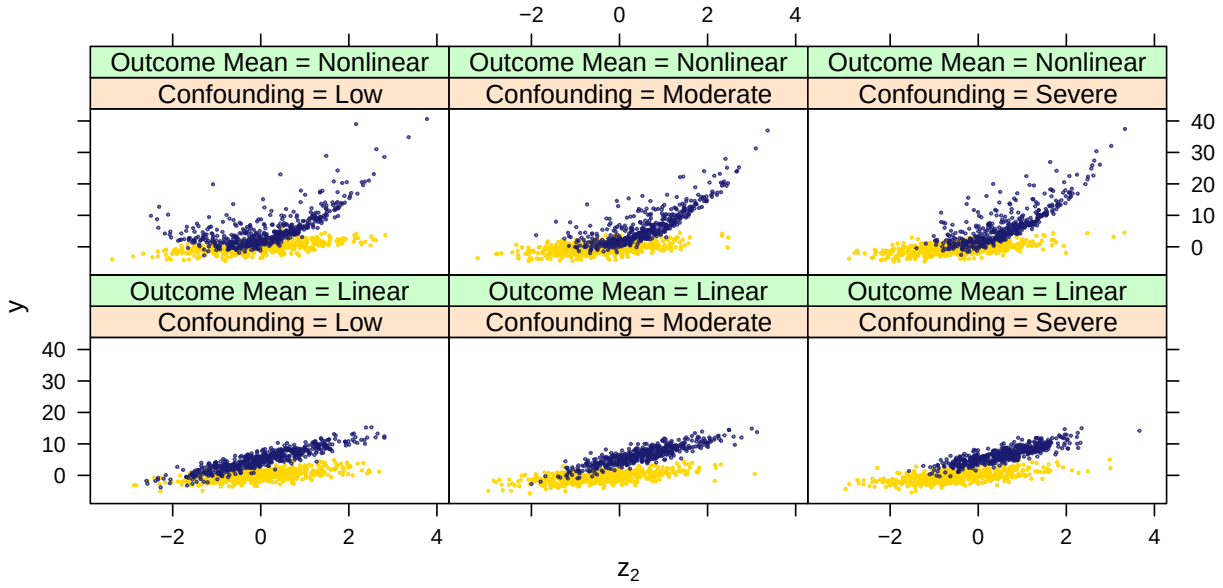

Figure 2: *Scatterplots of Outcome Variable Y as a Function of Treatment Status and the Confounder Z_2 Given the Degree of Confounding and the Form of the Mean Function for Y.* The darker blue points correspond to treated units and the lighter gold points correspond to control units. There are 1000 total units in each panel.

major problems in estimating the average treatment effect. However, for datasets with a nonlinear outcome mean and moderate or severe confounding we expect the estimates to be more adversely affected since there is very little information in the data about the counterfactual outcome mean under treatment for the control units with Z_2 less than about -1.5 and -1.0 respectively. We note in passing that the panels in Figure 2 are only conditioned on Z_2—the minimal confounder. More accurate estimates of the outcome mean function can be obtained by conditioning on Z_2 and Z_3.

4.1.2 Model Specifications

Now that we have discussed the data generation process under each of the 18 Monte Carlo scenarios we move on to discuss the six model specifications used in the Monte Carlo study. These are summarized in Table 3. Each specification consists of a propensity score model, an outcome model for treated units, and an outcome model for control units. Not all estimators will use all three models. For instance, the matching and IPW estimators will only use the propensity score model, while the regression estimator will only use the two outcome models. The propensity score model is a generalized additive model (GAM) for binomial

Specification	Propensity Score Model	Outcome Model (Treated)	Outcome Model (Control)
A	`X ~ s(Z1, Z2)`	`Y ~ s(Z2) + s(Z3)`	`Y ~ s(Z2) + s(Z3)`
B	`X ~ s(Z1, Z2) + s(Z3)`	`Y ~ s(Z1) + s(Z2) + s(Z3)`	`Y ~ s(Z1) + s(Z2) + s(Z3)`
C	`X ~ s(Z2)`	`Y ~ s(Z2)`	`Y ~ s(Z2)`
D	`X ~ s(Z2)`	`Y ~ s(Z2) + s(Z3)`	`Y ~ s(Z2) + s(Z3)`
E	`X ~ s(Z1)`	`Y ~ s(Z2)`	`Y ~ s(Z2)`
F	`X ~ s(Z2)`	`Y ~ s(Z3)`	`Y ~ s(Z3)`

Table 3: *Model Specifications Used in the Monte Carlo Study.* Each specification consists of a propensity score model, an outcome model for treated units, and an outcome model for control units. Not all estimators will use all three models. The propensity score model is a generalized additive model (GAM) for binomial outcomes with a probit link and the outcome models are GAMs for conditionally Gaussian outcomes with the identity link. The three cells to the right of a given specification consist of the `R` formula sent to the `gam` function in the `mgcv` package that performed the model fitting. Entries in black are sufficient adjustments to achieve consistent estimates of average treatment effects. Entries in red are not sufficient to control confounding bias. All four estimators under study (regression, matching, IPW, and AIPW) should be consistent for the average treatment effect under specifications A, B, C, and D. This will not be true for specifications E and F.

outcomes with a probit link and the outcome models are GAMs for conditionally Gaussian outcomes with the identity link.

We can think of these specifications as follows. In specification A, both the propensity score model and the outcome models are fully consistent with the true models that generated the data. Specification B includes all three Z variables in the propensity score model and the outcome models. Specification C can be thought of the minimal specification in that only the minimal confounder Z_2 enters into the propensity score model and the outcome models. Specification D consists of the minimal propensity score model and the true outcome models. Each of specifications A, B, C, and D is sufficient for consistent estimation of average treatment effects. Specifications E and F are partially misspecified. In specfication E the propensity score model is misspecified while the outcome models are specified in a way that is sufficient to control confounding. Thus we would expect that the use of specification E with either the matching or IPW estimator would result in biased and inconsistent estimates of causal effects. In specification F the propensity score model is specified in a way so as to control confounding but the outcome regressions omit the confounder Z_2 and are thus misspecified. We would thus expect that the use of this specification with the regression estimator would produced biased and inconsistent estimates of causal effects. Because not all estimators use all three pieces of a specification it will be the case that some specifications will be equivalent for a particular estimator.

For instance, specifications C, D, and F are equivalent for the matching estimator and the IPW estimator.

4.2 Results

4.2.1 Bias Under Specifications Consistent for ATE

We first look at results from specifications A, B, C, and D. Under any of these four specifications all of the estimators under study (matching, IPW, AIPW, regression) are consistent for the average treatment effect. Nonetheless, we do expect them to have different finite sample performance. Figure 3 presents Monte Carlo estimates of the bias of each of these four estimators across the various sample sizes, levels of confounding, (non)linearity of the outcome mean function, and specifications A, B, C, and D.[3]

Looking at the results under low confounding we see that all of the estimators appear to be essentially unbiased in any of the three sample sizes and with either true outcome mean function. There appears to be a very slight amount of downward bias in the matching estimator under specification B and, to a lesser extent, specification A. These specifications include more variables than necessary in the propensity score model and thus good matches become more difficult to find. Nevertheless, all four estimators perform well across all specifications and datasets with low confounding.

With moderate confounding the performance of the estimators begins to diverge. Looking first at the estimated bias under the true linear outcome model and moderate confounding we see that the regression and AIPW estimators are essentially unbiased at all sample sizes. The matching estimator exhibits some minor upward bias in small samples but this largely disappears in larger samples. The IPW estimator shows noticeable upward bias in small samples and a small amount of bias with $n = 1000$. This is true across specifications A, B, C, and D. Looking at the estimated bias under a true nonlinear outcome model and moderate confounding we see that all of the estimators show some signs of bias across the various specifications and samle sizes. As we would expect, bias decreases with sample size. All of the estimators perform similarly here with perhaps a slight edge in terms of bias going to the IPW estimator.

The results under moderate confounding are accentuated under severe confounding. Here, under the true

[3]All of the Monte Carlo analyses were conducted in `R` (R Development Core Team, 2007). We use the `gam` function in the `mgcv` package (Simon Wood, 2006) to estimate the propensity score and outcome models. The `Matching` package (Jasjeet S. Sekhon, 2006) was used to estimate the ATE using one to one nearest neighbor matching on the estimated propensity score.

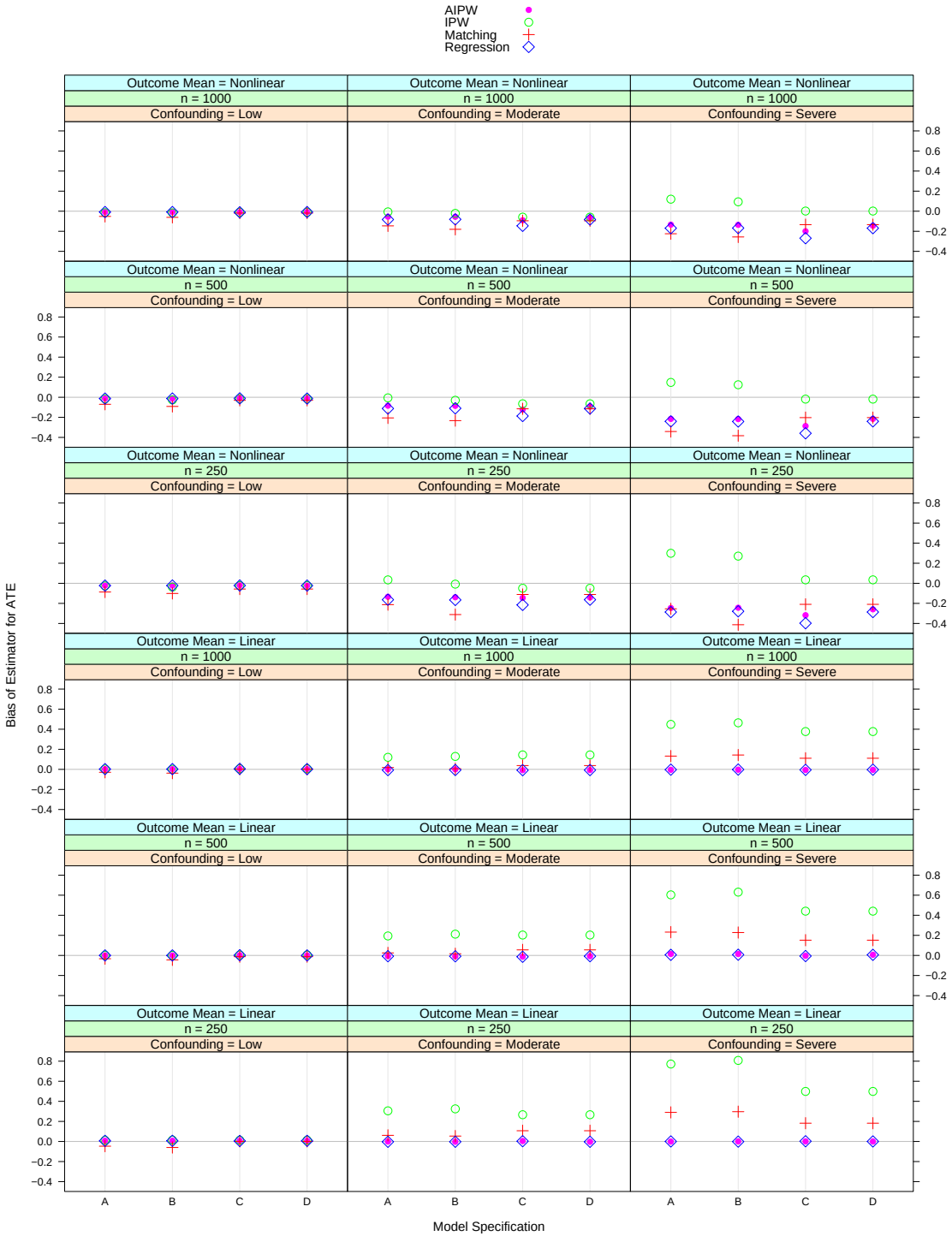

Figure 3: *Monte Carlo Estimates of Bias for Four Estimators of the Average Treatment Effect Across Four Different Model Specifications, Three Levels of Confounding, Three Sample Sizes, and Two Mean Functions for the Outcome.*

linear outcome model, the AIPW and regression estimators remain essentially unbiased in all sample sizes. However the IPW estimator becomes badly biased with the matching estimator somewhere in between. Under the true nonlinear outcome model we see that the patterns under moderate confounding are accentuated with all of the estimators showing noticeable bias but the bias diminishing as sample size increases. Here with $n = 1000$ and specification D none of the estimators display large amounts of bias.

In summary, the results here should not be that surprising. In situations with minimal confounding all four estimators are essentially unbiased under a range of specifications. With moderate or severe confounding and linear outcome mean functions the estimators that model the outcome mean function perform the best. In situations with moderate or severe confounding and nonlinear outcome mean functions all of the estimators exhibit some finite sample bias but this diminishes as sample size increases.

4.2.2 Root Mean Square Error (RMSE) Under Specifications Consistent for ATE

Looking just at the finite sample bias of correctly specified versions of the four estimators under study does not provide clear guidance as to which estimator is to be preferred. However, looking at the root mean square error (RMSE) of the estimators provides more relevant information. Figure 4 plots the RMSE of the IPW, matching, and regression estimators relative to the RMSE of the AIPW estimator. Here values greater than 1 indicate the estimator in question had a larger RMSE than the AIPW estimator while a value less than 1 indicates the estimator had a smaller RMSE than the AIPW estimator.

Looking at Figure 4 we see that the RMSE of the regression estimator is always about the same as that of the AIPW estimator. On the other hand, the RMSE of the matching and IPW estimators are typically much higher than that of the AIPW estimator. Thus if we were certain that we had a correct model specification either the AIPW or the regression estimator would appear to be superior to the matching or IPW estimators.

4.2.3 Bias Under Specifications Inconsistent for ATE

Of course we never know whether our model specification is sufficient to control confounding. For this reason, we would like to know how the various estimators perform when the model specifications are partially deficient in the sense that either (a) the treatment assignment model is misspecified and the outcome models

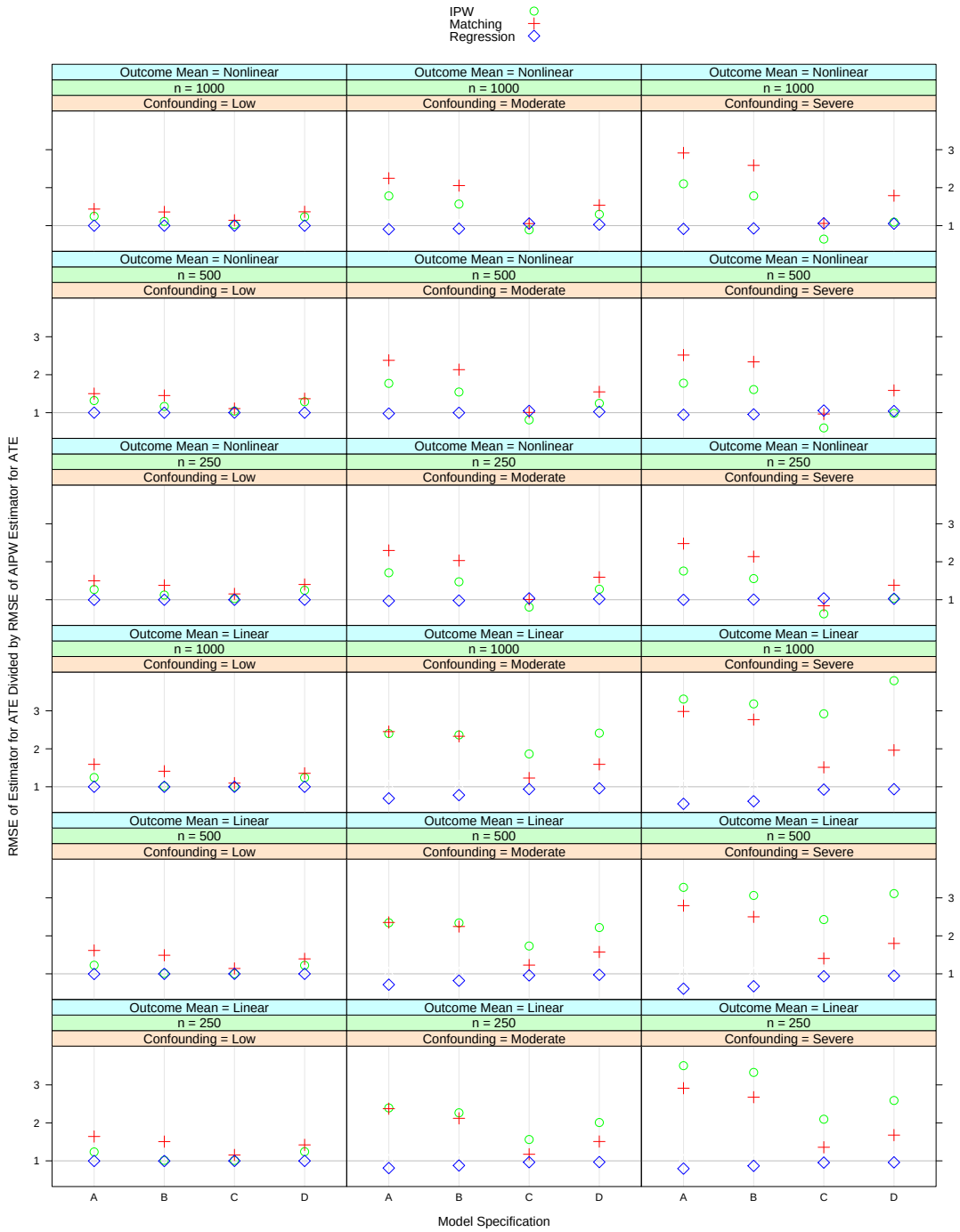

Figure 4: *Monte Carlo Estimates of Root Mean Square Error for Three Estimators of the Average Treatment Effect Relative to the Root Mean Square Error of the AIPW Estimator Across Four Different Model Specifications, Three Levels of Confounding, Three Sample Sizes, and Two Mean Functions for the Outcome.* A value of 1 implies the estimator in question has the same RMSE as the AIPW estimator, a value greater than 1 indicates that the estimator in question has a RMSE greater than the AIPW estimator.

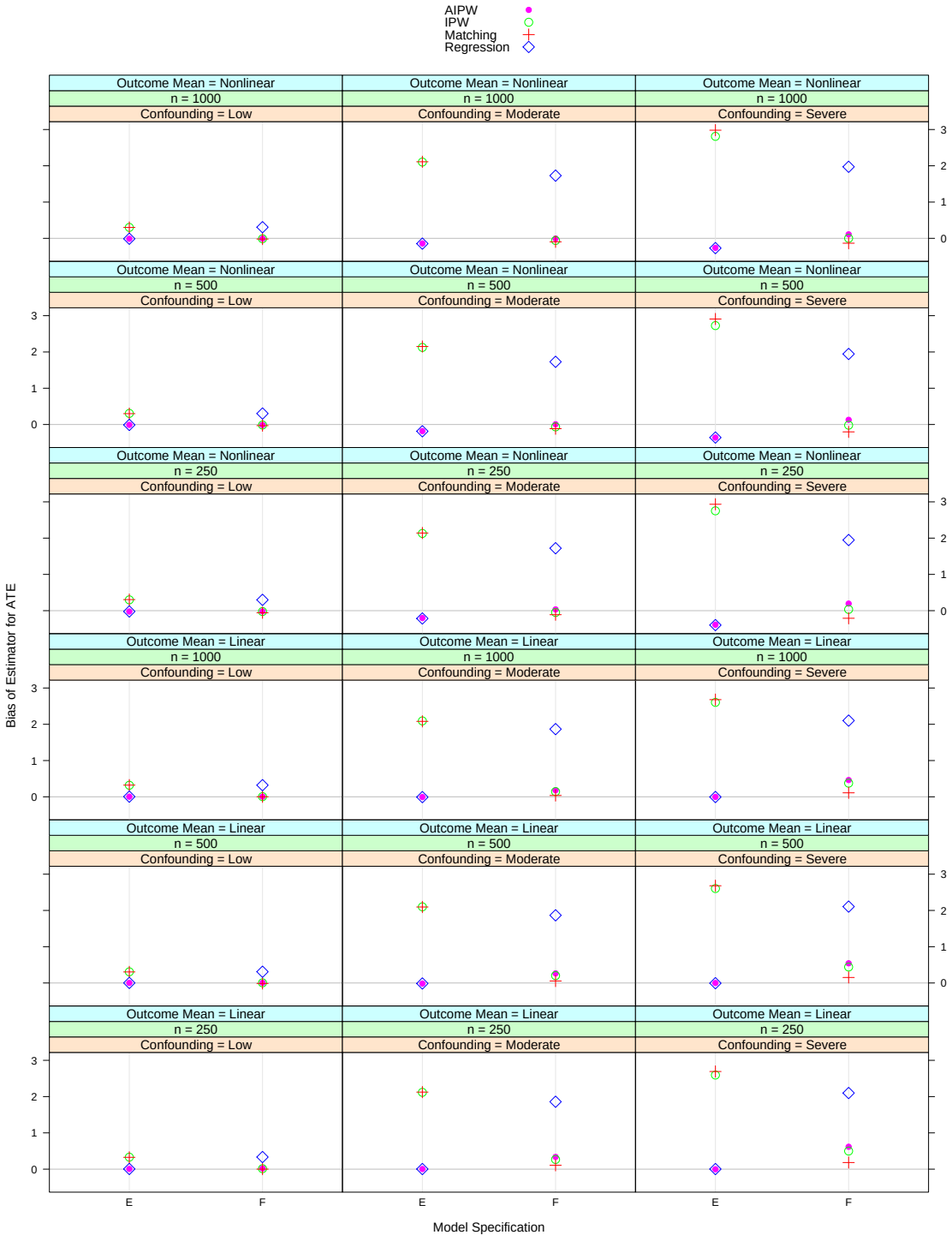

Figure 5: *Monte Carlo Estimates of Bias for Four Estimators of the Average Treatment Effect Across Two Different Model Specifications, Three Levels of Confounding, Three Sample Sizes, and Two Mean Functions for the Outcome.*

are correctly specified or (b) the treatment assignment model is correcty specified but the outcome models are misspecified. Specification E falls under category (a) while specification F falls under category (b).

Figure 5 shows the bias of the four estimators across the various Monte Carlo scenarios under specifications E and F. The general pattern here is quite clear– *only the AIPW estimator remains essentially unbiased across all scenarios and both model specifications.* With moderate or severe confounding the regression estimator will fail miserably if specification F is used, while the IPW and matching estimators will perform even worse if specification E is used. Nonetheless, as long as either the propensity score model or the outcome model is properly specified the AIPW estimator exhibits only small amounts of bias. Further, we know from theoretical results of (Scharfstein et al., 1999) that the AIPW will retain its consistency for the ATE under such partial misspecification.

4.2.4 RMSE Under Specifications Inconsistent for ATE

While the double robustness property of the AIPW illustrated above would seem to strongly favor its use over the IPW, matching, or regression estimators we might also be interested in its RMSE relative to other estimators under partial misspecification. Figure 6 plots the RMSE of the IPW, matching, and regression estimators relative to the RMSE of the AIPW estimator under specifications E and F. Consistent with the earlier Monte Carlo results we see that the RMSE of these other estimators are never much, if at all, below that of the AIPW estimator and under some circumstances the RMSE of these estimators is dramatically (10 to 15 times) higher than that of the AIPW estimator.

5 Discussion

In this paper we have shown that the AIPW performs about as well, or slightly better than, extant estimators under a fully correct specification. However, the AIPW estimator performs *dramatically* better than IPW, matching (one to one nearest neighbor matching on the estimated propensity score), or regression estimators under partial misspecification. Of course, this study should not be taken as comprehensive (other data generating models and ATE estimators should be considered in future studies). However, these initial results indicate the promise for estimators of this type. Since there is essentially no cost to using the AIPW

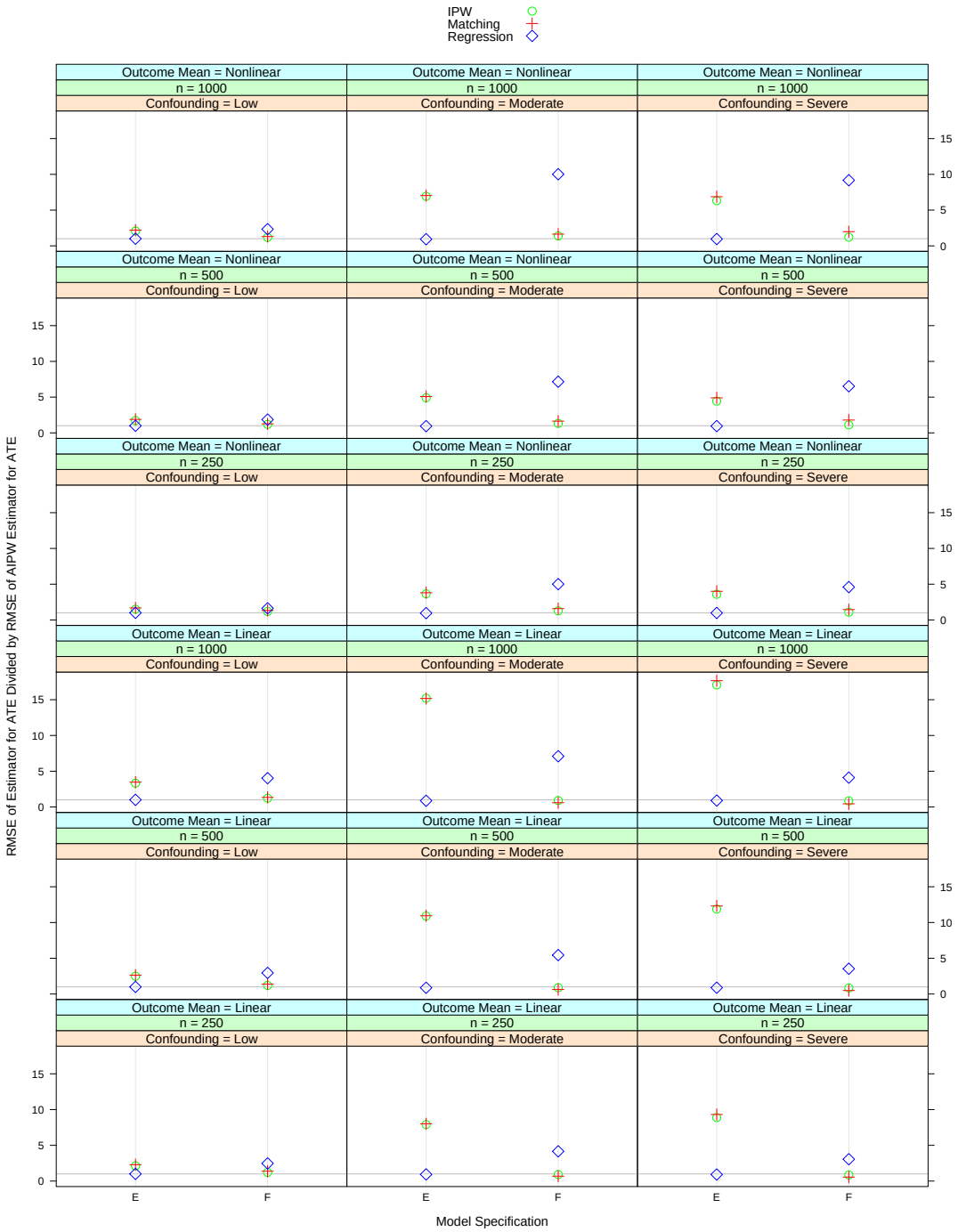

Figure 6: *Monte Carlo Estimates of Root Mean Square Error for Three Estimators of the Average Treatment Effect Relative to the Root Mean Square Error of the AIPW Estimator Across Two Different Model Specifications, Three Levels of Confounding, Three Sample Sizes, and Two Mean Functions for the Outcome.* A value of 1 implies the estimator in question has the same RMSE as the AIPW estimator, a value greater than 1 indicates that the estimator in question has a RMSE greater than the AIPW estimator.

estimator when one knows the correct specification and sizable advantages to using the AIPW estimator when the specification is partially deficient, it seems reasonable that most applied researchers should seriously consider using the AIPW estimator for their applied work on average treatment effects.

A Statistical Properties of the AIPW Estimator

A.1 Unbiasedness and Consistency of the AIPW Estimator

If we assume that the true propensity scores are regression functions are known, then the AIPW estimator can be shown to be unbiased for the ATE. This is easiest to demonstrate by first showing the in-sample unbiasedness of the IPW estimator, and then showing that the adjustment term of the AIPW estimator has in-sample expectation of zero.

$$\begin{aligned}
\mathbb{E}[\widehat{ATE}_{IPW}] &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\frac{X_i Y_i}{\pi(\mathbf{Z}_i)}\right] - \mathbb{E}\left[\frac{(1-X_i)Y_i}{1-\pi(\mathbf{Z}_i)}\right]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i(1)}{\pi(\mathbf{Z}_i)}\mathbb{E}[X_i] - \frac{Y_i(0)}{1-\pi(\mathbf{Z}_i)}\mathbb{E}[(1-X_i)]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i(1)}{\pi(\mathbf{Z}_i)}\pi(\mathbf{Z}_i) - \frac{Y_i(0)}{1-\pi(\mathbf{Z}_i)}1-\pi(\mathbf{Z}_i)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}
\end{aligned}$$

Therefore the IPW estimator is unbiased for the in-sample ATE. Unbiasedness in the population can be established by iterated expectation. Given this result, we can establish the in-sample unbiasedness of the AIPW estimator by showing that the adjustment term has expectation zero.

$$\begin{aligned}
\mathbb{E}[\widehat{ATE}_{AIPW}] = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\mathbb{E}\left[\frac{X_i Y_i}{\pi(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\pi(\mathbf{Z}_i)}\right] \\
&- \mathbb{E}\left[\frac{X_i - \pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\left[(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i) + \pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)\right]\right]\Bigg\} \\
= \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&[Y_i(1) - Y_i(0)] \\
&- \mathbb{E}\left[\frac{X_i - \pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\right]\left[(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i) + \pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)\right]\Bigg\} \\
= \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&[Y_i(1) - Y_i(0)] \\
&- \left[\frac{\pi(\mathbf{Z}_i) - \pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\right]\left[(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i) + \pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)\right]\Bigg\} \\
= \frac{1}{n}\sum_{i=1}^{n}&\{Y_i(1) - Y_i(0)\}
\end{aligned}$$

Again the unbiasedness of this estimator in the population can be established by iterated expectation. Consistency follows because the AIPW estimator is a sample average.

A.2 Double Robustness of the AIPW Estimator

The following proof of the double robustness of the AIPW estimator is directly from Chapter 13 of Tsiatis (2006). In order to facilitate exposition we will introduce slightly different notation than was used in the body of this paper. Here we write the propensity score as

$$\Pr(X = 1|\mathbf{Z}) = \pi(\mathbf{Z}, \psi)$$

where ψ is a finite dimensional parameter that governs the propensity score function. Similarly, we write the outcome regressions as:

$$\mathbb{E}(Y|X = 1, \mathbf{Z}) = \mu(X = 1, \mathbf{Z}, \xi)$$

and

$$\mathbb{E}(Y|X = 0, \mathbf{Z}) = \mu(X = 0, \mathbf{Z}, \xi)$$

where ξ is a finite dimensional parameter that governs the conditional expectation function of the outcome regression. With the new notation, the estimated propensity score function in a sample of size n is given by $\pi(\mathbf{Z}, \hat{\psi}_n)$ and the estimated outcome regression function in a sample of size n is given by $\mu(X, \mathbf{Z}, \hat{\xi}_n)$. It is assumed that $\hat{\psi}_n$ converges in probability to some value ψ^* and that $\hat{\xi}_n$ converges in probability to some value ξ^* as sample size goes to infinity. When $\psi^* = \psi_0$ we will say the propensity score model is correctly specified. Similarly, when $\xi^* = \xi_0$ we will say that the outcome regression is correctly specified.

Assume that the assumptions of SUTVA and strong ignorability of treatment assignment given $\mathbf{Z}$ hold. We wish to show that $\widehat{ATE}_{AIPW}$ is consistent for ATE if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$.

The AIPW estimator given by Equation 3 can be rewritten as:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i Y_i}{\pi(\mathbf{Z}_i, \hat{\psi}_n)} - \frac{\left[X_i - \pi(\mathbf{Z}_i, \hat{\psi}_n)\right]\mu(X = 1, \mathbf{Z}_i, \hat{\xi}_n)}{\pi(\mathbf{Z}_i, \hat{\psi}_n)} \right.$$
$$\left. - \frac{(1 - X_i)Y_i}{1 - \pi(\mathbf{Z}_i, \hat{\psi}_n)} - \frac{\left[X_i - \pi(\mathbf{Z}_i, \hat{\psi}_n)\right]\mu(X = 0, \mathbf{Z}_i, \hat{\xi}_n)}{1 - \pi(\mathbf{Z}_i, \hat{\psi}_n)}\right\}$$

Because this is a sample average, $\widehat{ATE}_{AIPW}$ converges in probability to:

$$\mathbb{E}\left[\frac{XY}{\pi(\mathbf{Z}, \psi^*)} - \frac{[X - \pi(\mathbf{Z}, \psi^*)]\,\mu(X = 1, \mathbf{Z}, \xi^*)}{\pi(\mathbf{Z}, \psi^*)} - \frac{(1 - X)Y}{1 - \pi(\mathbf{Z}, \psi^*)} - \frac{[X - \pi(\mathbf{Z}, \psi^*)]\,\mu(X = 0, \mathbf{Z}, \xi^*)}{1 - \pi(\mathbf{Z}, \psi^*)}\right]. \quad (6)$$

Using SUTVA and simple algebra we can write:

$$\frac{XY}{\pi(\mathbf{Z},\psi^*)} = \frac{XY(1)}{\pi(\mathbf{Z},\psi^*)} = Y(1) + \frac{[X - \pi(\mathbf{Z},\psi^*)]\,Y(1)}{\pi(\mathbf{Z},\psi^*)} \tag{7}$$

and

$$\frac{(1-X)Y}{1-\pi(\mathbf{Z},\psi^*)} = \frac{(1-X)Y(0)}{1-\pi(\mathbf{Z},\psi^*)} = Y(0) + \frac{[X - \pi(\mathbf{Z},\psi^*)]\,Y(0)}{1-\pi(\mathbf{Z},\psi^*)} \tag{8}$$

where $Y(1)$ denotes the potential outcome under treatment of a randomly chosen unit and $Y(0)$ denotes the potential outcome under control of a randomly chosen unit.

Next, we substitute 7 and 8 back into 6 to get:

$$\mathbb{E}\left[Y(1) - Y(0)\right] \tag{9}$$

$$+\mathbb{E}\left[\frac{[X - \pi(\mathbf{Z},\psi^*)]\,[Y(1) - \mu(X=1,\mathbf{Z},\xi^*)]}{\pi(\mathbf{Z},\psi^*)}\right] \tag{10}$$

$$+\mathbb{E}\left[\frac{[X - \pi(\mathbf{Z},\psi^*)]\,[Y(0) - \mu(X=0,\mathbf{Z},\xi^*)]}{1-\pi(\mathbf{Z},\psi^*)}\right]. \tag{11}$$

Note that 9 is the definition of ATE. Thus in order to prove that $\widehat{ATE}_{AIPW}$ is consistent for ATE if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$ it is sufficient to show that expectations 10 and 11 equal 0 if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$.

First consider the case where the $\psi^* = \psi_0$ (the propensity score model is correctly specified). Using the law of iterated conditional expectations one can write expectation 10 as

$$\mathbb{E}\left[\frac{\{\mathbb{E}\left[X|Y(1),\mathbf{Z}\right] - \pi(\mathbf{Z},\psi_0)\}\,\{Y(1) - \mu(X=1,\mathbf{Z},\xi^*)\}}{\pi(\mathbf{Z},\psi_0)}\right] \tag{12}$$

Conditional ignorability implies that:

$$\mathbb{E}[X|Y(1),\mathbf{Z}] = \mathbb{E}[X|\mathbf{Z}] = \pi(\mathbf{Z},\psi_0).$$

Substituting $\pi(\mathbf{Z},\psi_0)$ in for $\mathbb{E}[X|Y(1),\mathbf{Z}]$ in expectation 12 we see that expectation 10 is equal to 0 when $\psi^* = \psi_0$. Directly analogous calculations can be used to show that expectation 11 is also equal to 0 when $\psi^* = \psi_0$. Thus $\widehat{ATE}_{AIPW}$ is consistent for ATE when the propensity score model is correcty specified and the outcome regressions are misspecified.

We now turn our attention to the situation where $\xi^* = \xi_0$ (the outcome regressions are correctly specified). Using the law of iterated conditional expectations one can write expectation 10 as

$$\mathbb{E}\left[\frac{\{X - \pi(\mathbf{Z}, \psi^*)\}\{\mathbb{E}[Y(1)|X, \mathbf{Z}] - \mu(X = 1, \mathbf{Z}, \xi_0)\}}{\pi(\mathbf{Z}, \psi^*)}\right]. \tag{13}$$

The strong ignorability assumption allows us to write

$$\mathbb{E}[Y(1)|X, \mathbf{Z}] = \mathbb{E}[Y(1)|X = 1, \mathbf{Z}]$$

and SUTVA allows us to write

$$\mu(X = 1, \mathbf{Z}, \xi_0) = \mathbb{E}[Y|X = 1, \mathbf{Z}] = \mathbb{E}[Y(1)|X = 1, \mathbf{Z}].$$

Thus we can substitute $\mu(X = 1, \mathbf{Z}, \xi_0)$ in for $\mathbb{E}[Y(1)|X, \mathbf{Z}]$ in expectation 13. Doing this we see that expectation 10 is equal to 0 when $\xi^* = \xi_0$. Similar calculations can be used to show that 11 is also equal to 0 when $\xi^* = \xi_0$. Thus $\widehat{ATE}_{AIPW}$ is consistent for ATE when the outcome regression models are correcty specified and the propensity score model is misspecified. This completes the proof of double robustness.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.

Diamond, A., and J.S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *See http://sekhon. berkeley. edu/papers/GenMatch. pdf* .

Ho, D.E., K. Imai, G. King, and E.A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199.

Jasjeet S. Sekhon. 2006. *Multivariate and Propensity Score Matching with Balance Optimization*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:131–159.

Lunceford, Jared K., and Marie Davidian. 2004. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine* 23:2937–2960.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669–710.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robins, James M. 1999. "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models." *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* pp. 6–10.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors are Not Always Observed." *Journal of the American Statistical Association* 89:846–866.

Robins, J.M. 1986. "A new aproach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rubin, D.B. 2006. *Matched Sampling for Causal Effects*. Cambridge University Press.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. "Rejoinder to Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association* 94:1135–1146.

Simon Wood. 2006. *GAMs with GCV smoothness estimation and GAMMs by REML/PQL*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tsiatis, Anastasios A. 2006. *Semiparametric Theory and Missing Data*. New York: Springer.