Improving Modern Techniques of Causal Inference: Finite Sample Performance of ATM and ATO Doubly Robust Estimators, Variance Estimation for ATO Estimators, and Contextualized Tipping Point Sensitivity Analyses for Unmeasured Confounding

By

Lucy Elizabeth D'Agostino McGowan

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May 11, 2018

Nashville, Tennessee

Approved:

Frank Harrell, Jr., Ph.D.

Robert Alan Greevy, Jr., Ph.D.

Qingxia (Cindy) Chen, Ph.D.

Peter Rebeiro, Ph.D.

Copyright © 2018 by Lucy Elizabeth D'Agostino McGowan
All Rights Reserved

To my husband, Patrick,
&
our little one on the way

ACKNOWLEDGEMENTS

I would like to begin by thanking my committee members, Frank Harrell, Cindy Chen, Peter Rebeiro, and my advisor, Robert Greevy. I have learned so much from each of them as they provided invaluable insights and support throughout the process. I would especially like to thank Robert Greevy for his wisdom and kindness and for his constant belief in my ideas and aspirations.

I would like to thank Christianne Roumie and the Effectiveness team for providing me with outstanding applied experience throughout my time as a student and Jenny Bryan for teaching me tools for effective collaboration and elegant coding style. I am grateful for the R-Ladies movement and the friends and colleagues I have met through it over the past few years, for their friendship, technical acumen, and optimism. I would like to thank each member of my cohort, Alice Toll, Alice Curtis, Svetlana Eden, Jonathan Chipman, and Ryan Jarrett, and others from the Vanderbilt Biostatistics family, especially Allison Hainline, Jacquelyn Neal, Mark Giganti, and Hannah Weeks, for encouragement and support at every step of the way, from studying for exams, to conversations about life while sipping coffee, eating Thai food, or watching Wizard People – I am so grateful to have had the opportunity to study alongside each of them.

Finally, I would like to thank my family for their constant outpouring of love and support. My parents, Ralph and Carey, grandparents, LeiLanie and Ralph and Richard and Connie, and siblings, Serena, Sophia, Arel, Theresa, and Chiara. They are my cheerleaders, role models, and inspiration, and I am immensely grateful for each of them. Ultimate thanks go to my husband, Patrick, for his constant joy, patience, incredible selflessness, and kindness, celebrating my accomplishments and uplifting me when I need it most, he is the greatest partner and friend and I am so grateful to have him by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

APACHE Acute Physiology And Chronic Health Evaluation

ATC Average treatment effect among the controls

ATE Average treatment effect

ATM Average treatment effect among the matchable

ATO Average treatment effect for the overlap population

ATT Average treatment effect among the treated

DNR Do not resuscitate

DR Doubly robust

HR Hazard ratio

IPW Inverse probability weight

LB Limiting bound

LCL Lower confidence limit

OR Odds ratio

RHC Right heart catheterization

RMSE Root mean square error

RR Relative risk

SMD Standardized mean difference

SUPPORT Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments

UCL Upper confidence limit

CHAPTER 1

INTRODUCTION

This body of work systematically approaches unresolved aspects of the modern causal inference process in a sequential manner. In a causal inference framework, one first identifies the weighting scheme that will be used, then estimates the causal estimand of interest, along with its corresponding variance, and finally conducts sensitivity analyses to quantify how susceptible the result is to unknown factors, such as unmeasured confounding. In this mindset, we first examine the finite-sample properties of three weighting schemes, weights to estimate the average treatment effect (ATE), the average treatment effect among the matchable (ATM), and the average treatment effect for the overlap population (ATO). The latter two weighting schemes are relatively new to the field, introduced in 2013 and 2016 respectively (Li and Greene 2013; Li, Morgan, and Zaslavsky 2016). Once we have determined the appropriate weights, Chapter 3 provides derivations to estimate the quantity of interest as well as its variance. Finally, Chapter 4 focuses on contextualizing sensitivity to unmeasured confounding analyses. This work will be illustrated via applied examples.

In an observational study setting, inverse probability weighting (IPW) can be implemented to reduce bias in the causal estimate of interest. A seminal paper by Freedman and Berk (2008) revealed that weights designed to estimate the average treatment effect (ATE) could suffer from finite-sample biases and inefficiency (Freedman and Berk 2008). Revisiting the setting of this paper, in Chapter 2 we demonstrate that two new weighting approaches (ATM and ATO weights) do not have these downsides. We extend the setting to explore how large a sample size is required for good performance to be observed. Finally, we highlight the importance of identifying the true causal effect in studies like these, where simple interpretations of model coefficients can be misleading.

The methods and performance of IPW and IPW doubly robust estimators incorporating the recently defined ATO weights are important open questions in the field. In Chapter 3, we derive the large-sample variance estimator for the ATO estimator and doubly robust estimator for generalized linear models with identity, log, or logistic links. We then explore how this estimation compares to commonly used modeling and variance

estimation techniques under settings where the propensity score and outcome models are both correctly specified, when one is incorrectly specified, and when both are incorrectly specified.

The strength of evidence provided by epidemiological and observational studies is inherently limited by the potential for unmeasured confounding. Thus, we would expect every observational study to include a quantitative sensitivity to unmeasured confounding analysis. However, we reviewed 90 recent studies with "statistically significant" findings, published in top tier journals, and found 41 mentioned the issue of unmeasured confounding as a limitation, but only 4 included a quantitative sensitivity analysis. Moreover, the rule of thumb that considers hazard ratios, odds ratios, and relative risks of 2 or greater as robust can be misleading in being too low for studies missing an important confounder and being too high for studies that extensively control for confounding. In Chapter 4, we have worked to simplify the seminal work of Rosenbaum and Rubin (1983), Lin, Psaty, and Kronmal (1998), and Vanderweele and Ding (2017) to a formulation of a sensitivity to unmeasured confounding analysis that appeals to medical researchers. We offer guidelines to researchers for anchoring the tipping point analysis in the context of the study and provide examples.

CHAPTER 2

EXPLORING FINITE-SAMPLE BIAS IN PROPENSITY SCORE WEIGHTING CHOICES

2.1 Background

A primary goal of medical research is examining how a treatment will affect an outcome. This can be achieved via a randomized controlled trial, where participants are randomly assigned a treatment or control, and then their outcomes can be observed and directly compared. Alternatively, we can observe data that are collected on participants who are on the treatment or control for potentially nonrandom reasons, e.g. varying health care provider practices dependent on various patient characteristics. The latter setting, an observational study, is ubiquitous in the medical literature. Since deliberate randomization has not taken place, meaningful differences in observed covariates can exist between the treatment and control groups. These differences can bias the estimated effect of the treatment on the outcome of interest. Many popular methods to control for this utilize a propensity score model, estimating the probability that each participant would have received the treatment given observed pre-treatment covariates (Rosenbaum and Rubin 1983). These propensity scores can be incorporated in estimating the treatment-outcome effect in a variety of ways such as matching, stratification, adjusting, and weighting (D'Agostino 1998). This paper focuses on the final method, incorporating the conditional probability of treatment assignment in the treatment-outcome effect via propensity score weighting.

In a widely cited simulation study, Freedman and Berk (2008) explored the operating characteristics of propensity score weighting under specified conditions, with the propensity score model correctly specified and the outcome model incorrectly specified. The paper ultimately concludes that while the propensity score weighting does result in bias reduction, with realistic sample sizes the bias remains large (Freedman and Berk 2008). Despite attempts to refute this claim (Busso, DiNardo, and McCrary 2014; Busso, DiNardo, and McCrary 2009), and despite new methods for weighting (Li and Greene 2013; Li, Morgan, and Zaslavsky 2016) with improved efficiency and simple interpretability, propensity score weighting has not seen widespread use in research, perhaps partially due to these initial concerns. We replicate the simulations

of Freedman and Berk (2008) to examine the performance of two new weighting methods, ATM and ATO weights (Li and Greene 2013; Li, Morgan, and Zaslavsky 2016; Samuels 2017). We demonstrate that within the framework of this seminal paper that revealed poor finite-sample performance of weighting estimators, the ATM and ATO weights perform excellently.

2.1.1 Potential outcomes framework

The potential outcomes framework, first put forth by Neyman in 1923 (Neyman 1923; Imbens and Rubin 2015), and applied to the observational study setting by Rubin in 1974 (Rubin 1974), describes methods for estimating quantities based on unobserved, potential events. For example, in a study measuring the efficacy of a treatment, we would like to know how a participant's outcome would differ had they received the treatment versus the control. In the potential outcomes framework, we have two quantities, two potential outcomes, for each individual (i), their outcome if they had received the control, $Y_i(0)$, and their outcome had they received the treatment, $Y_i(1)$. The estimand of interest then is $Y_i(1) - Y_i(0)$, the difference in outcomes dependent on which treatment participant i received. It is almost always the case, however, that only one of these two paired potential outcomes $(Y_i(0), Y_i(1))$ is observed, dependent on whether you received the treatment (denoted as Z), or the control (denoted as $1 - Z$).

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \tag{2.1}$$

Often this is handled fundamentally as a missing data problem (Imbens and Rubin 2015). While only one of these paired outcomes is observed, we can estimate what the other, the counterfactual, would have been had the participant received the opposite treatment.

Ultimately, we are interested in some estimate of the treatment effect, for example we may want to know what the average treatment effect is across all participants, or we may want to know what the average treatment effect is among participants who received the treatment. At times these estimates will yield the same result, for example if the distribution of the conditional probability of receiving treatment is the same across both the treatment and control group, the average treatment effect across all participants will be the same as the average treatment effect among the treated.

The various potential estimators of interested are described below, along with the associated weighting schemes that can be used to estimate them.

2.1.1.1 Average treatment effect

The average treatment effect (ATE) is defined in the potential outcomes framework as the seen in Equation (2.2) (Imbens and Rubin 2015).

$$ATE = E[Y(1) - Y(0)] \tag{2.2}$$

In order to mimic the elegant results of a randomized controlled trial to observe an unbiased treatment effect in observational studies, we make the assumption that the outcomes under each treatment, $Y(1)$ and $Y(0)$, do not depend on which treatment was actually received, given the observed covariates. Formally, this is expressed as Equation (2.3)

$$(Y(1), Y(0)) \perp Z|\mathbf{X} \tag{2.3}$$

This is known as the strongly ignorable treatment assignment or the assumption of no unmeasured confounders (Imbens and Rubin 2015; D'Agostino 1998; Li, Morgan, and Zaslavsky 2016).

In the observational study setting, the probability of receiving treatment is no longer constant, as in a randomized controlled trial. We can use propensity scores to estimate this probability of treatment using observed pre-treatment covariates. Let Z be the treatment effect, where $Z = 1$ indicates the participant received the treatment and $Z = 0$ indicates the participant received the control. Pre-treatment covariates are denoted as $\mathbf{X}$. The propensity score, the conditional probability of receiving treatment given the observed covariates, is written as seen in Equation (2.4) (D'Agostino 1998; Austin and Stuart 2015; Li, Morgan, and Zaslavsky 2016).

$$e_i = P(Z_i = 1|\mathbf{X}) \tag{2.4}$$

When estimating the ATE, the target population is the whole population, both treated and controlled, and therefore inference is drawn with this in mind. The ATE can be estimated using the inverse probability weight of receiving treatment, derived from

the propensity score, e_i, as follows,

$$w_{ATE} = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i} \tag{2.5}$$

While this is often declared as the population of interest, it is not always the medically or scientifically appropriate population (Li, Morgan, and Zaslavsky 2016; Imbens and Wooldridge 2009; Rosenbaum 2012; Crump et al. 2009). Estimating the ATE assumes that every participant can be switched from their current treatment to the opposite (Li, Morgan, and Zaslavsky 2016). This is not always sensible, for example it may not be medically appropriate for every participant who didn't receive a treatment to receive it.

2.1.1.2 Average treatment effect among the treated

The average treatment effect among the treated (ATT) sets all subjects in the treated population to hold a weight of 1, and weights the control population accordingly. Here the inference is made with the treated group as the target population.

$$ATT = E[Y(1) - Y(0)|Z = 1] \tag{2.6}$$

The weights are defined as,

$$w_{ATT} = \frac{e_i Z_i}{e_i} + \frac{e_i(1 - Z_i)}{1 - e_i} \tag{2.7}$$

2.1.1.3 Average treatment effect among the controls

The average treatment effect among the controls (ATC) sets all subjects in the control population to hold a weight of 1, and weights the treated population accordingly. Here the inference is made with the control group as the target population.

$$ATC = E[Y(1) - Y(0)|Z = 0] \tag{2.8}$$

The weights are defined as,

$$w_{ATC} = \frac{(1 - e_i)Z_i}{e_i} + \frac{(1 - e_i)(1 - Z_i)}{1 - e_i} \tag{2.9}$$

2.1.1.4 Average treatment among the evenly matchable

The average treatment effect among the evenly matchable (ATM) was formally defined by Samuels as (Samuels 2017)

$$ATM_d = E[Y(1) - Y(0)|M_d = 1] \tag{2.10}$$

Like the ATE, ATT, and ATC, the ATM is a population average treatment effect. To define whether a subject is evenly matchable ($M_d = 1$) for a given matching process, d, for example propensity score caliper matching, consider a random sample from the general population. Let the sample size go to infinity. A subject is evenly matchable if the limit of the ratio of the number of subjects from the opposite treatment to the number from its own treatment is greater than 1 within the localized region of the covariate space around the subject defined by d. As with the ATE, ATT, and ATC, the population defined by the ATM is usually not observable; it is estimated from the observed sample. For the ATE, all of the subjects in the sample estimate the population. For the ATT, only the exposed subjects in the sample estimate the population. For the ATM, the population is estimated by the subjects who meet the evenly matchable criterion within the sample. If the ratio of exposed to unexposed subjects in a given region of the covariate space is exactly 1, without loss of generality, the exposed are considered the evenly matchable subjects. In practice, the estimated population is nearly equivalent to the cohort formed by one-to-one pair matching using d. Thus, the population defined by the ATM may be thought of as the population formed by pair matching. An exciting version of the ATM weight was introduced by Li and Greene as, (Li and Greene 2013)

$$w_{ATM} = \frac{\min\{e_i, 1 - e_i\}}{Z_i e_i + (1 - Z_i)(1 - e_i)} \tag{2.11}$$

Samuels demonstrates that these ATM weights are equal to the minimum of the ATT (Equation (2.7)) and ATC (Equation (2.9)) weights, $w_{ATM} = min(w_{ATT}, w_{ATC})$ (Samuels 2017).

2.1.1.5 Average treatment effect for the overlap population

The average treatment effect for the overlap population (ATO) was first formally introduced by Li, Morgan, and Zaslavsky (Li, Morgan, and Zaslavsky 2016). The

ATO is not easy to describe in the same way as the estimators above – it essentially creates a pseudo-population that has excellent variance properties. In practice, the reference populations created by ATM and ATO weights will look similar due to the weights themselves being similar. In best practice, a paper using either method should include a detailed description of this reference population, for example via a thorough "Table 1".

The weights are defined as,

$$w_{AT0} = (1 - e_i)Z_i + e_i(1 - Z_i) \tag{2.12}$$

Li et al demonstrate how this compares to known estimators showing when the conditional probability of receiving treatment is small ($e_i \approx 0$), the w_{ATO} approximates the w_{ATT}.

$$(1 - e_i, e_i) \approx \left(1, \frac{e_i}{1 - e_i}\right)$$

Similarly, if the conditional probability of receiving control is small ($1 - e_i \approx 0$), the w_{AT0} approximates the w_{ATC}.

$$(1 - e_i, e_i) \approx \left(\frac{1 - e_i}{e_i}, 1\right)$$

Finally, if there is nearly a 50-50 chance of receiving treatment or control ($e_i \approx 0.5$), in other words the treatment and control groups are balanced in distribution and size, the w_{ATO} approximates the w_{ATE}.

$$(1 - e_i, e_i) \approx \left(\frac{0.25}{e_i}, \frac{0.25}{1 - e_i}\right)$$

2.2 Doubly Robust Estimators

In order to generate the estimators described above, we first fit a propensity score model to estimate e_i, and then fit an outcome model, applying the weight specified for the given estimator of interest. A doubly robust estimator is one that is robust when either the propensity score model *or* the outcome model is correctly specified (Scharfstein, Rotnitzky, and Robins 1999; Robins 2000; Robins, Rotnitzky, and Laan

2000; Laan and Robins 2003; Neugebauer and Laan 2005). Doubly robust estimators for the ATE have been described in great detail (Lipsitz, Ibrahim, and Zhao 1999; Lunceford and Davidian 2004; Neugebauer and Laan 2005; Bang and Robins 2005; Kang and Schafer 2007; Robins et al. 2007; Robins, Rotnitzky, and Zhao 2012; Funk et al. 2011). The doubly robust ATE is estimated by first estimating e_i via the propensity score model, $e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$, then fitting the outcome model separately among the treated population and the control population, and using these models to predict outcome values, $(\hat{Y}_i(1) = m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1), \hat{Y}_i(0) = m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0))$, for all participants. These estimates are then combined using the weights specified above. This can be generalized to the following doubly robust estimator for any of the above stated quantities, resulting in the augmented estimator, $\hat{\Delta}_{DR,w}$.

$$\begin{aligned}\hat{\Delta}_{DR,w} =& \frac{\sum_{i=1}^{n} w_i(m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1) - m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n} w_i} + \\ & \frac{\sum_{i=1}^{n} w_i Z_i(Y_i - m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1))}{\sum_{i=1}^{n} w_i Z_i} - \frac{\sum_{i=1}^{n} w_i(1 - Z_i)(Y_i - m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n} w_i(1 - Z_i)}\end{aligned} \tag{2.13}$$

Where w represents the weight for the estimator of interest. For example, if we were interested in the ATO estimate, we would use w_{ATO} for each participant. The doubly robust estimator has been formally proven for the ATE (Lunceford and Davidian 2004) and ATM (Li and Greene 2013). The proof for the ATO logically follows, however we include a formal derivation in Chapter 3.

Using the framework described in Freedman and Berk (2008), we will examine the properties of the doubly estimators for the ATE, ATO, and ATM under various scenarios. R code for computing these weights and doubly robust estimators is included in Appendix A.

2.3 Methods

Simulations are conducted using R version 3.4.3 (R Core Team 2017).

2.3.1 Freedman and Berk simulation setting

Freedman and Berk set up their simulation with X as the exposure and $\mathbf{Z}$ as the pre-treatment covariates. This is the opposite of how we have defined these quantities above, as we have attempted to remain consistent with the majority of the potential outcomes literature. For simplicity, we are changing their notation to match ours used elsewhere.

2.3.1.1 Continuous outcome

The outcome model is a linear model with normally distributed errors, $U \sim N(0, 1)$ defined as

$$Y = 1 + Z + X_1 + 2X_2 + U \tag{2.14}$$

The propensity score model is defined as a probit selection model with normally distributed errors, $V \sim N(0, 1)$.

$$Z = \begin{cases} 1 & 0.5 + 0.25X_1 + 0.75X_2 + V > 0 \\ 0 & o.w. \end{cases} \tag{2.15}$$

$\mathbf{X}$ is bivariate normal, defined as $\mathbf{X} \sim MVN\left(\begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$.

This setup results in populations with about 83.12% exposed to the treatment and 16.88% controlled.

Drawing from the distributions stated above, we fix the propensity score model as specified, but estimate the outcome model leaving out the confounder X_2 to examine the effect this has on the bias of the estimated coefficient for our exposure of interest, Z. We run this simulation, varying the sample size by 10 from 100 to 10,000 to examine the properties and rate of convergence of this finite-sample bias using 50,000 simulations for sample sizes 100 to 1,000 and 10,000 simulations at each subsequent sample size. In addition to the w_{ATE} we examine the w_{ATM} and w_{ATO}. In this continuous setting, the true ATE, ATM, and ATO are all the same, the coefficient specified for Z in the outcome model specified, 1, therefore all models will be compared to this true value.

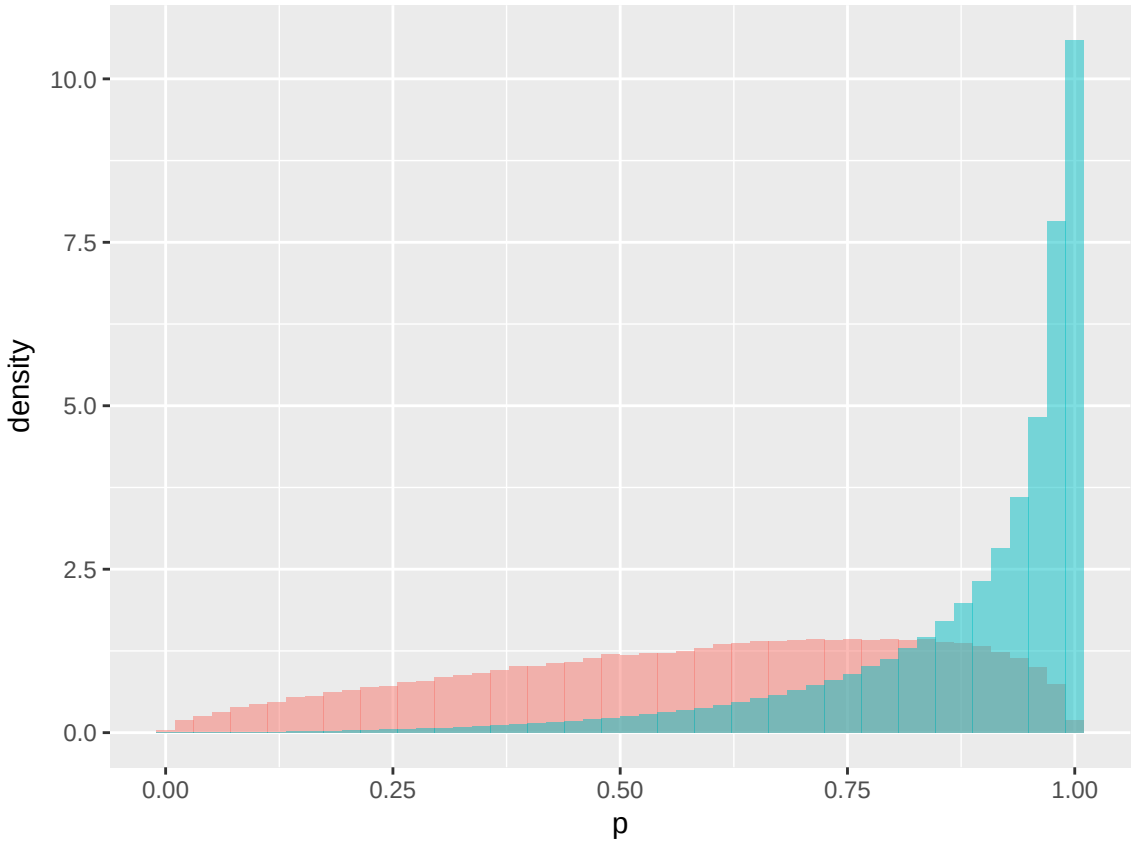

Figure 2.1: Distribution of the propensity score for the continous outcome model. The blue represents the exposed population, where $Z = 1$, the red represents where $Z = 0$, and the grey represents the overlap between the two.

We examine the bias for the exposure effect using these three weighting schemes: w_{ATE}, w_{ATM} and w_{ATO}. Freedman and Berk (2008) reports a bias of 1.131 in the unweighted model with X_2 excluded from the outcome model, and a remaining bias of 0.1366 using the ATE weights with the sample size, n, set to 1,000 and 250 simulations. While we do not have the original paper's simulation code, under the same model specifications and sample size of 1,000, we see similar biases. In our unweighted model, we observe a bias of 1.129 and in our model weighted with ATE weights, we observe a bias of 0.143. Using the ATM and ATO weights, however, we observe attenuated biases of -0.001 and -0.001, respectively. Figure 2.2 displays this relationship over sample sizes ranging from 100 to 10,000. While there is a clear finite-sample bias present for the ATE weights, the ATM and ATO weights have negligible bias. Even at a quite large sample (n = 10,000), the ATE weights remain slightly biased. It has been pointed out that this is likely due in part to the particular simulation setting chosen by Freedman and Berk (Busso, DiNardo, and McCrary 2014), however it is worth noting that not all weighting methods are subject to this finite-sample bias, and indeed it is nearly negligible when using the ATM and ATO weights.

Additionally, Freedman and Berk (2008) comments on the standard error of the w_{ATE} method, demonstrating that in the case where the outcome model is correctly specified and there is no bias to reduce, the w_{ATE} adjusted models are inefficient, and in fact the standard error is double that of the unweighted model. The standard error here refers to the observed standard deviation across the 10,000 simulations. In their simulation with a sample size of 1,000, the standard error of the exposure coefficient in the unweighted model is 0.0974 and the standard error of the exposure coefficient in the w_{ATE} weighted model is 0.2130 when the outcome model is correctly specified. We observe similar values in our simulation, 0.096 and 0.22 for the unweighted and ATE weighted models, respectively. The standard error for the ATM and ATO weighted models, however, are much more efficient, with standard errors of 0.1 and 0.102, respectively. Figure 2.3 demonstrates the relationship between the standard error by sample size and weighting method. Again, the ATM and ATO weights represent a superior method.

2.3.1.2 Binary outcome

In the logistic case, the story is considerably more complicated. Here the outcome model has errors with a standard logistic distribution, $U \sim logis(0, 1)$. The model is specified as,

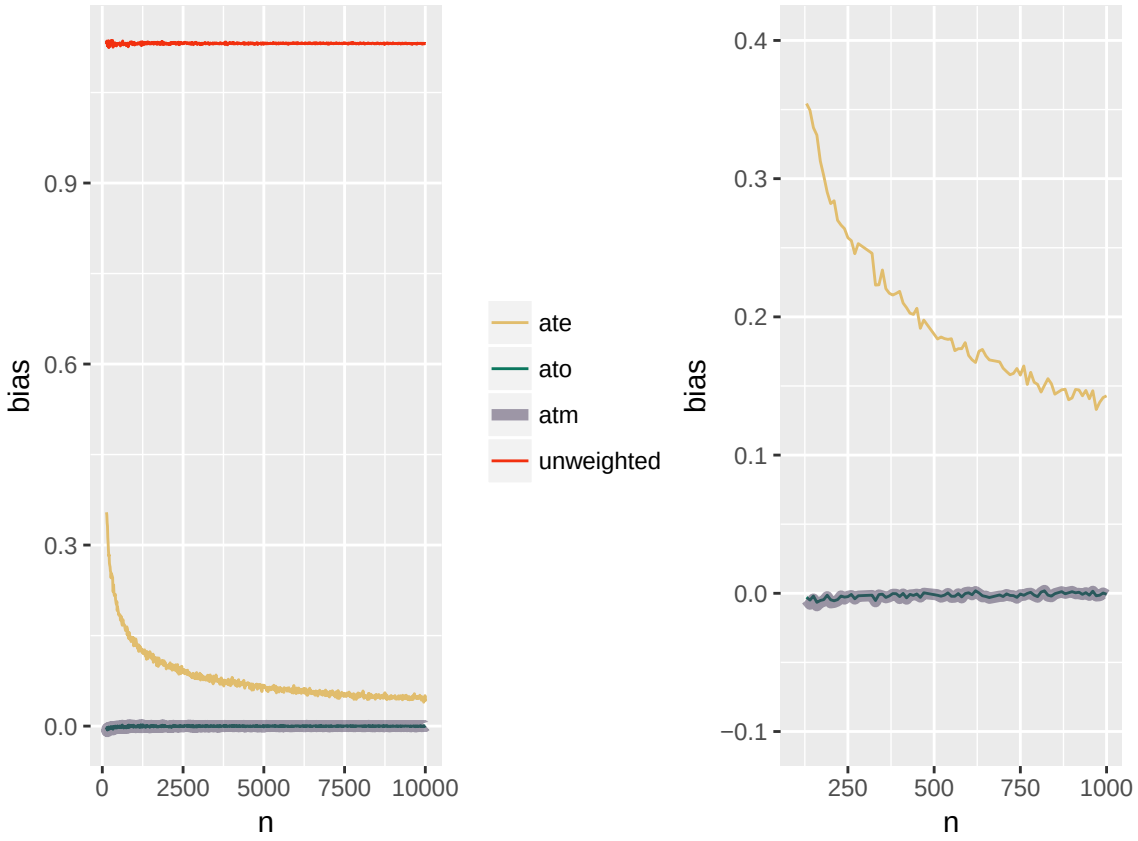

Figure 2.2: Bias in the causal estimate, introduced by excluding a confounder, X_2, from the continuous outcome model, by sample size. The unweighted estimate (red) has the most bias, the ATE weighted estimate (yellow) is next, and finally the ATO (green) and ATM (thick, purple) weighted estimates show little to no bias. The left panel shows the full plot, the right panel zooms in on the weighted estimates for n between 100 and 1,000.

$$Y = \begin{cases} 1 & 1 + Z + X_1 + 2X_2 + U > 0 \\ 0 & o.w. \end{cases} \tag{2.16}$$

The propensity score model errors also have a standard logistic distribution, $V \sim logis(0, 1)$, and the model is specified as,

$$Z = \begin{cases} 1 & 0.5 + 0.25X_1 + 0.75X_2 + V > 0 \\ 0 & o.w. \end{cases} \tag{2.17}$$

Same as the continuous setting, $\mathbf{X}$ is bivariate normal, defined as $\mathbf{X} \sim MVN\left(\begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$.

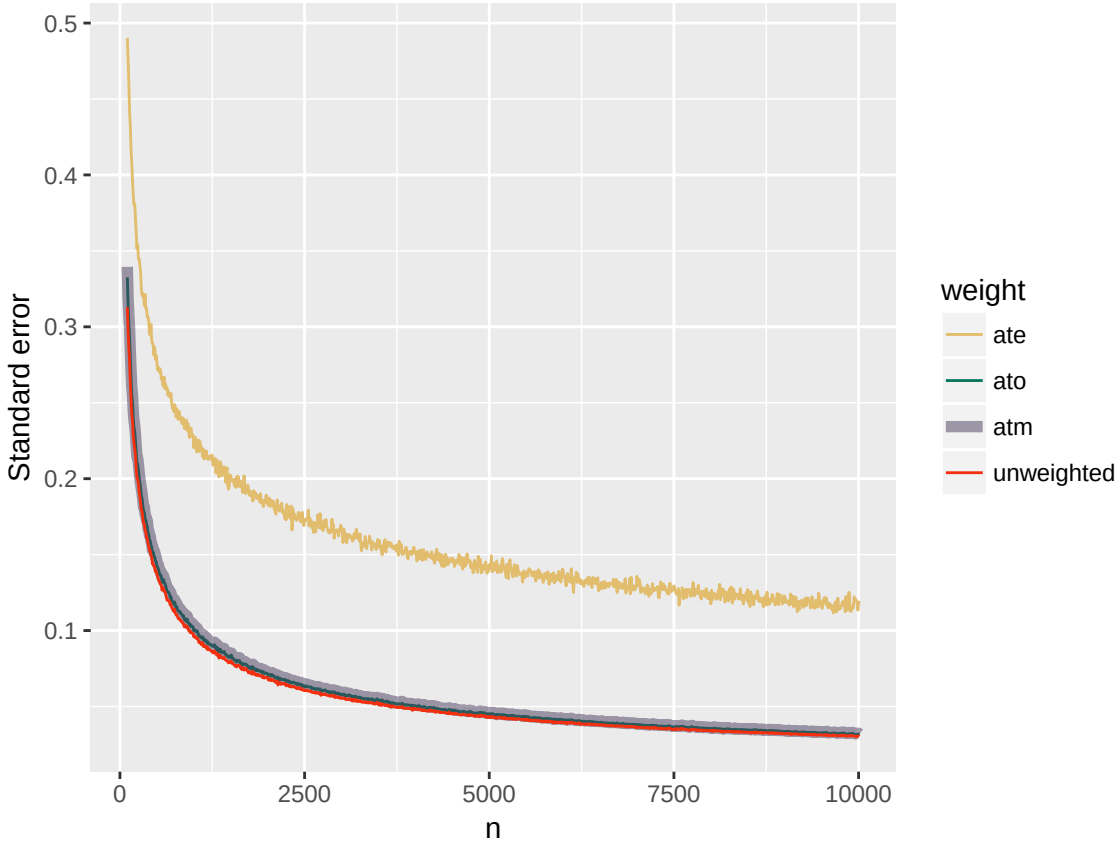

Figure 2.3: Standard error of the causal estimate by sample size. The ATE has the largest variability (yellow), with the ATM (thick, purple) and ATO (green) with a smaller variability, almost directly on top of eachother (the ATO is slightly below the ATM), and the unweighted (red) just below that.

This simulation setting results in about 86.85% outcome events with about 75.75% of the population exposed.

Freedman and Berk (2008) states,

> "The bad behavior of the weighted simple logistic regression is not a small-sample problem. It is quite reproducible. We think it is due to occasional large weights." (Freedman and Berk 2008)

While it is true that this issue is not *only* a small-sample problem, the constant bias seen across large sample sizes may indeed be due to the collapsibility of the odds ratio they are drawing inference on. In the case where the odds ratio will approximate the relative risk, this indeed will yield similar results as the continuous case, the bias noticed will be due to a small-sample problem. In the case where the odds ratio does not approximate the relative risk, as we have in this simulation where the outcome is not rare, the relationship is more nuanced. In fact, in these scenarios, it has

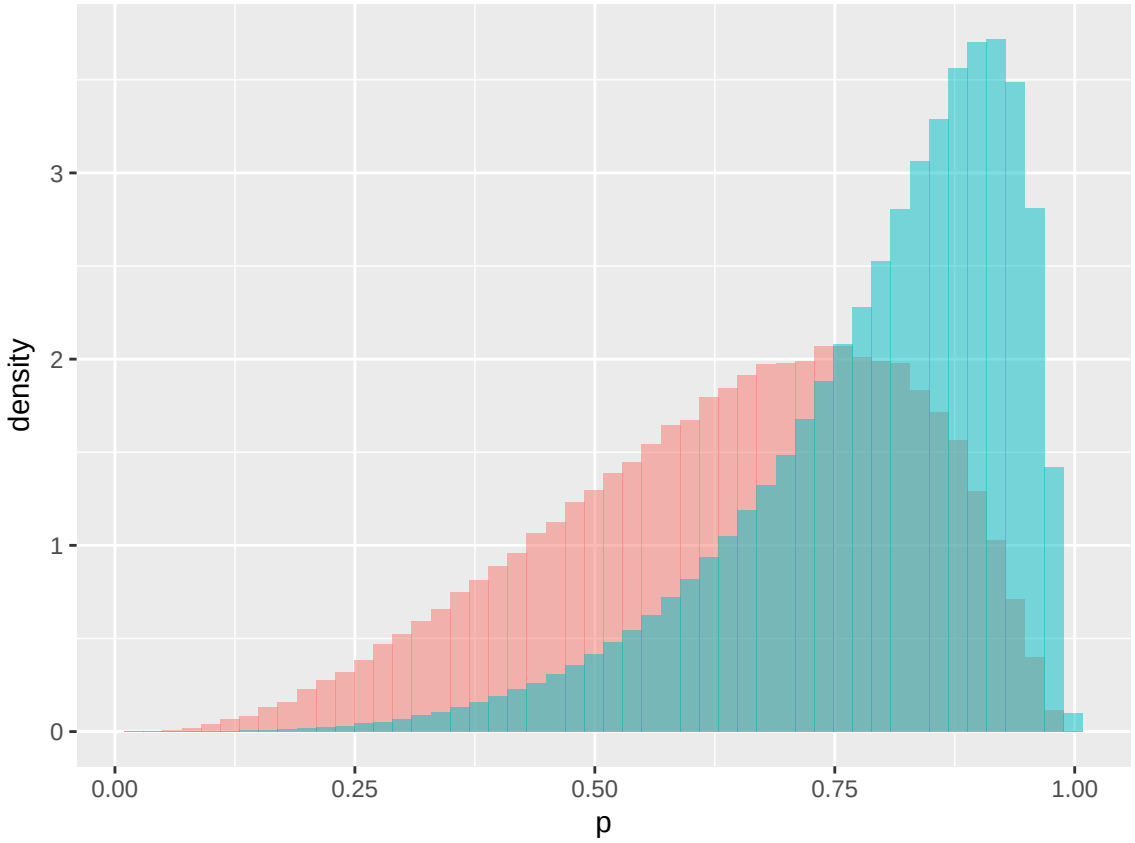

Figure 2.4: Distribution of the propensity score for the binary outcome model. The blue represents the exposed population, where $Z = 1$, and red represents where $Z = 0$.

been suggested by VanderWeele and others that perhaps logistic regression should be avoided all together, and rather a log-linear model should be fit (VanderWeele 2015). In order to adequately examine the properties here, rather than examining the odds ratio, we examine the mean difference in the probability of the outcome, again using the doubly robust estimators for the weighted models. By difference in the probability of the outcome, we mean we are estimating the following for each individual,

$$P(Y_i(1) = 1) - P(Y_i(0) = 1)$$

and then taking the average across all individuals. Our estimator of interest is on the probability scale, and therefore is no longer a linear estimator, therefore the "true" values for the ATE, ATM, and ATO will differ. To illustrate this, Figure 2.5 plots the distribution of these risk differences in each population, ATE, ATM, and ATO, respectively.

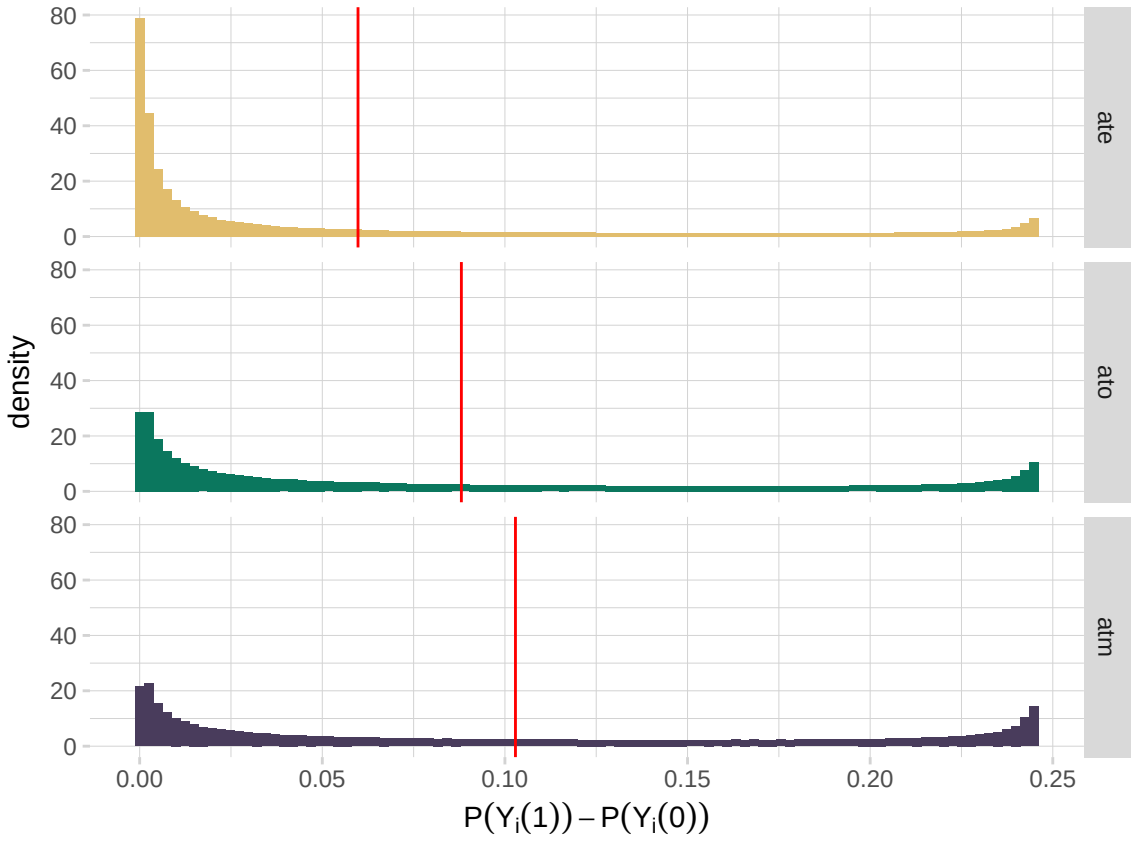

Figure 2.5: Distribution of risk difference for each population, ATE, ATM, and ATO. Red line indicates the mean risk difference within each population.

Drawing from the distributions stated in Equations (2.16) and (2.17), as with the continuous outcome model we fix the propensity score model as specified, but estimate

the outcome model leaving out the confounder X_2 to examine the effect this has on the bias of the estimated effect of our exposure of interest, Z. Again, we run this simulation varying the sample size by 10 from 100 to 10,000. We use 50,000 simulations for sample sizes 100 to 1,000, and 10,000 simulations for all subsequent sample sizes. We compare each method to its "true" value for the mean risk difference, 0.05938, 0.08698, and 0.10156 for the ATE, ATO, and ATM, respectively. Here, all of the weighting schemes perform similarly (Figure 2.6). This is likely because this is an easier setting for the ATE to perform well. The bias due to missing the confounder X_2 is smaller in this setting. Additionally, the propensity model is specified in a slightly different manner, with logit errors versus normal, causing the underlying model to have more overlap in propensity scores between the treated and control population in the binary outcome case versus the continuous (Figure 2.1, under the probit propensity score model, compared to Figure 2.4 under the logit propensity score model).

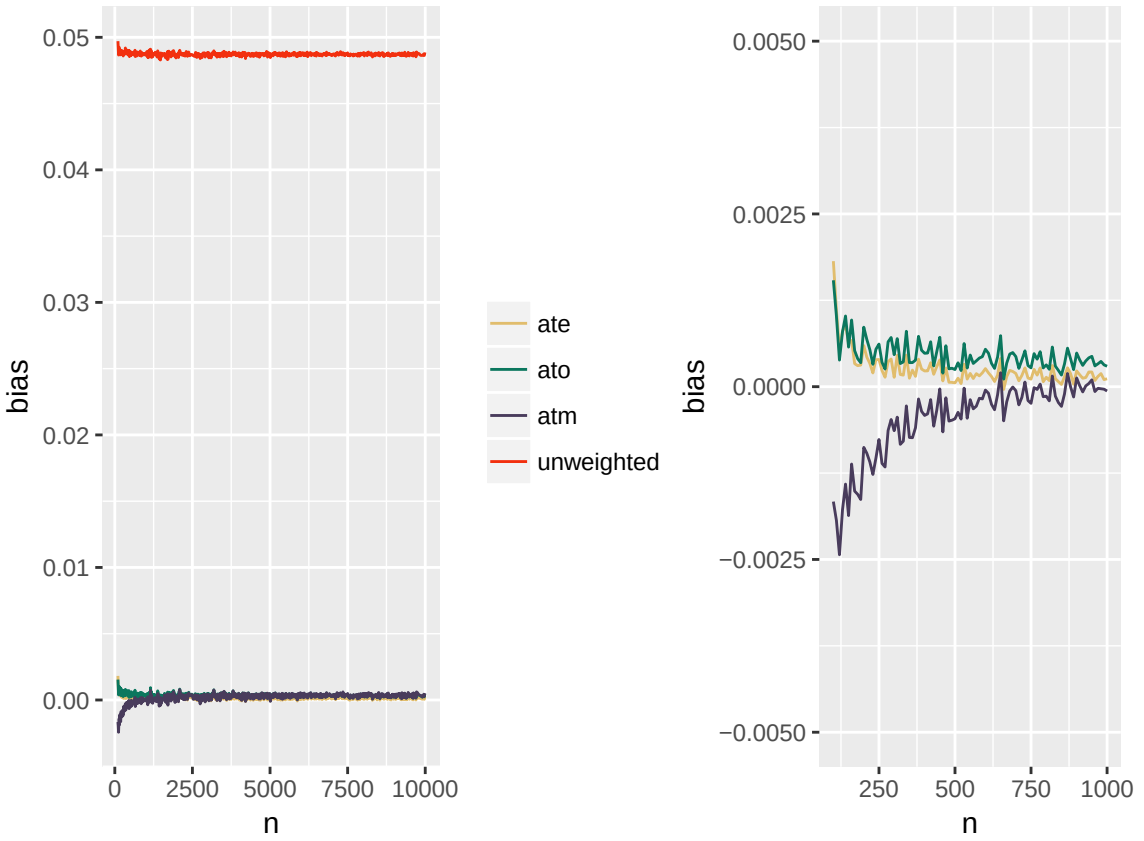

Figure 2.6: Bias in the causal estimate, introduced by excluding a confounder, X_2, from the binary outcome model, by sample size. The unweighted estimate (red) has the most bias, the ATE weighted estimate (yellow), ATO (green), and ATM (purple) weighted estimates overlap, showing little to no bias. The left panel shows the full plot, the right panel zooms in on the weighted estimates for n between 100 and 1000.

2.3.1.3 Binary Outcome (revised)

In replicating the setup for the binary outcome model specified in Freedman and Berk (2008), we were able to replicate the paper's original findings but found surprisingly good overlap between the propensity scores for the treatment and control group (Figure 2.4), as well as a relatively small mean risk difference (0.05938, 0.08698, and 0.10156 for the ATE, ATO, and ATM, respectively). We revise this simulation to decrease the overlap, using the same probit model for the propensity score distribution (Equation (2.15)), as specified in the continuous case, as well as update the outcome model by changing the coefficient for Z from 1 to 3. As previously specified, the errors have a standard logistic distribution, $U \sim logis(0, 1)$. The outcome model is specified as,

$$Y = \begin{cases} 1 & 1 + 3Z + X_1 + 2X_2 + U > 0 \\ 0 & o.w. \end{cases} \tag{2.18}$$

Same as the previous settings, $\mathbf{X}$ is bivariate normal, defined as $\mathbf{X} \sim MVN\left(\begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$.

This simulation setting results in about 90.72% outcome events with about 83.16% of the population exposed. The distribution of the propensity scores here are identical to those specified in the continuous setting (Figure 2.2).

We conduct the same simulation as the binary case, varying the sample size by 10 from 100 to 10,000 with 50,000 simulations for sample sizes 100 to 1,000, and 10,000 simulations for all of the subsequent sample sizes. We compare each method to it's "true" value for the mean risk difference, 0.1294, 0.2727, and 0.3104 for the ATE, ATO, and ATM, respectively (Figure 2.7). The weighting schemes now perform similarly to the continuous setting (Figure 2.8). The ATE weights take slight longer to converge to their "true" value, where as the ATM and ATO weights converge a bit faster, although the difference is not as noticeable as the continuous case.

2.4 Discussion

While we replicated the results seen in Freedman and Berk (2008) for the ATE weights, we demonstrate that this seminal paper ought not to be viewed as the final word on

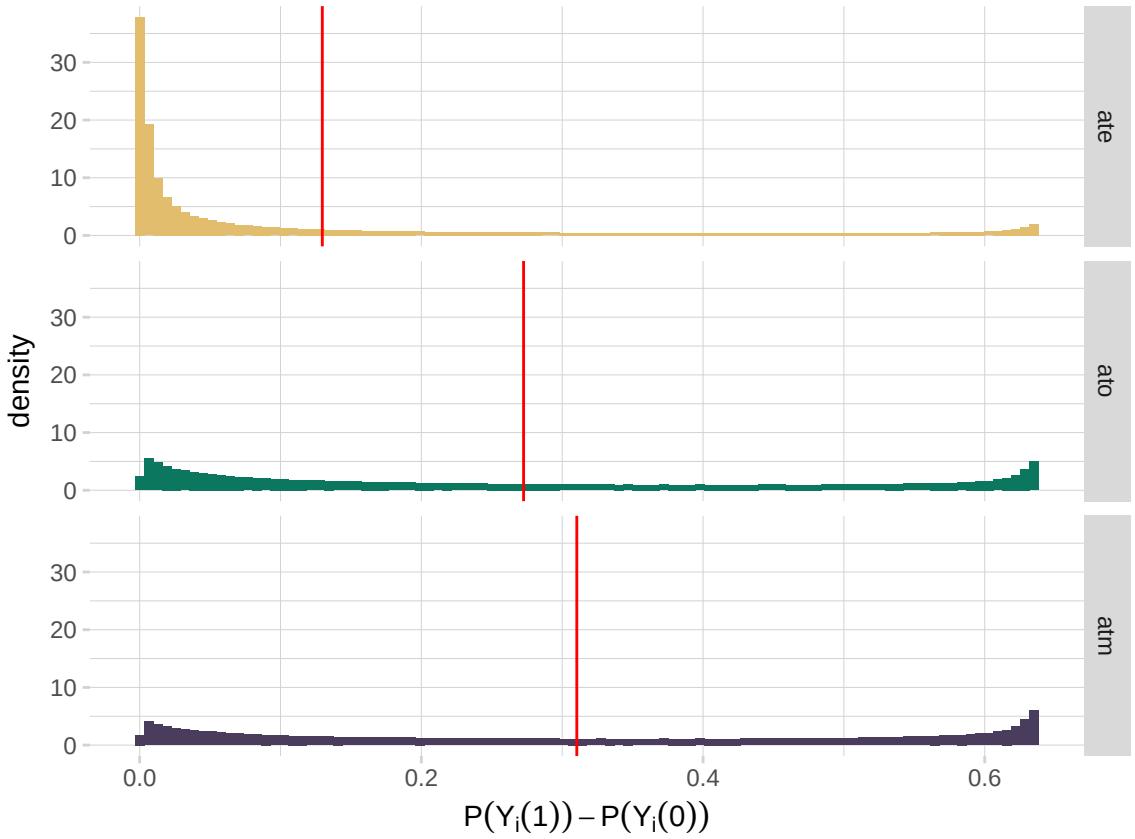

Figure 2.7: Distribution of risk difference for each population, ATE, ATM, and ATO in the revised binary outcome setting. Red line indicates the mean risk difference within each population.

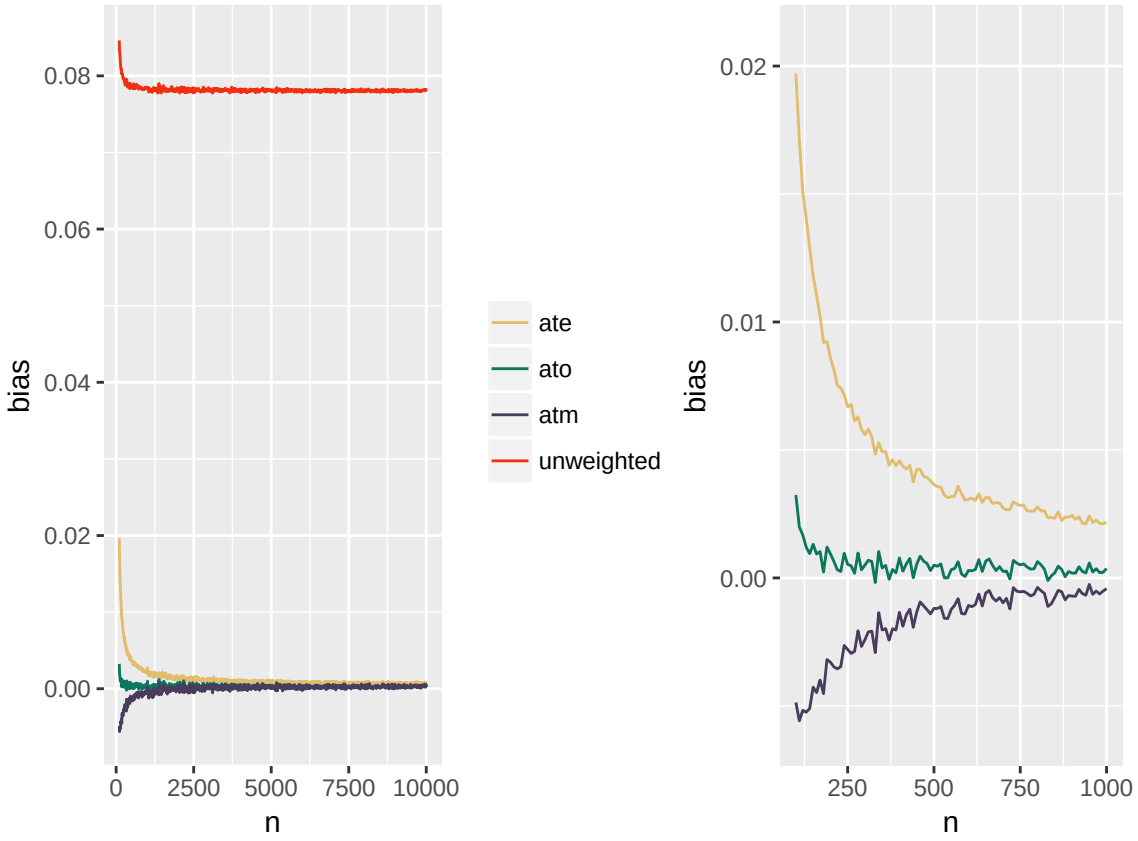

Figure 2.8: Bias in the causal estimate, introduced by excluding a confounder, X_2, from the revised binary outcome model, by sample size. The unweighted estimate (red) has the most bias, the ATE weighted estimate (yellow) the next most, followed by the ATM (purple), and finally ATO (green). The left panel shows the full plot, the right panel zooms in on the weighted estimates for n between 100 and 1000.

the use of weighting methods in general. In particular, new more stable weighting schemes, ATM and ATO weights, do not suffer from the poor finite-sample properties seen in the ATE weights. Thus, we have demonstrated that the utility of weighting depends upon the weighting method used. These findings have important implications on how analyses should be performed in medical research, particularly with the large number of observational studies emerging in the literature.

Through the two simulation settings presented in Freedman and Berk (2008), we have also demonstrated that the degree of overlap in the propensity scores between the treatment and control groups is important when assessing the preferable weighting method. In the first (continuous) example, the setting is "favorable to weighting" because the propensity score model is correctly specified, it is unfavorable for the ATE weights because the degree of overlap in propensity scores is poor (Busso, DiNardo, and McCrary 2009). It has been shown elsewhere that under more favorable conditions

(i.e. strong overlap in propensity scores between the treatment and control groups), the ATE weights do quite well (Busso, DiNardo, and McCrary 2009; Busso, DiNardo, and McCrary 2014). It turns out, as we demonstrate, that stable weights, such as the ATM and ATO weights, work quite well in both settings. In the continuous outcome setting explored here, the ATM and ATO estimands are the same as the ATE, since we are dealing with a linear outcome and a constant treatment effect. However, this no longer holds when moving into the binary outcome. There ATE, ATM, and ATO are different quantities because the standardizing population influences the average risk difference. This simulation setting is notably different from the continuous setting. When comparing Figure 2.1 and Figure 2.4, although the coefficients for the propensity score model are the same, we notice the degree of overlap is greater due to the underlying propensity score model having a logit distribution, as opposed to a probit distribution in the continuous case. Because of this greater degree of overlap, the ATE, ATM, and ATO have similar characteristics when comparing their finite-sample bias. In settings such as this, where the degree of overlap in propensity scores between the treatment and control group is adequate, the choice of weighting method becomes one of preference, often based on the population of interest. When overlap is adequate such that it is plausible that anyone in the population could have received either exposure with a nontrivial probability, the ATE may be philosophically appealing. However, when overlap is poor such that the researchers could see a one-to-one matched cohort as a compelling design, the ATM weights become more appealing. We examine a more extreme binary outcome case than that of the original study, in the sense of less overlap between the propensity score distributions and a larger treatment-outcome effect. We see a similar, but less extreme, result as in the continuous case, further solidifying our claim that the ATM and ATO weights do not suffer from this finite-sample bias.

In addition to demonstrating the decreased finite-sample bias in the ATM and ATO estimators, we have also shown that these methods have lower variance. Freedman and Berk (2008) states that "weighting is likely to increase random error by a substantial amount"; while we are able to replicate the magnitude of the variability they see using the ATE weights in the continuous case, we show that the variability for the ATM and ATO weights is nearly the same as those of the correctly specified unweighted model. The correctly specified unweighted model, as shown in Figure 2.3, has the minimum possible variance, as it is fit using ordinary least squares. The ATO estimator has nearly identical standard errors, while the ATM standard errors are only slightly larger. A topic of future research will be exploring methods for estimating these variances

appropriately under various scenarios.

We demonstrate compelling evidence to show that the ATM and ATO weights perform well in the simulation settings presented in Freedman and Berk (2008). However, there are other simulation settings which have not been explored, and therefore need further research. For example, this paper does not explore the implications of having neither model correctly specified. Nevertheless we have demonstrated that the conclusions of the original study should be reconsidered in light of these new weighting estimates; propensity score weighting does not unilaterally suffer from finite-sample bias. Therefore, weighting should be considered as a viable option in research studies that use propensity score methodology.

2.5 Conclusion

Revisiting the setup of the seminal paper Freedman and Berk (2008), where poor finite sample properties of ATE weights were revealed, we demonstrate that ATM and ATO weights have excellent finite sample properties. We hope that these findings will persuade researchers to reconsider the benefits of propensity score based weighting methods.

2.6 Appendix A

The following provides a coding walk through in R for calculating these doubly robust estimators. The data are generated using a single simulation from the settings detailed in this paper, with n = 1,000. There are two datasets, the first simulated with a continuous outcome, and the second simulated with a binary outcome. The simulation code and datasets are provided on GitHub (https://github.com/LucyMcGowan/dr-example-code).

2.6.1 Continuous outcome

```
df_url <- "http://bit.ly/df_continuous"
load(url(df_url))

## Fit the propensity score ----
```

```r
p_1 <- predict(
  glm(z ~ x_1 + x_2,
      data = df_continuous,
      family = binomial("probit")),
  type = "response"
)

## Create weights ----

## Calculate the probability of receiving control
p_0 <- 1 - p_1

### Calculate the probability of being assigned the treatment
### you received
p_assign <- ifelse(df_continuous$z == 1, p_1, p_0)

### ATE
w_ate <- 1 / p_assign

### ATM
w_atm <- pmin(p_1, p_0) / p_assign

### ATO
w_ato <- 1 - p_assign

## Fit outcome models ----

m_1 <- predict(
  glm(y ~ x_1, data = df_continuous[df_continuous$z == 1, ]),
  newdata = df_continuous
)

m_0 <- predict(
  glm(y ~ x_1, data = df_continuous[df_continuous$z == 0, ]),
  newdata = df_continuous
)
```

The DR estimator is estimated using Equation (2.13), plugging in the given weight for w_i.

$$\hat{\Delta}_{DR,w} = \frac{\sum_{i=1}^{n} w_i (m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1) - m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i Z_i (Y_i - m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1))}{\sum_{i=1}^{n} w_i Z_i} - \frac{\sum_{i=1}^{n} w_i (1 - Z_i)(Y_i - m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n} w_i (1 - Z_i)}$$

For example, we can plug in `w_atm` to get $\hat{\Delta}_{DR,ATM}$

```
(dr_atm <- (sum(w_atm * (m_1 - m_0)) / sum(w_atm)) +
    (sum(w_atm * df_continuous$z * (df_continuous$y - m_1)) /
       sum(w_atm * df_continuous$z)) -
    (sum(w_atm * (1 - df_continuous$z) * (df_continuous$y - m_0)) /
       sum(w_atm * (1 - df_continuous$z)))
)
```

```
## [1] 1.003834
```

Rather than re-typing this equation each time, we can create a function that will calculate it.

```
dr <- function(weight, y, m_1, m_0, z) {
  (sum(weight * (m_1 - m_0)) / sum(weight)) +
    (sum(weight * z * (y - m_1)) / sum(weight * z)) -
    (sum(weight * (1 - z) * (y - m_0)) / sum(weight * (1 - z)))
}

dr(w_ate, df_continuous$y, m_1, m_0, df_continuous$z)
```

```
## [1] 1.381388
```

```
dr(w_atm, df_continuous$y, m_1, m_0, df_continuous$z)
```

```
## [1] 1.003834
```

```
dr(w_ato, df_continuous$y, m_1, m_0, df_continuous$z)
```

```
## [1] 1.0054
```

2.6.2 Binary outcome

```
df_url <- "http://bit.ly/df_binary"
load(url(df_url))

## Fit the propensity score ----
p_1 <- predict(
  glm(z ~ x_1 + x_2, data = df_binary, family = binomial),
  type = "response"
)

## Create weights ----

## Calculate the probability of receiving control
p_0 <- 1 - p_1

### Calculate the probability of being assigned the treatment
### you received
p_assign <- ifelse(df_binary$z == 1, p_1, p_0)

### ATE
w_ate <- 1 / p_assign

### ATM
w_atm <- pmin(p_1, p_0) /p_assign

### ATO
w_ato <- 1 - p_assign

## Fit outcome models ----

m_1 <- predict(
  glm(y ~ x_1,
      data = df_binary[df_binary$z == 1, ],
      family = binomial),
  newdata = df_binary,
```

```
  type = "response"
)

m_0 <- predict(
  glm(y ~ x_1,
      data = df_binary[df_binary$z == 0, ],
      family = binomial),
  newdata = df_binary,
  type = "response"
)
```

Using the function defined above, we calculate the DR estimator.

```
dr(w_ate, df_binary$y, m_1, m_0, df_binary$z)
```

```
## [1] 0.06305217
```

```
dr(w_atm, df_binary$y, m_1, m_0, df_binary$z)
```

```
## [1] 0.1056484
```

```
dr(w_ato, df_binary$y, m_1, m_0, df_binary$z)
```

```
## [1] 0.09410479
```

CHAPTER 3

DOUBLY ROBUST AND LARGE SAMPLE VARIANCE ESTIMATOR FOR OVERLAP WEIGHTS

3.1 Background

In 2016, Li et al proposed an overlap weight, a method for propensity score weighting with improved variance properties (Li, Morgan, and Zaslavsky 2016). In Chapter 2, we explored the improved finite sample properties of a doubly robust estimator for the average treatment effect for the overlap population (ATO). In this chapter, we prove the doubly robust property of this estimator, that is a form that is robust when either the propensity score model *or* the outcome model is correctly specified (Scharfstein, Rotnitzky, and Robins 1999; Robins 2000; Robins, Rotnitzky, and Laan 2000; Laan and Robins 2003; Neugebauer and Laan 2005). In addition we derive a large-sample variance estimator for the IPW and doubly robust ATO estimator, extending the Williamson variance (Williamson, Forbes, and White 2013), similar to the sandwich estimator proposed for the ATE by Lunceford and Davidian (Lunceford and Davidian 2004) as well as the sandwich estimator proposed for the ATM by Li and Greene (Li and Greene 2013). Our extension can by applied to outcomes modeled with a generalized linear model with a identity, log, or logistic link. We then use a simulation setup similar to that in Chapter 2 to compare this doubly robust estimator and large-sample variance estimator to two conditions, a naive model and a naive sandwich estimator.

For calculating the uncertainty of an estimator that incorporates both propensity scores and conditional outcome models, we demonstrate the utility of a large-sample variance estimator that accounts for the propensity scores being themselves estimated. Alternatively, the bootstrap has been recommended to acquire the appropriate standard errors (Li, Morgan, and Zaslavsky 2016; Li and Greene 2013; Funk et al. 2011). We focus on weighting as the method for incorporating the propensity score, however there are other ways such as matching, stratification, or covariate adjustment. McCandless et al. (2009) introduce a Bayesian propensity score methodology that adjusts for the propensity score as a covariate, allowing for the incorporation of the uncertainty associated with fitting the propensity score model (McCandless, Gustafson, and Austin

2009). Zigler et al. (2013) highlight the challenges that can occur with combining the propensity score estimation and outcome model into a joint estimation process within a Bayesian framework. They present strategies for augmenting the propensity score adjustment to prevent these problems (Zigler et al. 2013). Zigler (2016) addresses the natural tension in using Bayesian methods to estimate causal effects and elucidates how propensity score adjustment fits into the Bayesian framework (Zigler 2016).

3.2 Methods

3.2.1 ATO estimator and large-sample variance

The estimator of the average treatment effect for the overlap population (ATO) is as follows (Li, Morgan, and Zaslavsky 2016).

$$\hat{\tau}_{ATO} = \frac{\sum_{i=1}^{n}(1 - e(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))Z_i Y_i}{\sum_{i=1}^{n}(1 - e(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))Z_i} - \frac{\sum_{i=1}^{n} e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})(1 - Z_i)Y_i}{\sum_{i=1}^{n} e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})(1 - Z_i)} \tag{3.1}$$

where $e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ is the estimated propensity score, Z_i is the indicator for treatment, and Y_i is the observed outcome. The first term above can be defined as $\hat{\mu}_1$ and the second term as $\hat{\mu}_0$.

The point estimate can be estimated by plugging in the propensity score obtained from a logistic regression for $e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$. Alternatively, as Williamson et al. mention, a generalized linear model can be fit for the outcome on the treatment, applying the ATO weight, $Z_i(1-e_i)+(1-Z_i)e_i$, and appropriate link function (Williamson, Forbes, and White 2013). Of note, this is distinct from the doubly robust estimator in that here the outcome model does not include any additional covariates, only the treatment indicator.

Williamson et al. demonstrate a general form for a variance estimator for inverse probability of treatment weighted estimators (estimators using ATE weights) (Williamson, Forbes, and White 2013). Their large-sample variance estimator, was generalized to include models fit with the identity, log, or logit link function. These large sample variance estimators are distinct from a "naive" sandwich estimator calculated using only the outcome model in that they account for the propensity score estimation. When the propensity score model is correctly specified, incorporating the estimation of the propensity score model in the variance calculation leads to smaller, more accurate

variances (Williamson, Forbes, and White 2013; Funk et al. 2011).

Following the process of Williamson et al., we can solve the following estimating equations $\sum_{i=1}^{n} \mathbf{u}(\boldsymbol{\theta}(Y_i, Z_i, X_i)) = 0$ for $\boldsymbol{\theta} = (\mu_1, \mu_0, \boldsymbol{\beta}^T)^T$ to calculate $\hat{\tau}_{ATO}$ (Equation (3.1)) where,

$$\mathbf{u}(\boldsymbol{\theta}; Y, Z, \mathbf{X}) = \begin{pmatrix} (Y - \mu_1)Z(1 - e(\mathbf{X}, \boldsymbol{\beta})) \\ (Y - \mu_0)(1 - Z)e(\mathbf{X}, \boldsymbol{\beta}) \\ \mathbf{X}(Z - e(\mathbf{X}, \boldsymbol{\beta})) \end{pmatrix} \tag{3.2}$$

This estimator is an M-estimator with an asymptotically normal distribution. Using the delta method, the estimator has a large-sample variance equal to:

$$\text{var}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-T} \tag{3.3}$$

where $\mathbf{A} = -E[\partial \mathbf{u}/\partial \boldsymbol{\theta}^T]$ and $\mathbf{B} = E[\mathbf{u}\mathbf{u}^T]$. These can be estimated by

$$\hat{\text{var}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-T} \tag{3.4}$$

where $\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n} -\partial \mathbf{u}/\partial \boldsymbol{\theta}^T$ and $\hat{B}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}\mathbf{u}^T$.

Using the notation of Williamson et al., this derivation results in the following large-sample variance estimator.

$$\hat{\text{var}}(\hat{\tau}_{ATO}) = K_{1j}^2\hat{\text{var}}(\hat{\mu}_1) + K_{0j}^2\hat{\text{var}}(\hat{\mu}_0) - 2K_{0j}K_{1j}\hat{\text{var}}(\hat{\mu}_1, \hat{\mu}_0) \tag{3.5}$$

where $\hat{K}_{01} = 1$ and $\hat{K}_{11} = 1$ are used if $\hat{\tau}_{ATO}$ is calculated using a generalized linear model with the identity link, $\hat{K}_{02} = \hat{\mu}_0^{-1}$ and $\hat{K}_{12} = \hat{\mu}_1^{-1}$ for a log link, and $\hat{K}_{03} = (\hat{\mu}_0(1 - \hat{\mu}_0))^{-1}$ and $\hat{K}_{04} = (\hat{\mu}_1(1 - \hat{\mu}_1))^{-1}$ for a logit link.

The full derivation of $\mathbf{A}$ and $\mathbf{B}$ as well as the final variance estimator can be found in Appendix A1.

3.2.2 ATO doubly robust estimator

The doubly robust estimator for the ATM is derived by Li and Greene (Li and Greene 2013). We extend the derivation to demonstrate that the same holds when

implementing the ATO weights. The doubly robust estimator for the ATO weights, denoted is $\hat{\Delta}_{DR,ATO}$, is as follows,

$$\begin{aligned}\hat{\Delta}_{DR,ATO} =& \frac{\sum_{i=1}^{n}((1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i))(m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1) - m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} + \\ & \frac{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i(Y_i - m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1))}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} - \\ & \frac{\sum_{i=1}^{n} e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)(Y_i - m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^{n} e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)}\end{aligned} \tag{3.6}$$

where $e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})$ is the estimated propensity score, often obtained from a logistic regression, $m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)$ is the predicted value of the outcome obtained from the outcome model fit among those who received the treatment ($Z_i = 1$), and $m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)$ is the predicted value of the outcome obtained from the outcome model fit among those who did not receive the treatment ($Z_i = 0$). The outcome models can be fit with a generalized linear model with the identity, log, or logit link as appropriate.

Theorem 2.1.1

When the outcome model is correctly specified, that is $m_1(\mathbf{X},\boldsymbol{\alpha}_1)$ and $m_0(\mathbf{X},\boldsymbol{\alpha}_0)$ are correctly specified, $\hat{\Delta}_{DR,ATO}$ (Equation (3.6)) will yield an unbiased estimator for the ATO effect (Δ).

The first term of $\hat{\Delta}_{DR,ATO}$ is consistent for

$$\frac{E[((1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-Z))(m_1(\mathbf{X},\boldsymbol{\alpha}_1) - m_0(\mathbf{X},\boldsymbol{\alpha}_0))]}{E[((1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} \tag{3.7}$$

Therefore, the first term of $\hat{\Delta}_{DR,ATO}$ converges to Δ.

The second and third terms of $\hat{\Delta}_{DR,ATO}$ converge to $E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y - E[Y|Z = 1,\mathbf{X}])]$ and $E[(e(\mathbf{X},\boldsymbol{\beta}))(1-Z)(Y - E[Y|Z = 0,\mathbf{X}])]$ which both equal 0, as seen below.

$$E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y - E[Y|Z = 1,\mathbf{X}])] = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_1|\mathbf{X}] - E[Y_1|\mathbf{X}])] = 0 \tag{3.8}$$

The full proof follows in Appendix B1.

Theorem 2.1.2

When the propensity score model is correctly specified, that is $e(\mathbf{X}, \boldsymbol{\beta})$ is correctly specified, $\hat{\Delta}_{DR,ATO}$ will yield an unbiased estimator for the ATO effect (Δ).

We can rewrite $\hat{\Delta}_{DR,ATO}$ as follows.

$$\begin{aligned}\hat{\Delta}_{DR,ATO} = &\left\{ \frac{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_iY_i}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} - \frac{\sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)Y_i}{\sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right\} \\ &+ \left\{ \frac{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1) + \sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + \sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} - \right. \\ &\left. \frac{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} \right\} \\ &+ \left\{ \frac{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0) + \sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)}{\sum_{i=1}^{n}(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + \sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} - \right. \\ &\left. \frac{\sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)}{\sum_{i=1}^{n}e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right\}\end{aligned} \tag{3.9}$$

Since we know the propensity score model is correctly specified, the first term is an unbiased estimator for Δ (as it is equivalent to $\hat{\tau}_{ATO}$, Equation (3.1), laid out in the section above), therefore, we just need to show that the second and third term are 0.

$e(\mathbf{X}, \boldsymbol{\beta}) = e(\mathbf{X}) = E[Z|\mathbf{X}] = E[Z|Y_1, \mathbf{X}]$ by no unmeasured confounders.

The $E[(1 - e(\mathbf{X}, \boldsymbol{\beta}))Zm_1(\mathbf{X}, \boldsymbol{\alpha}_1)] = E[(e(\mathbf{X}, \boldsymbol{\beta}))(1 - Z)m_1(\mathbf{X}, \boldsymbol{\alpha}_1)]$ and $E[(1 - e(\mathbf{X}, \boldsymbol{\beta}))Z] = E[e(\mathbf{X}, \boldsymbol{\beta})(1 - Z)]$, therefore the second term converges to Equation (3.10).

$$\begin{aligned}&\frac{E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)] + E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X}))Z] + E[(1-e(\mathbf{X}))Z]} \\ &- \frac{E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X}))Z]} = 0\end{aligned} \tag{3.10}$$

Similarly, the third term, $E[(1 - e(\mathbf{X}))Zm_0(\mathbf{X}, \boldsymbol{\alpha}_0)] = E[(e(\mathbf{X}))(1 - Z)m_0(\mathbf{X}, \boldsymbol{\alpha}_0)]$,

converges to Equation (3.11).

$$\begin{aligned}&\frac{E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)]+E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)]}{E[(1-e(\mathbf{X}))Z]+E[(1-e(\mathbf{X}))Z]}\\&-\frac{E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)]}{E[(1-e(\mathbf{X}))Z]}=0\end{aligned} \tag{3.11}$$

Therefore, $\hat{\Delta}_{ATO,DR} \to_p \Delta$. The full proof follows in Appendix B2.

3.2.3 ATO doubly robust large-sample variance estimator

The large-sample variance estimator (a sandwich estimator) for the ATE doubly robust estimator was originally proposed by Lunceford and Davidian (Lunceford and Davidian 2004). Li and Greene derive a large-sample variance estimator for the ATM doubly robust estimator following the same M-estimation process (Li and Greene 2013; Mao and Li 2018). We extend this to the ATO doubly robust estimator.

The doubly robust ATO estimator derived in Equation (3.6), $\hat{\Delta}_{DR,ATO}$, can be written as $\hat{\delta}_1 + \hat{\delta}_2 - \hat{\delta}_3$, where

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n((1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i))(m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1) - m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^n(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \tag{3.12}$$

$$\hat{\delta}_2 = \frac{\sum_{i=1}^n(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i(Y_i - m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1))}{\sum_{i=1}^n(1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} \tag{3.13}$$

$$\hat{\delta}_3 = \frac{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)(Y_i - m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \tag{3.14}$$

Using these quantities, we can solve the following estimating equations $\sum_{i=1}^n \mathbf{u}(\boldsymbol{\theta}(Y_i, Z_i, X_i, V_i)) = 0$ for $\boldsymbol{\theta} = (\delta_1, \delta_2, \delta_3, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_0^T, \boldsymbol{\beta}^T)^T$, similar to Equation (3.2) where,

$$\mathbf{u}(\boldsymbol{\theta}; Y, Z, \mathbf{X}, \mathbf{X}) = \begin{pmatrix} (m_1(\mathbf{X}, \boldsymbol{\alpha}_1) - m_0(\mathbf{X}, \boldsymbol{\alpha}_0) - \delta_1)(Z(1 - e(\mathbf{X}, \boldsymbol{\beta})) + (1 - Z)e(\mathbf{X}, \boldsymbol{\beta})) \\ (Y - m_1(\mathbf{X}, \boldsymbol{\alpha}_1) - \delta_2)Z(1 - e(\mathbf{X}, \boldsymbol{\beta})) \\ (Y - m_0(\mathbf{X}, \boldsymbol{\alpha}_0) - \delta_3)(1 - Z)e(\mathbf{X}, \boldsymbol{\beta}) \\ (Y - m_1(\mathbf{X}, \boldsymbol{\alpha}_1))Z\mathbf{X} \\ (Y - m_0(\mathbf{X}, \boldsymbol{\alpha}_0))(1 - Z)\mathbf{X} \\ \mathbf{X}(Z - e(\mathbf{X}, \boldsymbol{\beta})) \end{pmatrix} \tag{3.15}$$

Using the same method detailed above, we can solve for the large-sample variance using the delta method.

The variance of our estimator, $\hat{\Delta}_{DR,ATO}$ will be the variance of $\hat{\delta}_1 + \hat{\delta}_2 - \hat{\delta}_3$, which can be estimated by Equation (3.16).

$$\hat{\text{var}}(\hat{\Delta}_{DR,ATO}) = (1, 1, -1, \mathbf{0}, \mathbf{0}, \mathbf{0})\frac{1}{n}\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-T}(1, 1, -1, \mathbf{0}, \mathbf{0}, \mathbf{0})^T \tag{3.16}$$

Full derivations of $\mathbf{A}$ and $\mathbf{B}$ are included in Appendix A2. R code to calculate these quantities is included in Appendix C.

3.2.4 Simulations

Simulations are conducted using R version 3.4.3 (R Core Team 2017).

We conducted simulations using both a continuous and binary outcome, based on the simulations in Chapter 2, adapted from Freedman and Berk (Freedman and Berk 2008). The parameters are slightly modified. In both cases, the propensity score model is defined as a probit selection model with normally distributed errors, $V \sim N(0, 1)$.

$$Z = \begin{cases} 1 & 0.5 + 0.25X_1 + 0.75X_2 + V > 0 \\ 0 & o.w. \end{cases} \tag{3.17}$$

$\mathbf{X}$ is bivariate normal, defined as $\mathbf{X} \sim MVN\left(\begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$.

For the continuous case, the outcome model is the following with normally distributed errors, $U \sim N(0, 1)$, defined as

$$Y = 1 + Z + X_1 + 3X_2 + U \tag{3.18}$$

For the binary case, the outcome model has errors with a standard logistic distribution, $U \sim logis(0, 1)$. The model is specified as,

$$Y = \begin{cases} 1 & 1 + Z + X_1 + 3X_2 + U > 0 \\ 0 & o.w. \end{cases} \tag{3.19}$$

For each outcome, we fit three models. We fit a naive model, defined as a weighted generalized linear model, a naive model with robust standard errors, defined as a weighted generalized linear model with variances estimated using the sandwich estimator (not including the propensity score estimation), and a doubly robust model with our derived large-sample variance estimator. For the continuous outcome, the models were fit with the identity link; the binary models were fit with a log link. The point estimates for the naive model and the naive model with robust standard errors are calculated by maximizing the sum of weighted likelihood function, with respect to $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$. This can be done using the following score function, the derivative of the log likelihood function, $\mathbf{u}$.

$$\mathbf{u}(\boldsymbol{\theta}) = w_i \frac{\{y_i - b'(\boldsymbol{\theta}_i)\}}{\boldsymbol{\varphi}} \tag{3.20}$$

where for the models fit with the identity link, $\boldsymbol{\varphi} = \boldsymbol{\Sigma}$, $\boldsymbol{\theta} = \mathbf{X}_i\boldsymbol{\alpha}$, and $b'(\boldsymbol{\theta}) = \boldsymbol{\theta}$, and for the models fit with the log link, $\varphi = 1$, $\boldsymbol{\theta} = \log(\mathbf{X}_i\boldsymbol{\alpha})$, and $b'(\boldsymbol{\theta}) = \exp(\boldsymbol{\theta})$. Here, w_i is our ATO weight, $Z_i(1 - e(\mathbf{X}_i, \boldsymbol{\beta})) + (1 - Z_i)e(\mathbf{X}_i, \boldsymbol{\beta})$.

These can be estimated using the `glm` function in R along with the `weights` parameter (Venables and Ripley 2002).

The variance for the naive model is estimated using $b''(\boldsymbol{\theta})\boldsymbol{\varphi}/w_i$, whereas the variance for the naive model with robust standard errors in calculated using the sandwich estimator, that is $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-T}$, where $\mathbf{A}$ is estimated using the observed model variance, here $b''(\boldsymbol{\theta})\boldsymbol{\varphi}/w_i$ and $\mathbf{B}$ is the expectation of the squared estimating equation $\mathbf{u}$ (Equation (3.20)). This can be estimated in R using the `sandwich` function (Zeileis 2004; Zeileis 2006).

We also add a doubly robust estimator fit with the logit link to demonstrate how this

fit compares to the doubly robust estimator with a log link. We then fit each of these under four model conditions: "correct-correct", where both the propensity score model and the outcome model are correctly specified, "correct-wrong" where the propensity score model is correctly specified, but the outcome model is missing the covariate X_2, "wrong-correct" where the propensity score model is misspecified, missing the covariate X_2 and the outcome model is correctly specified, and "wrong-wrong" where X_2 is missing from both the propensity score and the outcome model. We fit two additional "wrong-wrong" models, reducing the "true" coefficient for X_2 to 1 and then 0.25 to demonstrate how this model performs when a less severe confounder is missed. We also examine each simulation under three sample sizes, 200, 1000, and 5000, each carried out 1000 times.

For each simulation, we report the bias, the ratio of the estimated standard error and the "true" Monte Carlo standard error, the root mean square error (RMSE) of the standard error, and the 95% coverage.

3.3 Results

Figures 3.1-3.6 display the ratio of the estimated standard error and the "true" standard error for the continuous outcome for each of the model states. Figures 3.7-3.12 display the ratio of the estimated standard error and the "true" standard error for the binary outcome for each of the model states. In addition we report the bias, the ratio of the estimated standard error and the "true" Monte Carlo standard error, the root mean square error (RMSE) of the standard error, and the 95% coverage. The bias for the continuous model is calculated based on our "true" effect of 1, as set up in Equation (3.18); the bias for the binary model is calculated based on the "true" risk difference in the ATO population, 0.105, as determined by Equation (3.17) and Equation (3.19). Tables 3.1 - 3.3 display the results for the continuous outcome, simulated via Equation (3.18) and Tables 3.4-3.6 display the results for the binary outcome, simulated via Equation (3.19).

The model "c-c" represents the correct-correct simulation, where both the propensity score model and the outcome model are correctly specified. The model "c-w" represents the correct-wrong simulation, where the propensity score model is correctly specified, but the outcome model is incorrectly specified via missing the covariate X_2. The model "w-c" represents the wrong-correct simulation, where the propensity score model is incorrectly specified, missing the covariate X_2, and the outcome model is correctly

specified. The model "w-w (a = 0.25)" represents the wrong-wrong simulation, where both models are missing X_2 and the coefficient for X_2 in the outcome model is 0.25. The model "w-w (a = 1)" is the wrong-wrong simulation, where both models are missing X_2 and the coefficient for X_2 in the outcome model is 1. The model "w-w (a = 3)" is the wrong-wrong simulation, where both models are missing X_2 and the coefficient for X_2 in the outcome model is 3. All simulations were carried out 1000 times.

3.3.1 Continuous outcome

In the continuous outcome case, the doubly robust estimator slightly outperforms the naive model in terms of bias in the small sample (n = 200), however the bias is the same in the larger sample sizes, regardless of the model specification (Tables 3.1-3.3). The variance estimation and coverage appear slightly improved using the large-sample variance for the doubly robust estimator for correct-correct and correct-wrong models. The naive estimator with the robust variance performs similarly to the large-sample variance for the doubly robust estimator in the correct-correct (Figure 3.1) and wrong-correct case (Figure 3.3), where the propensity score model is misspecified, but the outcome model is correctly specified, however in the correct-wrong case, the naive estimator with the robust variance is notable conservative (Figure 3.2). In the case of the three wrong-wrong models (Figures 3.4 - 3.6), it appears that the large-sample for the doubly robust estimator and the naive estimator with the robust variance perform well in terms of the "true" variance, however eventually, regardless of the size of the unmeasured confounder, although the variance estimate may be close, the bias will over-rule, as seen in the coverage (Tables 3.1 - 3.3), particularly as n increases. The naive variance underestimates the variance in all cases except when the propensity score model is correctly specified and the outcome model is incorrectly specified, in which case it overestimates the variance.

Table 3.1: Monte Carlo results for the simulation of the continuous outcome, n = 200

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR (robust variance*)	-0.006	0.954	0.031	0.93
c-c	naive	-0.007	0.594	0.011	0.74
c-c	naive (robust variance)	-0.007	0.938	0.030	0.93

model	method	bias	se / true se	RMSE(se)	coverage
c-w	DR (robust variance*)	0.001	0.932	0.030	0.93
c-w	naive	0.003	1.217	0.026	0.98
c-w	naive (robust variance)	0.003	1.927	0.060	1.00
w-c	DR (robust variance*)	0.009	0.952	0.032	0.94
w-c	naive	0.009	0.654	0.013	0.80
w-c	naive (robust variance)	0.009	0.953	0.033	0.94
w-w (a=0.25)	DR (robust variance*)	-0.151	0.968	0.027	0.88
w-w (a=0.25)	naive	-0.151	0.649	0.011	0.70
w-w (a=0.25)	naive (robust variance)	-0.151	0.960	0.026	0.88
w-w (a=1)	DR (robust variance*)	-0.594	0.979	0.032	0.34
w-w (a=1)	naive	-0.594	0.657	0.013	0.17
w-w (a=1)	naive (robust variance)	-0.594	0.971	0.031	0.34
w-w (a=3)	DR (robust variance*)	-1.768	0.976	0.056	0.04
w-w (a=3)	naive	-1.769	0.659	0.023	0.01
w-w (a=3)	naive (robust variance)	-1.769	0.968	0.056	0.04

Table 3.2: Monte Carlo results for the continuous outcome, n = 1000

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR (robust variance*)	0.005	0.976	0.006	0.94
c-c	naive	0.005	0.614	0.002	0.76
c-c	naive (robust variance)	0.005	0.973	0.006	0.94

model	method	bias	se / true se	RMSE(se)	coverage
c-w	DR (robust variance*)	-0.002	0.952	0.006	0.94
c-w	naive	-0.002	1.249	0.005	0.98
c-w	naive (robust variance)	-0.002	1.986	0.012	1.00
w-c	DR (robust variance*)	0.002	0.980	0.007	0.95
w-c	naive	0.002	0.666	0.003	0.81
w-c	naive (robust variance)	0.002	0.988	0.007	0.94
w-w (a=0.25)	DR (robust variance*)	-0.146	1.021	0.005	0.67
w-w (a=0.25)	naive	-0.146	0.684	0.002	0.42
w-w (a=0.25)	naive (robust variance)	-0.146	1.019	0.005	0.67
w-w (a=1)	DR (robust variance*)	-0.582	1.040	0.006	0.00
w-w (a=1)	naive	-0.582	0.699	0.002	0.00
w-w (a=1)	naive (robust variance)	-0.582	1.038	0.006	0.00
w-w (a=3)	DR (robust variance*)	-1.763	0.990	0.011	0.00
w-w (a=3)	naive	-1.763	0.669	0.005	0.00
w-w (a=3)	naive (robust variance)	-1.763	0.987	0.011	0.00

Table 3.3: Monte Carlo results for the continuous outcome, n = 5000

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR (robust variance*)	0.001	1.005	0.001	0.95
c-c	naive	0.001	0.632	0.000	0.79
c-c	naive (robust variance)	0.001	1.004	0.001	0.95

model	method	bias	se / true se	RMSE(se)	coverage
c-w	DR (robust variance*)	-0.002	0.939	0.001	0.94
c-w	naive	-0.002	1.235	0.001	0.98
c-w	naive (robust variance)	-0.002	1.961	0.003	1.00
w-c	DR (robust variance*)	-0.001	0.978	0.001	0.94
w-c	naive	-0.001	0.653	0.001	0.79
w-c	naive (robust variance)	-0.001	0.973	0.001	0.94
w-w (a=0.25)	DR (robust variance*)	-0.148	0.993	0.001	0.07
w-w (a=0.25)	naive	-0.148	0.665	0.000	0.02
w-w (a=0.25)	naive (robust variance)	-0.148	0.993	0.001	0.07
w-w (a=1)	DR (robust variance*)	-0.588	1.030	0.001	0.00
w-w (a=1)	naive	-0.588	0.692	0.000	0.00
w-w (a=1)	naive (robust variance)	-0.588	1.029	0.001	0.00
w-w (a=3)	DR (robust variance*)	-1.761	1.027	0.002	0.00
w-w (a=3)	naive	-1.761	0.694	0.001	0.00
w-w (a=3)	naive (robust variance)	-1.761	1.026	0.002	0.00

3.3.2 Binary outcome

Generally, we expect the naive variance estimator to underestimate the variance, and the naive sandwich estimator to be more conservative. In the case of a Poisson model, however, if the model is *underdispersed*, as is the case in our simulations, the opposite will be true, the naive variance estimator will overestimate the variance. We see this here, since our simulated outcome has a variance smaller than its mean, whereas the

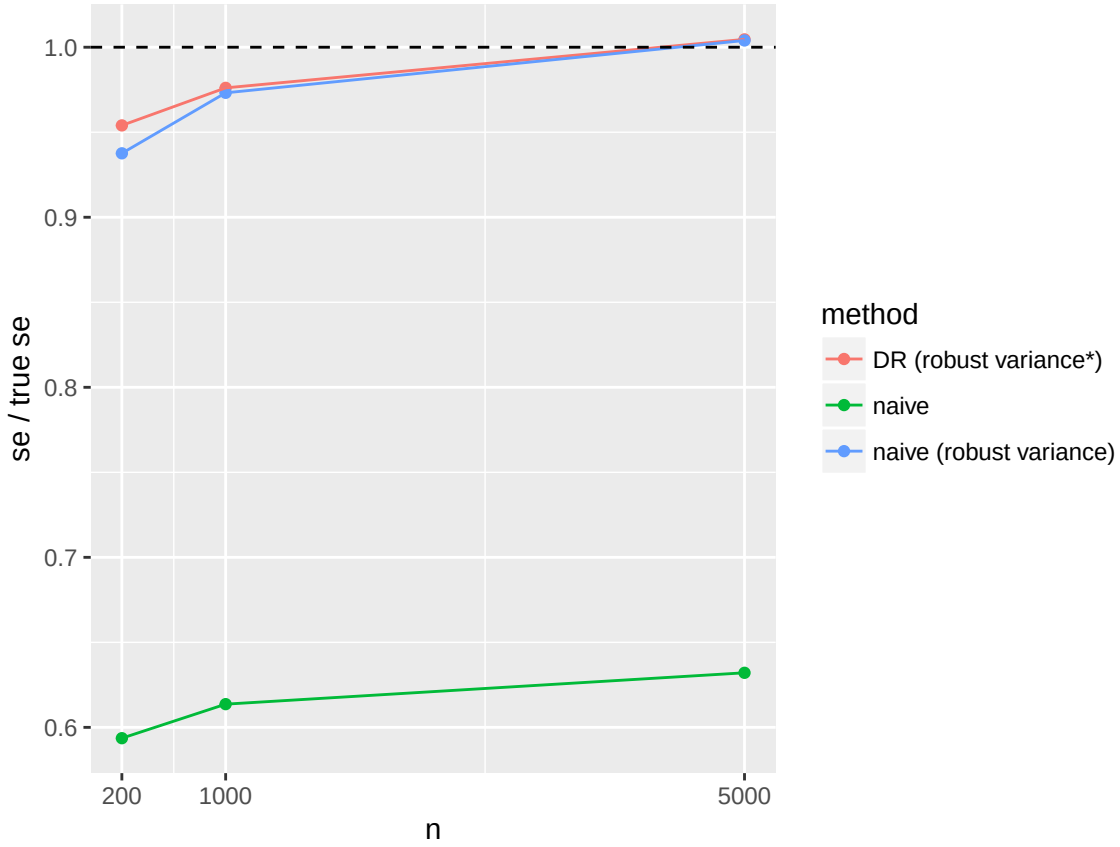

Figure 3.1: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are correctly specified at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

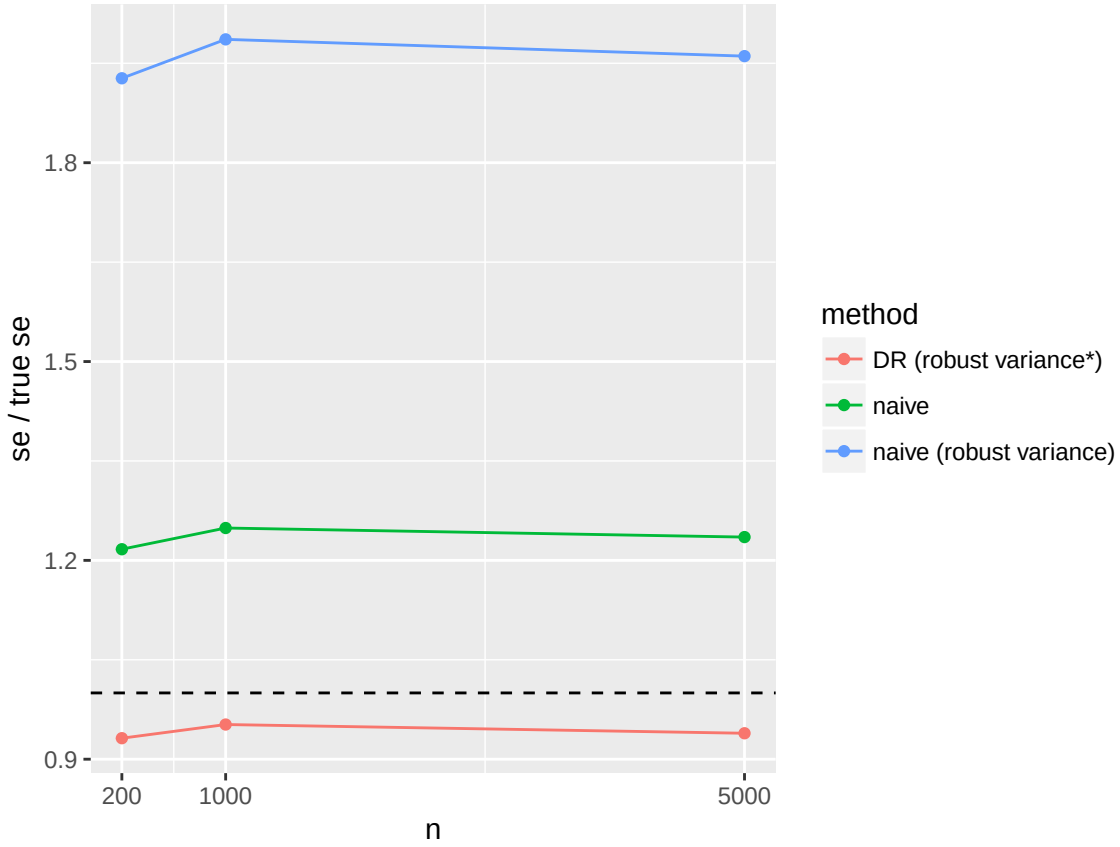

Figure 3.2: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where the propensity score model is correctly specified and the outcome model is incorrectly specified at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

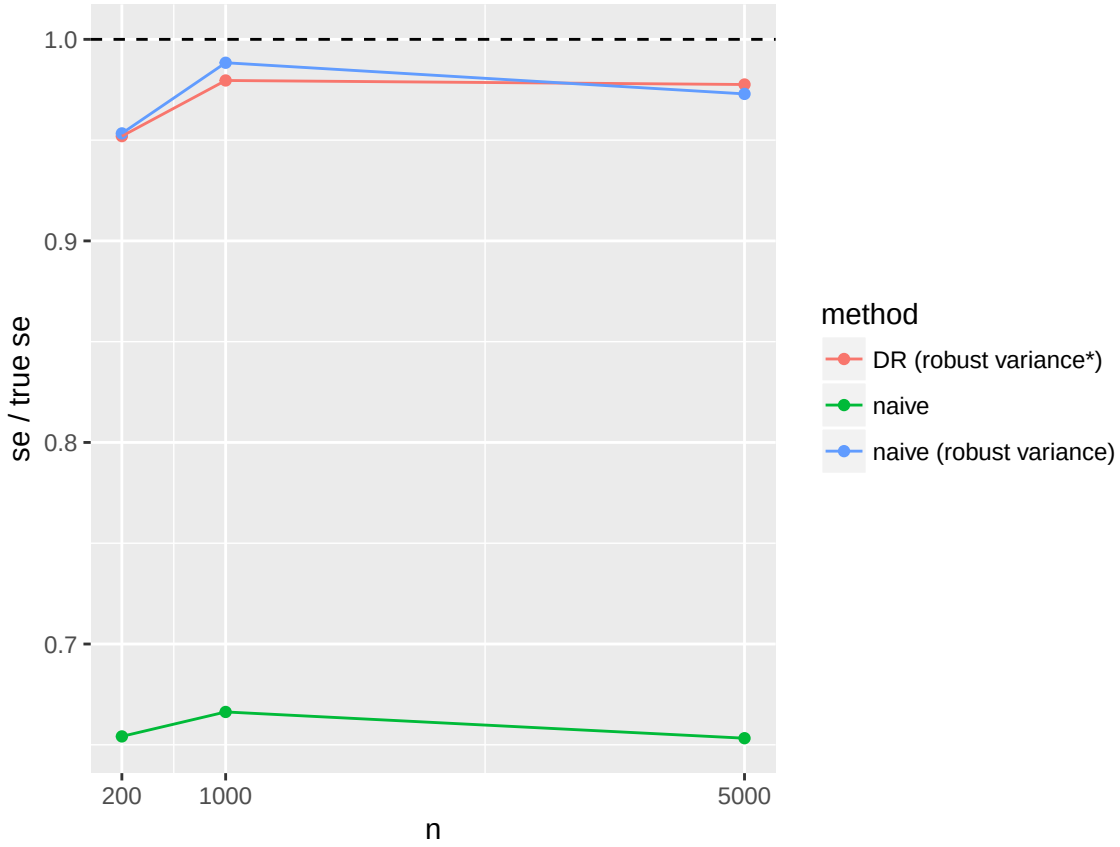

Figure 3.3: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where the propensity score model is incorrectly specified and the outcome model is correctly specified at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

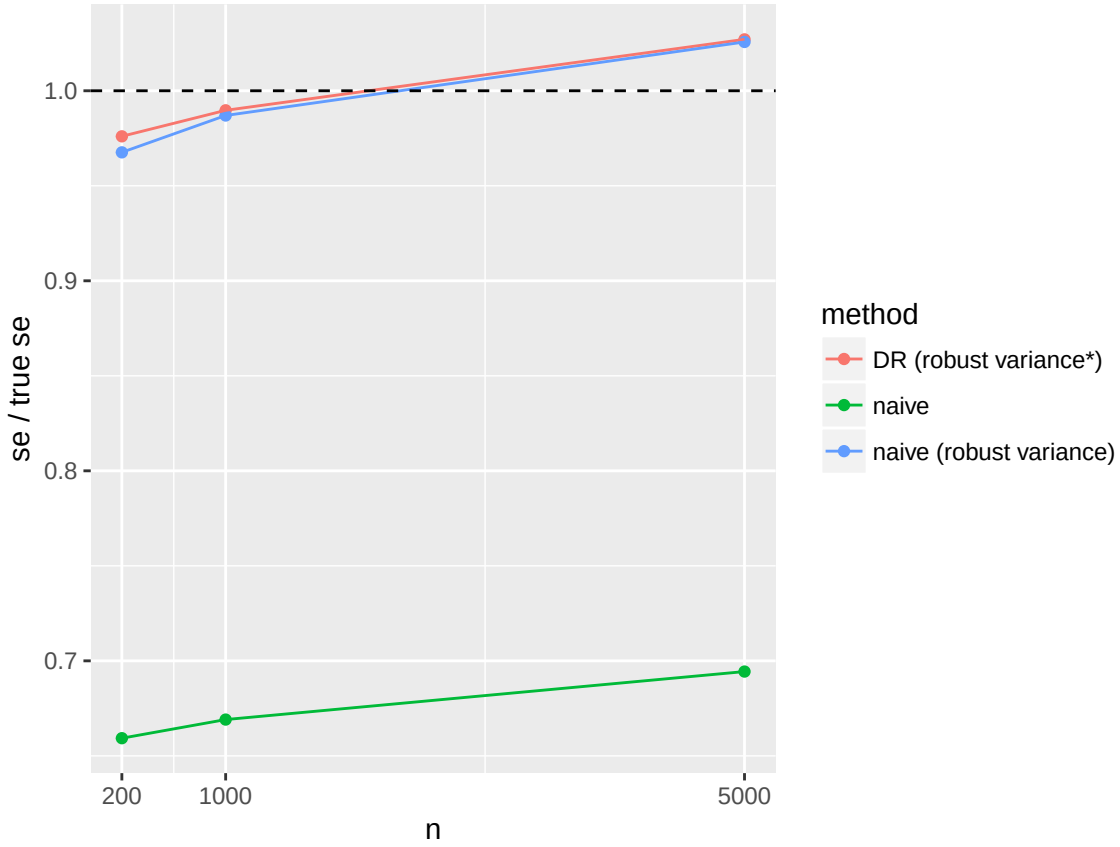

Figure 3.4: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 3, at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

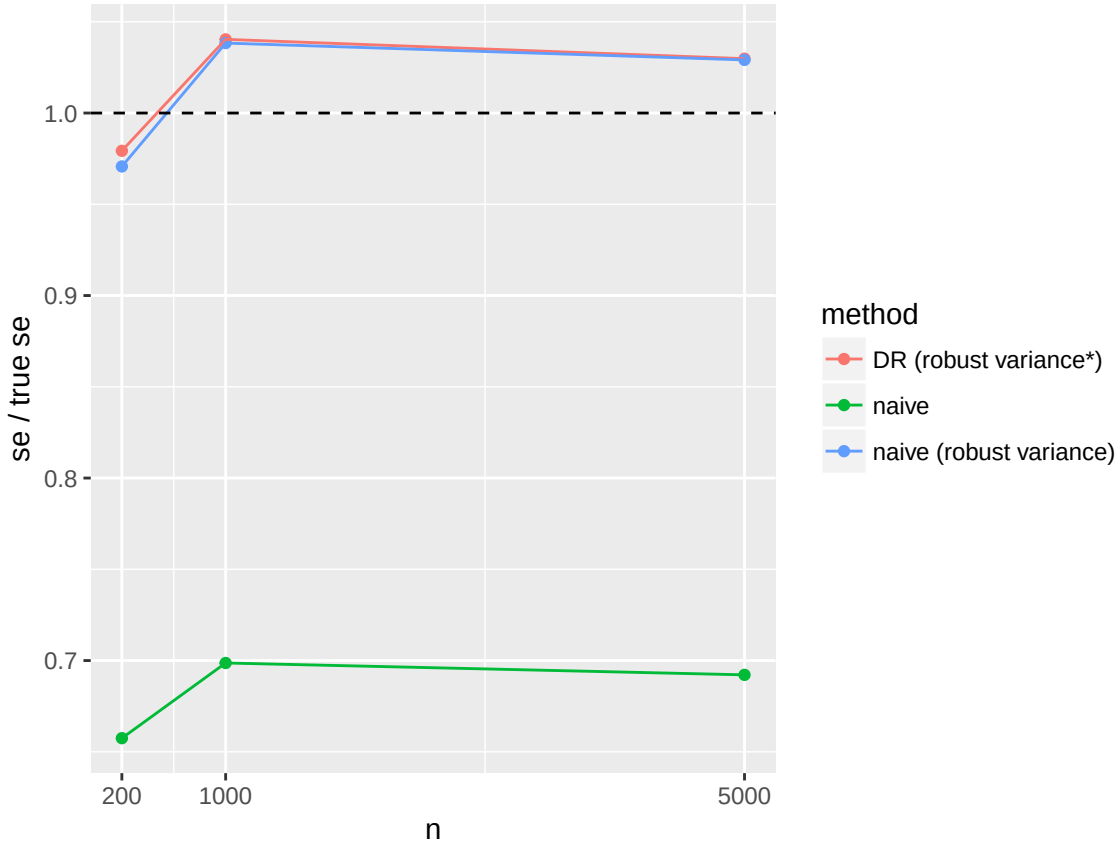

Figure 3.5: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 1, at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

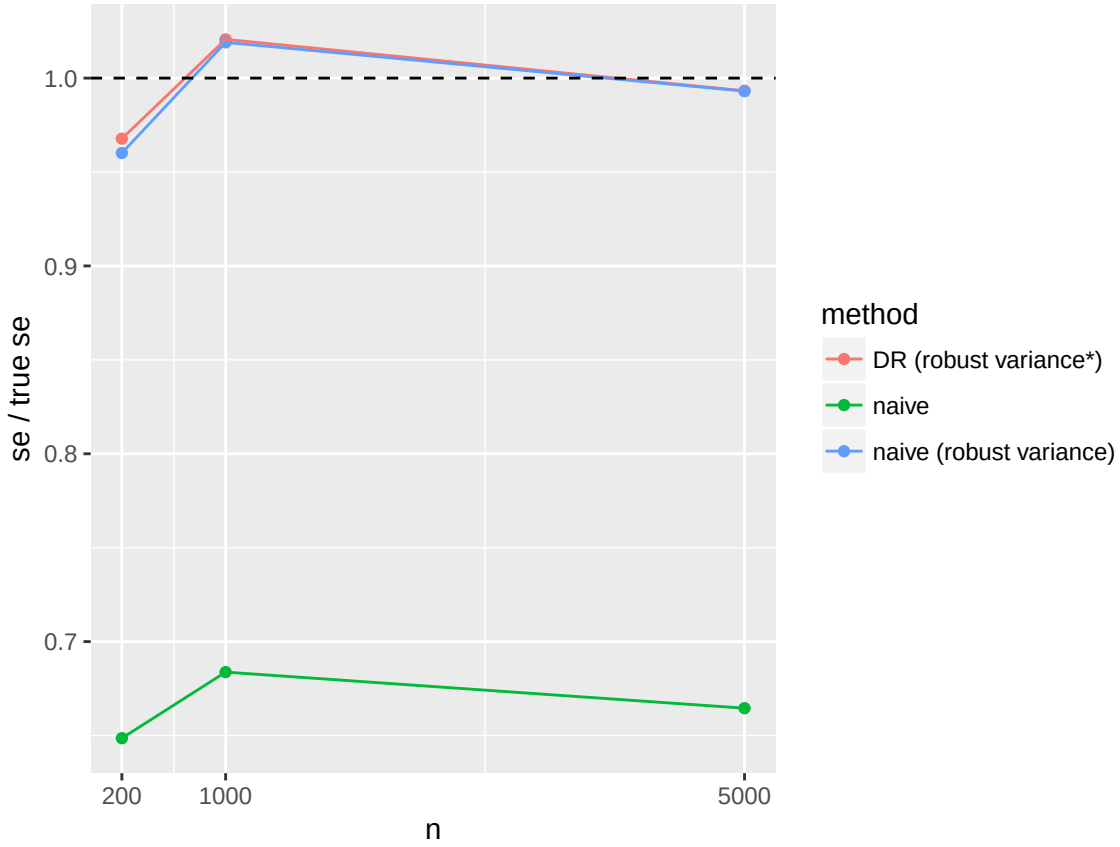

Figure 3.6: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 0.25, at n = 200, 1000, and 5000 for the continuous outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator, the green line indicates the ratio for the standard errors obtained from the naive model, and the blue line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

Poisson model assumes they are equal.

In the binary case, the bias is slightly smaller in the doubly robust estimator using the Logistic models compared to the doubly robust estimator using the Poisson models. Generally, the doubly robust estimators are less biased than the naive models, with the exception of the scenario where the propensity score model is incorrect and the outcome model is correct, in which case the doubly robust estimator using the Poisson models is slightly more biased (Tables 3.4 - 3.6). In all scenarios, the model variance for the naive model is very conservative (Figures 3.7-3.12). In the case where both the propensity score model and the outcome model are correctly specified, the large-sample variance for the doubly robust estimator for both the Poisson and Logistic models as well as the robust variance for the naive model perform similarly (Figure 3.7). In the case where the propensity score model is correctly specified and the outcome model is incorrectly specified, the large-sample variance for the doubly robust estimator for both the Poisson and Logistic models slightly underestimate the variance and the robust variance for the naive model slightly overestimates the variance (Figure 3.8). Because the doubly robust estimators for the Poisson and Logistic models are less biased in this case, however, the coverage is much better compared to that of the robust variance for the naive model (Tables 3.4 - 3.6). In the case where the propensity score model is incorrectly specified and the outcome model is correctly specified, the large-sample variance for the doubly robust estimator for both the Poisson and Logistic models as well as the robust variance for the naive model perform similarly, with the Poisson model underestimating the variance at the smaller sample size (n = 200) (Figure 3.9). Here the coverage is similar for the large-sample variance for the doubly robust estimator for the Logistic model and the robust variance for the naive model (Tables 3.4 - 3.6). Again examining the models where both the propensity score model and the outcome model are incorrectly specified, while the variance estimation appears close for the large-sample variance for the doubly robust estimator for both the Poisson and Logistic models as well as the robust variance for the naive model (Figures 3.10-3.12), the bias over-rules the variance in the coverage, regardless of the size of the unmeasured confounder (Tables 3.4 - 3.6).

model	method	bias	se / true se	RMSE(se)	coverage

Table 3.4: Monte Carlo results for the simulation of the binary outcome, n = 200

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR logistic (robust variance*)	0.003	0.932	0.014	0.92
c-c	DR poisson (robust variance*)	0.018	0.926	0.025	0.94
c-c	naive	-0.034	2.936	0.039	1.00
c-c	naive (robust variance)	-0.034	0.947	0.026	0.95
c-w	DR logistic (robust variance*)	0.006	0.959	0.011	0.94
c-w	DR poisson (robust variance*)	0.014	0.967	0.011	0.93
c-w	naive	-0.030	3.087	0.037	1.00
c-w	naive (robust variance)	-0.030	1.090	0.029	0.98
w-c	DR logistic (robust variance*)	0.009	0.893	0.013	0.90
w-c	DR poisson (robust variance*)	0.121	0.569	0.299	0.89
w-c	naive	-0.028	2.899	0.046	1.00
w-c	naive (robust variance)	-0.028	0.934	0.033	0.95
w-w (a=0.25)	DR logistic (robust variance*)	-0.083	0.944	0.010	0.86

model	method	bias	se / true se	RMSE(se)	coverage
w-w (a=0.25)	DR poisson (robust variance*)	-0.079	0.946	0.010	0.87
w-w (a=0.25)	naive	-0.169	2.325	0.033	1.00
w-w (a=0.25)	naive (robust variance)	-0.169	0.920	0.031	0.85
w-w (a=1)	DR logistic (robust variance*)	-0.091	0.942	0.010	0.82
w-w (a=1)	DR poisson (robust variance*)	-0.088	0.953	0.010	0.84
w-w (a=1)	naive	-0.178	2.419	0.033	1.00
w-w (a=1)	naive (robust variance)	-0.178	0.938	0.030	0.82
w-w (a=3)	DR logistic (robust variance*)	-0.136	0.944	0.010	0.65
w-w (a=3)	DR poisson (robust variance*)	-0.133	0.937	0.009	0.67
w-w (a=3)	naive	-0.241	2.365	0.034	1.00
w-w (a=3)	naive (robust variance)	-0.241	0.915	0.031	0.62

Table 3.5: Monte Carlo results for the binary outcome, n = 1000

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR logistic (robust variance*)	0.002	0.965	0.002	0.93
c-c	DR poisson (robust variance*)	0.003	0.958	0.002	0.93

model	method	bias	se / true se	RMSE(se)	coverage
c-c	naive	-0.039	2.936	0.007	1.00
c-c	naive (robust variance)	-0.039	0.961	0.005	0.90
c-w	DR logistic (robust variance*)	0.003	0.944	0.002	0.94
c-w	DR poisson (robust variance*)	0.004	0.938	0.002	0.94
c-w	naive	-0.038	2.997	0.007	1.00
c-w	naive (robust variance)	-0.038	1.065	0.005	0.94
w-c	DR logistic (robust variance*)	0.012	0.989	0.002	0.92
w-c	DR poisson (robust variance*)	0.049	0.927	0.007	0.83
w-c	naive	-0.020	3.031	0.008	1.00
w-c	naive (robust variance)	-0.020	0.971	0.005	0.94
w-w (a=0.25)	DR logistic (robust variance*)	-0.081	0.977	0.002	0.49
w-w (a=0.25)	DR poisson (robust variance*)	-0.080	0.982	0.002	0.50
w-w (a=0.25)	naive	-0.161	2.494	0.006	0.99
w-w (a=0.25)	naive (robust variance)	-0.161	0.977	0.006	0.26
w-w (a=1)	DR logistic (robust variance*)	-0.092	0.959	0.002	0.37

model	method	bias	se / true se	RMSE(se)	coverage
w-w (a=1)	DR poisson (robust variance*)	-0.092	0.954	0.002	0.40
w-w (a=1)	naive	-0.176	2.420	0.006	0.98
w-w (a=1)	naive (robust variance)	-0.176	0.943	0.006	0.18
w-w (a=3)	DR logistic (robust variance*)	-0.140	0.979	0.002	0.05
w-w (a=3)	DR poisson (robust variance*)	-0.139	0.981	0.002	0.06
w-w (a=3)	naive	-0.239	2.474	0.006	0.90
w-w (a=3)	naive (robust variance)	-0.239	0.959	0.006	0.02

Table 3.6: Monte Carlo results for the binary outcome, n = 5000

model	method	bias	se / true se	RMSE(se)	coverage
c-c	DR logistic (robust variance*)	0.001	0.988	0.000	0.94
c-c	DR poisson (robust variance*)	0.002	0.967	0.000	0.94
c-c	naive	-0.040	2.931	0.001	1.00
c-c	naive (robust variance)	-0.040	0.966	0.001	0.66
c-w	DR logistic (robust variance*)	0.001	0.938	0.000	0.94
c-w	DR poisson (robust variance*)	0.001	0.952	0.000	0.94

model	method	bias	se / true se	RMSE(se)	coverage
c-w	naive	-0.041	3.005	0.001	1.00
c-w	naive (robust variance)	-0.041	1.069	0.001	0.69
w-c	DR logistic (robust variance*)	0.012	1.018	0.000	0.86
w-c	DR poisson (robust variance*)	0.040	0.957	0.002	0.57
w-c	naive	-0.021	3.065	0.002	1.00
w-c	naive (robust variance)	-0.021	0.984	0.001	0.85
w-w (a=0.25)	DR logistic (robust variance*)	-0.081	1.015	0.000	0.00
w-w (a=0.25)	DR poisson (robust variance*)	-0.081	1.014	0.000	0.00
w-w (a=0.25)	naive	-0.161	2.588	0.001	0.22
w-w (a=0.25)	naive (robust variance)	-0.161	1.016	0.001	0.00
w-w (a=1)	DR logistic (robust variance*)	-0.091	0.962	0.000	0.00
w-w (a=1)	DR poisson (robust variance*)	-0.091	0.955	0.000	0.00
w-w (a=1)	naive	-0.174	2.435	0.001	0.12
w-w (a=1)	naive (robust variance)	-0.174	0.945	0.001	0.00
w-w (a=3)	DR logistic (robust variance*)	-0.138	1.034	0.000	0.00

model	method	bias	se / true se	RMSE(se)	coverage
w-w (a=3)	DR poisson (robust variance*)	-0.138	1.033	0.000	0.00
w-w (a=3)	naive	-0.236	2.633	0.001	0.00
w-w (a=3)	naive (robust variance)	-0.236	1.018	0.001	0.00

3.4 Discussion

We have derived a doubly robust estimator for the ATO estimand, as well as the large-sample variance for both the ATO estimator and the doubly robust ATO estimator. We perform a Monte Carlo simulation to compare the large-sample variance for the doubly robust estimator to two other variance estimation techniques. Although this large-sample variance estimator is only intended for the case when both the propensity score model and the outcome model are correctly specified, it appears that in our simulation settings it performs relatively well as long as at least one of the two models is correctly specified. Similarly, it seems that incorporating the propensity score estimation in the variance does generally improve the coverage properties when compared to the naive model with robust standard errors when the propensity score model is correctly specified, but the outcome model is incorrectly specified. This has some significance, as use of these sandwich estimators (or, nearly equivalently, estimating the variances using the `survey` package in R (Lumley 2011)), is commonly seen in the literature. When both models are correct, or the propensity score model is incorrectly specified but the outcome model is correctly specified, incorporating the propensity score estimation in the variance performs similarly to the naive model with robust standard errors in the continuous case, and slightly outperforms the naive model with robust standard errors in terms of coverage the binary case, due to a decrease in bias. When the models are incorrectly specified, the variance estimation does not seem negatively impacted, however the bias over-rules the variance and thus the coverage is essentially 0 as n increases. Based on these results, we would recommend using the large-sample variance for the doubly robust estimator when intending to incorporate both a propensity score and outcome model in the estimation process. In the case of model uncertainty, should the researcher find themselves in

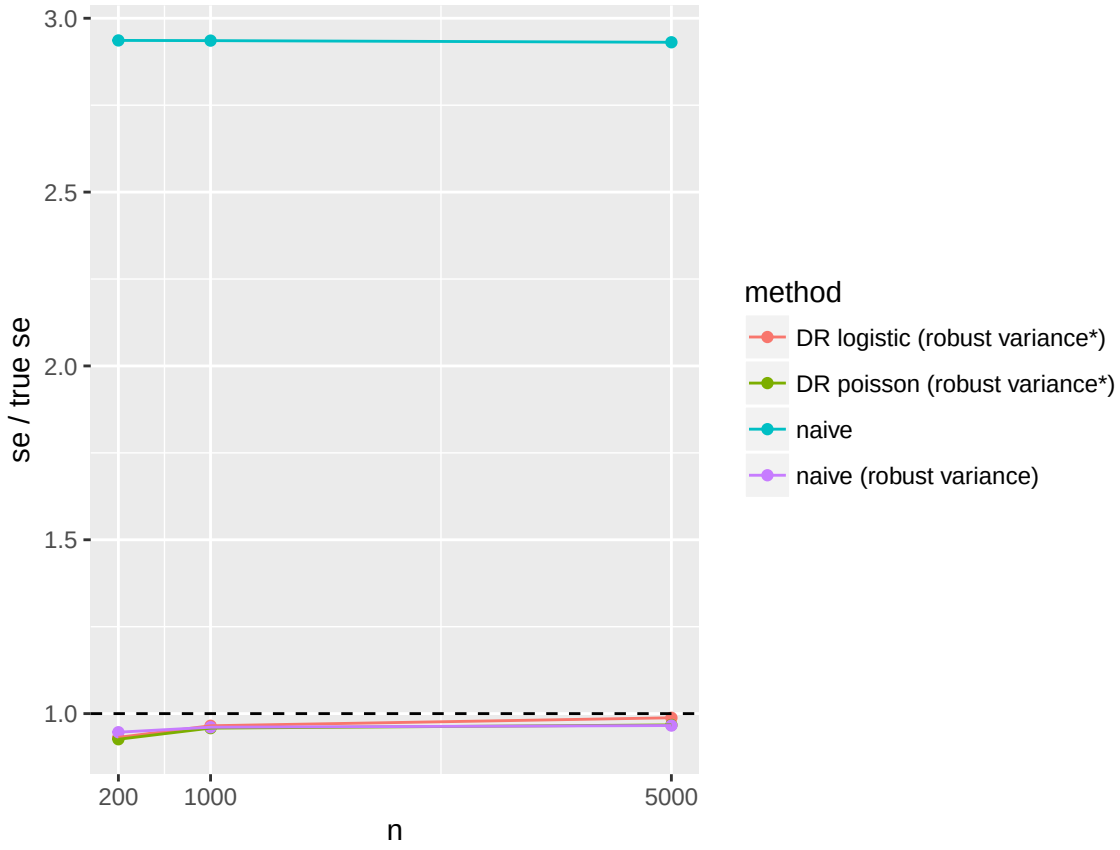

Figure 3.7: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are both correctly specified at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

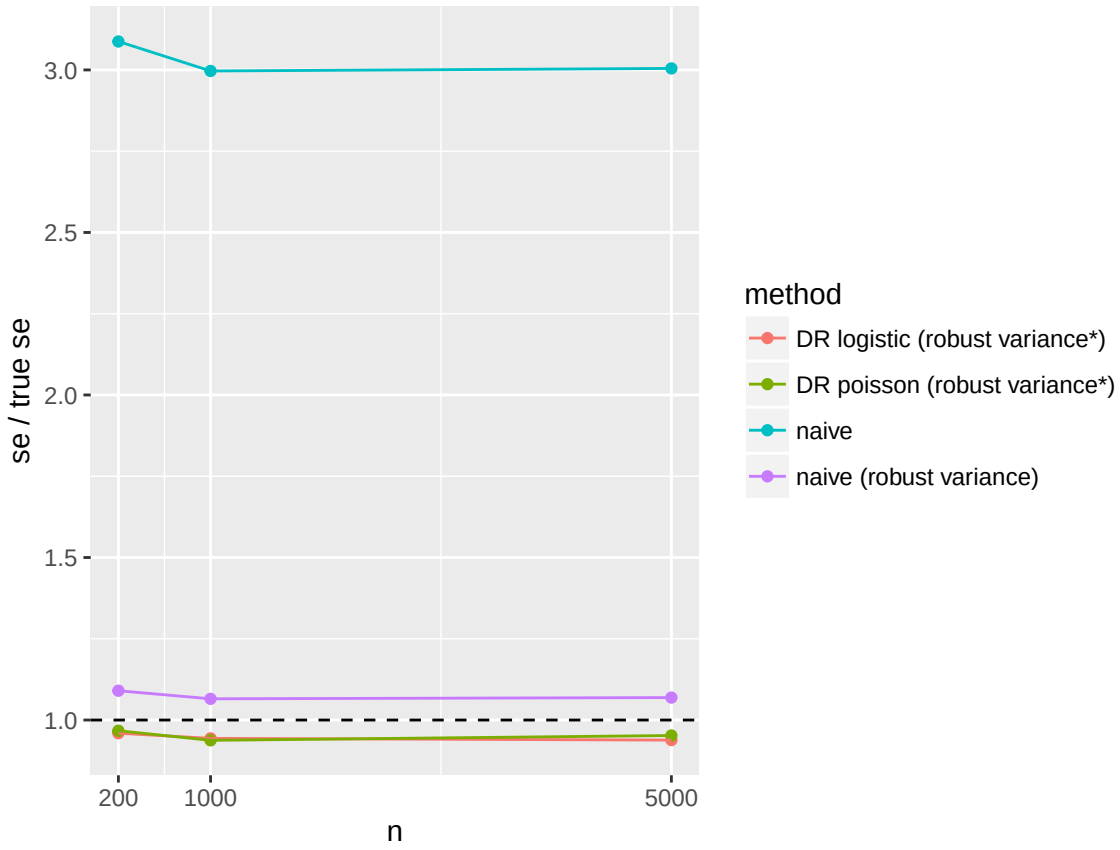

Figure 3.8: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where the propensity score model is correctly specified and the outcome model are is incorrectly specified at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

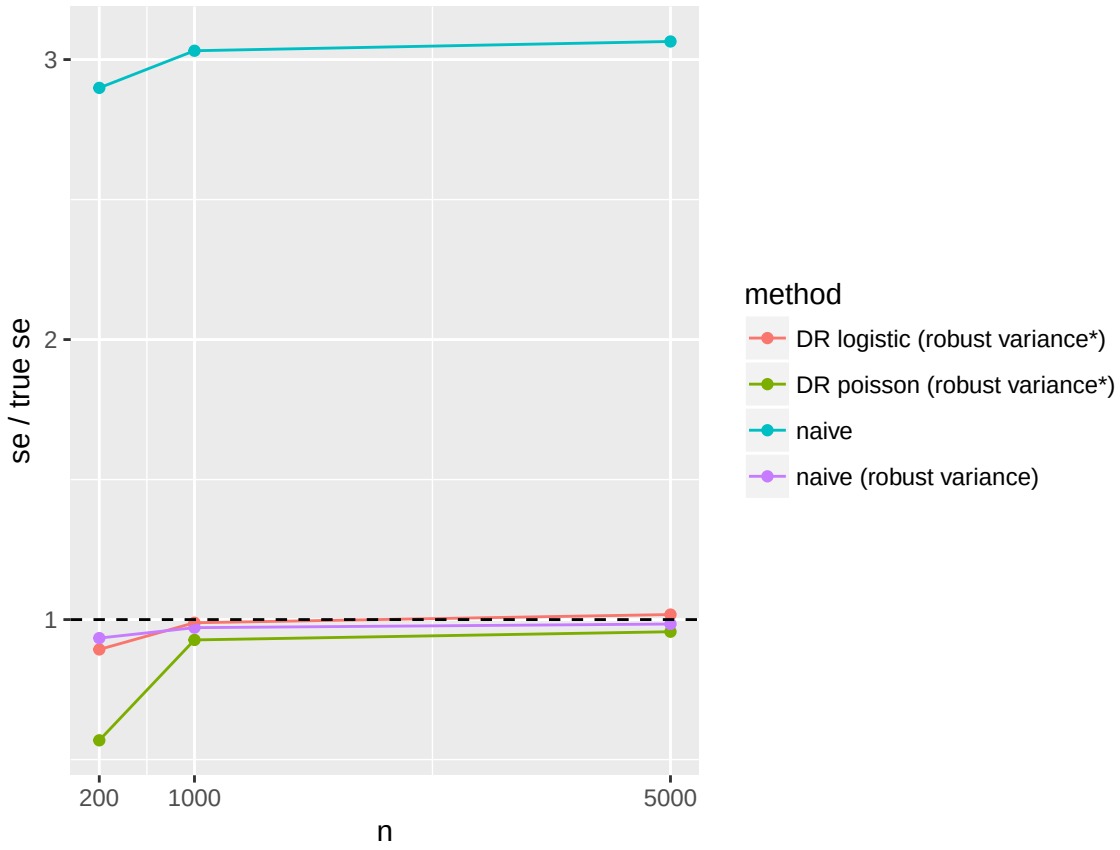

Figure 3.9: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where the propensity score model is incorrectly specified and the outcome model are is correctly specified at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

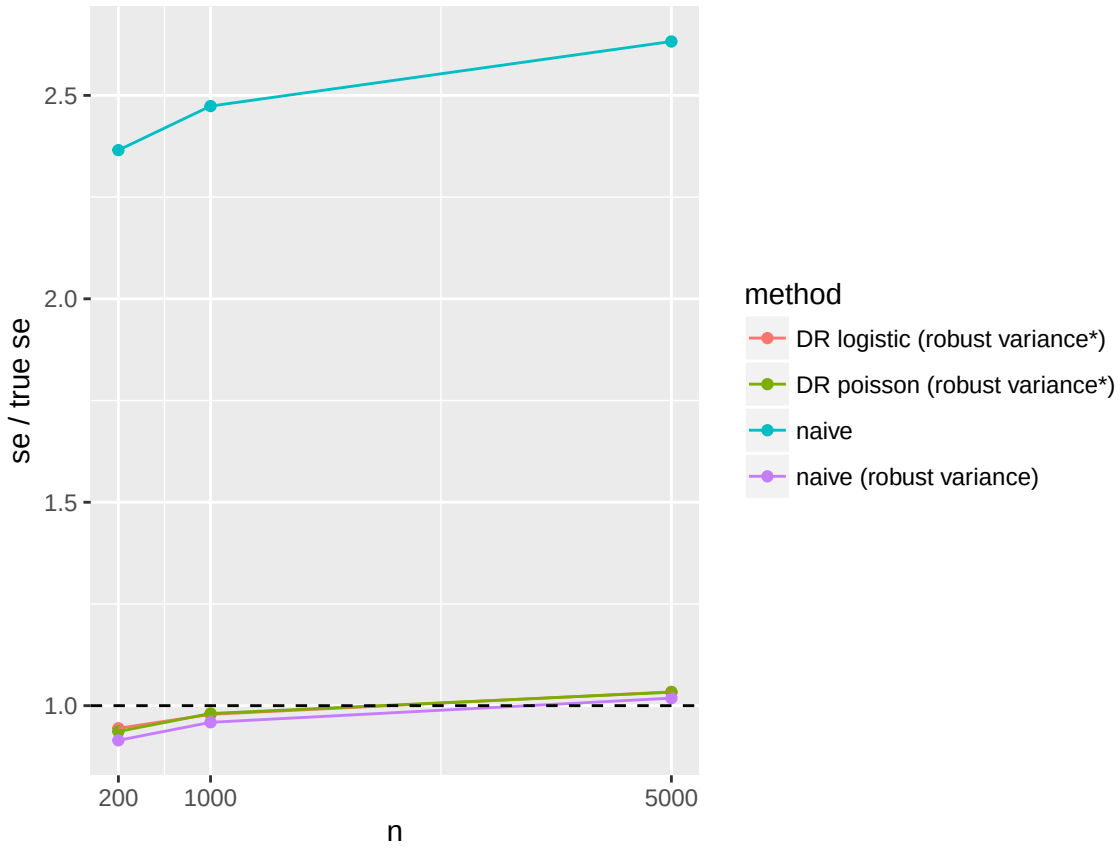

Figure 3.10: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 3, at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

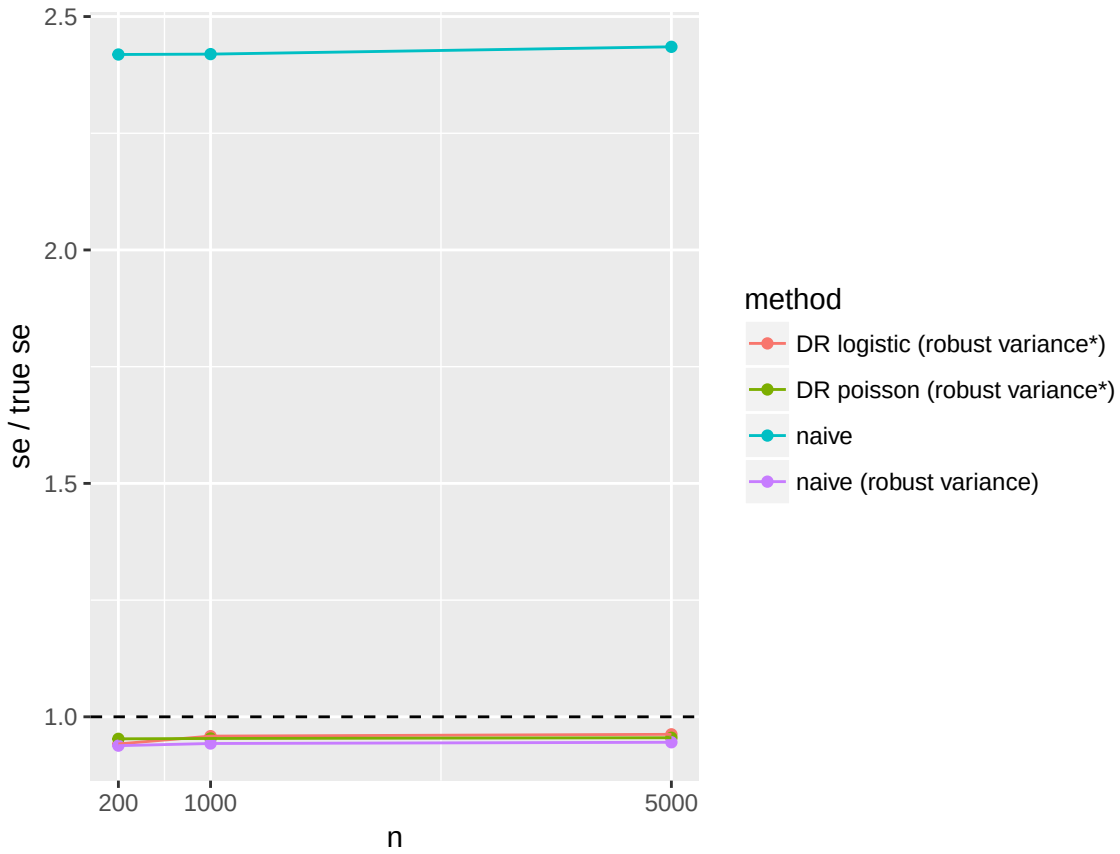

Figure 3.11: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 1, at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

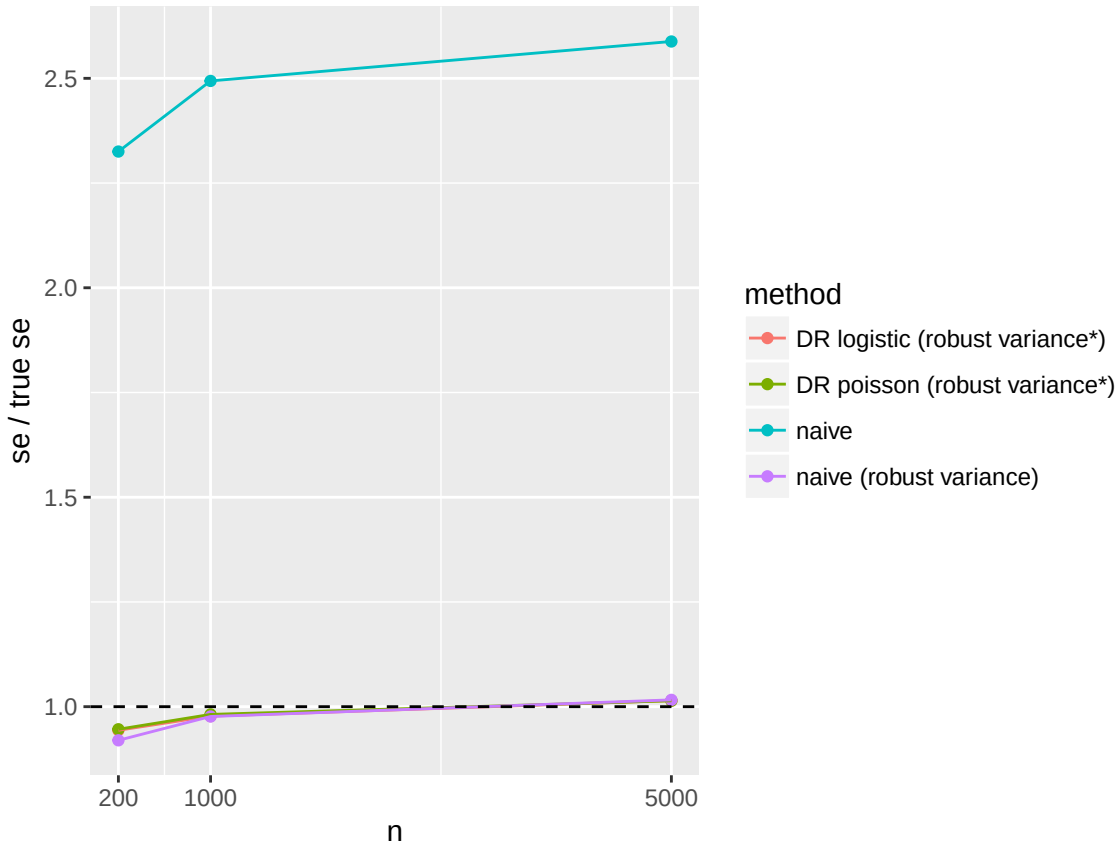

Figure 3.12: The ratio of the estimated standard error to the "true" Monte Carlo standard error for the simulation where both the propensity score model and the outcome model are incorrectly specified and the missing covariate has a coefficient of 0.25, at n = 200, 1000, and 5000 for the binary outcome. The red line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the logit link, the green line indicates the ratio for the standard error estimated using our large-sample variance for the doubly robust estimator estimated using the log link, the blue line indicates the ratio for the standard errors obtained from the naive model, and the purple line indicates the standard errors obtained from the naive model with robust standard errors (sandwich estimator) applied.

the wrong-wrong scenario, we recommend sensitivity analyses to assess the impact of a potential unmeasured confounder on the bias of the estimate of interest. This is discussed further in Chapter 4.

3.5 Appendix A1. Derivation of the large-sample variance for the ATO estimator

Using the estimating equations specified in Equation (3.2), we can solve for $\mathbf{A}$ and $\mathbf{B}$ where $\mathbf{A} = -E[\partial \mathbf{u}/\partial \boldsymbol{\theta}^T]$ and $\mathbf{B} = E[\mathbf{u}\mathbf{u}^T]$.

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$$

$$\begin{aligned}
a_{11} &= E[Z(1 - e(\mathbf{X}, \boldsymbol{\beta}))] \\
a_{13} &= E[\mathbf{X}^T(Y - \mu_1)Ze(\mathbf{X}, \boldsymbol{\beta})(1 - e(\mathbf{X}, \boldsymbol{\beta}))] \\
a_{23} &= -E[\mathbf{X}^T(Y - \mu_0)(1 - Z)e(\mathbf{X}, \boldsymbol{\beta})(1 - e(\mathbf{X}, \boldsymbol{\beta}))] \\
a_{22} &= E[(1 - Z)e(\mathbf{X}, \boldsymbol{\beta})] \\
a_{33} &= E[\mathbf{X}\mathbf{X}^T e(\mathbf{X}, \boldsymbol{\beta})(1 - e(\mathbf{X}, \boldsymbol{\beta}))]
\end{aligned}$$

$$\mathbf{B} = \begin{pmatrix} b_{11} & 0 & \\ 0 & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

$$\begin{aligned}
b_{11} &= E[(Y - \mu_1)^2 Z(1 - e(\mathbf{X}, \boldsymbol{\beta}))^2] \\
b_{13} &= E[\mathbf{X}^T(Y - \mu_1)Z(1 - e(\mathbf{X}, \boldsymbol{\beta}))^2] \\
b_{22} &= E[(Y - \mu_0)^2(1 - Z)e(\mathbf{X}, \boldsymbol{\beta})^2] \\
b_{23} &= -E[\mathbf{X}^T(Y - \mu_0)(1 - Z)e(\mathbf{X}, \boldsymbol{\beta})^2] \\
b_{31} &= E[(Y - \mu_1)Z(1 - e(\mathbf{X}, \boldsymbol{\beta}))^2\mathbf{X}] \\
b_{32} &= -E[(Y - \mu_0)(1 - Z)e(\mathbf{X}, \boldsymbol{\beta})^2\mathbf{X}] \\
b_{33} &= E[\mathbf{X}\mathbf{X}^T(Z - e(\mathbf{X}, \boldsymbol{\beta}))^2]
\end{aligned}$$

$$\text{var}(\hat{\mu_1}) = a_{11}^{-2}(b_{11} - 2a_{13}a_{33}^{-1}b_{13} + a_{13}a_{33}^{-1}(a_{13}a_{33}^{-1}b_{33}))$$

$$\text{var}(\hat{\mu_0}) = a_{22}^{-2}(b_{22} - 2a_{23}a_{33}^{-1}b_{23} + a_{23}a_{33}^{-1}(a_{23}a_{33}b_{33}))$$

$\text{cov}(\hat{\mu_1}, \hat{\mu_0}) = a_{11}^{-1} a_{22}^{-1} (a_{23} a_{33}^{-1} b_{13} + a_{13} a_{33}^{-1} (b_{23} - a_{23} a_{33}^{-1} a_{33})^T)$

We can plug in the sample estimators for these expectations to estimate the variance as follows.

$\hat{\text{var}}(\hat{\tau}_{ATO}) = K_{1j}^2 \hat{\text{var}}(\hat{mu}_1) + K_{0j}^2 \hat{\text{var}}(\hat{\mu}_0) - 2K_{0j} K_{1j} \hat{\text{var}}(\hat{\mu}_1, \hat{\mu}_0)$

where $\hat{K}_{01} = 1$ and $\hat{K}_{11} = 1$ are used for when $\hat{\tau}_{ATO}$ is estimated using a generalized linear model with the identity link, $\hat{K}_{02} = \hat{\mu}_0^{-1}$ and $\hat{K}_{12} = \hat{\mu}_1^{-1}$ for when $\hat{\tau}_{ATO}$ is fit with a log link, and $\hat{K}_{03} = (\hat{\mu}_0(1 - \hat{\mu}_0))^{-1}$ and $\hat{K}_{04} = (\hat{\mu}_1(1 - \hat{\mu}_1))^{-1}$ for $\hat{\tau}_{ATO}$ fit with a logit link.

3.6 Appendix A2. Derivation of the large-sample variance for the ATO doubly robust estimator

Using the estimating equations specified Equation (3.15), we can solve for $\mathbf{A}$ and $\mathbf{B}$ as follows.

$\mathbf{A} = -E[\partial \mathbf{u} / \partial \boldsymbol{\theta}^T]$, so we will be taking the derivative of $\mathbf{u}$ with respect to $\boldsymbol{\theta} = (\delta_1, \delta_2, \delta_3, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_0^T, \boldsymbol{\beta}^T)^T$.

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & 0 & a_{14} & a_{15} & a_{16} \\ 0 & a_{22} & 0 & a_{24} & 0 & a_{26} \\ 0 & 0 & a_{33} & 0 & a_{35} & a_{36} \\ 0 & 0 & 0 & a_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{66} \end{pmatrix}$$

$$
\begin{aligned}
a_{11} &= E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-Z)] \\
a_{14} &= -E\left[\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1}((1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-Z))\right] \\
a_{15} &= E\left[\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0}((1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-Z))\right] \\
a_{16} &= -E[\mathbf{X}(-e(\mathbf{X},\boldsymbol{\beta})(1-e(\mathbf{X},\boldsymbol{\beta}))Z + e(\mathbf{X},\boldsymbol{\beta})(1-e(\mathbf{X},\boldsymbol{\beta}))(1-Z)) \\
&\qquad (m_1(\mathbf{X},\boldsymbol{\alpha}_1) - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_1)] \\
a_{22} &= E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z] \\
a_{24} &= E\left[\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1}(1-e(\mathbf{X},\boldsymbol{\beta}))Z\right] \\
a_{26} &= E[\mathbf{X}e(\mathbf{X},\boldsymbol{\beta})(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y - m_1(\mathbf{X},\boldsymbol{\alpha}_1) - \delta_2)] \\
a_{33} &= E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)] \\
a_{35} &= E\left[\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0}e(\mathbf{X},\boldsymbol{\beta})(1-Z)\right] \\
a_{36} &= -E[\mathbf{X}e(\mathbf{X},\boldsymbol{\beta})(1-e(\mathbf{X},\boldsymbol{\beta}))(1-Z)(Y - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_3)] \\
a_{44} &= E\left[\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1}\mathbf{X}^T Z\right] \\
a_{55} &= E\left[\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0}\mathbf{X}^T(1-Z)\right] \\
a_{66} &= E[\mathbf{X}\mathbf{X}^T e(\mathbf{X},\boldsymbol{\beta})(1-e(\mathbf{X},\boldsymbol{\beta}))]
\end{aligned}
$$

where $\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1}$ and $\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0}$ are defined as follows, depending on the form of the outcome model.

Since $\mathbf{B} = E[\mathbf{u}\mathbf{u}^T]$, we define $\mathbf{B}$ as the following.

identity link	$\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1} = \mathbf{X}$
	$\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} = \mathbf{X}$
log link	$\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1} = m_1(\mathbf{X},\boldsymbol{\alpha}_1)\mathbf{X}$
	$\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} = m_0(\mathbf{X},\boldsymbol{\alpha}_0)\mathbf{X}$
logit link	$\frac{\partial m_1(\mathbf{X},\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1} = m_1(\mathbf{X},\boldsymbol{\alpha}_1)(1-m_1(\mathbf{X},\boldsymbol{\alpha}_1))\mathbf{X}$
	$\frac{\partial m_0(\mathbf{X},\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} = m_0(\mathbf{X},\boldsymbol{\alpha}_0)(1-m_0(\mathbf{X},\boldsymbol{\alpha}_0))\mathbf{X}$

$$\mathbf{B} = \begin{pmatrix} E[(m_1(\mathbf{X},\boldsymbol{\alpha}_1) - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_1)(Z(1-e(\mathbf{X},\boldsymbol{\beta}) + (1-Z)e(\mathbf{X},\boldsymbol{\beta}))] \\ E[(Y - m_1(\mathbf{X},\boldsymbol{\alpha}_1) - \delta_2)Z(1-e(\mathbf{X},\boldsymbol{\beta}))] \\ E[(Y - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_3)(1-Z)e(\mathbf{X},\boldsymbol{\beta})] \\ E[(Y - m_1(\mathbf{X},\boldsymbol{\alpha}_1))Z\mathbf{X}] \\ E[(Y - m_0(\mathbf{X},\boldsymbol{\alpha}_0))(1-Z)\mathbf{X}] \\ E[\mathbf{X}(Z - e(\mathbf{X},\boldsymbol{\beta}))] \end{pmatrix} \times$$

$$\begin{pmatrix} E[(m_1(\mathbf{X},\boldsymbol{\alpha}_1) - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_1)(Z(1-e(\mathbf{X},\boldsymbol{\beta}) + (1-Z)e(\mathbf{X},\boldsymbol{\beta}))] \\ E[(Y - m_1(\mathbf{X},\boldsymbol{\alpha}_1) - \delta_2)Z(1-e(\mathbf{X},\boldsymbol{\beta}))] \\ E[(Y - m_0(\mathbf{X},\boldsymbol{\alpha}_0) - \delta_3)(1-Z)e(\mathbf{X},\boldsymbol{\beta})] \\ E[(Y - m_1(\mathbf{X},\boldsymbol{\alpha}_1))Z\mathbf{X}] \\ E[(Y - m_0(\mathbf{X},\boldsymbol{\alpha}_0))(1-Z)\mathbf{X}] \\ E[\mathbf{X}(Z - e(\mathbf{X},\boldsymbol{\beta}))] \end{pmatrix}^T$$

We can estimate these quantities by plugging in the sample average for each expectation. Since $\hat{\Delta}_{DR,ATO} = \hat{\delta}_1 + \hat{\delta}_2 - \hat{\delta}_3$, the variance is estimated by the following.

$\hat{\text{var}}(\hat{\Delta}_{DR,ATO}) = (1,1,-1,\mathbf{0},\mathbf{0},\mathbf{0})\frac{1}{n}\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-T}(1,1,-1,\mathbf{0},\mathbf{0},\mathbf{0})^T$

Appendix C includes R code to calculate this.

3.7 Appendix B1. Proof of the doubly robust property of the ATO doubly robust estimator when the outcome model is correctly specified

We will first prove that when the outcome model is correctly specified, that is $m_1(\mathbf{X},\boldsymbol{\alpha}_1)$ and $m_0(\mathbf{X},\boldsymbol{\alpha}_0)$ are correctly specified, (3.6) will yield an unbiased estimator for the ATO effect (Δ).

$m_1(\mathbf{X},\boldsymbol{\alpha}_1) = E[Y|Z=1,\mathbf{X}] = E[Y_1|Z=1,\mathbf{X}] = E[Y_1|Z,\mathbf{X}] = E[Y_1|\mathbf{X}]$, similarly $m_0(\mathbf{X},\boldsymbol{\alpha}_0) = E[Y|Z=0,\mathbf{X}] = E[Y_0|Z=0,\mathbf{X}] = E[Y_0|Z,\mathbf{X}] = E[Y_0|\mathbf{X}]$, and $(m_1(\mathbf{X},\boldsymbol{\alpha}_1), m_0(\mathbf{X},\boldsymbol{\alpha}_0)) \perp Z|\mathbf{X}$ since we are assuming that there are no unmeasured confounders.

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n ((1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i))(m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1) - m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0))}{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)}$$

$\hat{\delta}_1$ is consistent for

$$
\begin{aligned}
&\frac{E[((1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z))(m_1(\mathbf{X},\boldsymbol{\alpha}_1)-m_0(\mathbf{X},\boldsymbol{\alpha}_0))]}{E[((1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} \\
&= \frac{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z+[e(\mathbf{X},\boldsymbol{\beta})(1-Z))(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} \\
&= \frac{E[(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])((1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z))]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} \\
&= \frac{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])]+E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z]+E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} \\
&= \frac{E[E[(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])((1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z))|Z,\mathbf{X}]]}{E[E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z)|Z,\mathbf{X}]} \\
&= \frac{E[E[(E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])|Z,\mathbf{X}]]E[E[((1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z))|Z,\mathbf{X}]]}{E[E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z+e(\mathbf{X},\boldsymbol{\beta})(1-Z)|Z,\mathbf{X}]} \\
&= E[E[Y_1|\mathbf{X}]-E[Y_0|\mathbf{X}])] \\
&= \Delta
\end{aligned}
$$

Therefore, $\hat{\delta}_1$ converges to Δ.

Since $m_1(\mathbf{X},\boldsymbol{\alpha}_1)$ is correctly specified, $E[Y_1 - m_1(\mathbf{X},\boldsymbol{\alpha}_1)] = 0$ and $(Y_1 - m_1(\mathbf{X},\boldsymbol{\alpha}_1)) \perp Z|\mathbf{X}$

$$
\begin{aligned}
&E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y-E[Y|Z=1,\mathbf{X}])] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y_1-E[Y|Z=1,\mathbf{X}])] \\
&\quad = E[E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y_1-E[Y|Z=1,\mathbf{X}])|Z,\mathbf{X}]] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))ZE[(Y_1-E[Y|Z=1,\mathbf{X}])|Z,\mathbf{X}]] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_1|Z,\mathbf{X}]-E[Y|Z=1,\mathbf{X}])] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_1|\mathbf{X}]-E[Y_1|\mathbf{X}])] = 0
\end{aligned}
$$

Similarly, since $m_0(\mathbf{X},\boldsymbol{\alpha}_0)$ is correctly specified, $E[Y_0 - m_0(\mathbf{X},\boldsymbol{\alpha}_0)] = 0$ and $(Y_0 - m_0(\mathbf{X},\boldsymbol{\alpha}_0)) \perp Z|\mathbf{X}$

$$
\begin{aligned}
&E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y-E[Y|Z=0,\mathbf{X}])] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y_0-E[Y|Z=0,\mathbf{X}])] \\
&\quad = E[E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(Y_0-E[Y|Z=0,\mathbf{X}])|Z,\mathbf{X}]] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))ZE[(Y_0-E[Y|Z=0,\mathbf{X}])|Z,\mathbf{X}]] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_0|Z,\mathbf{X}]-E[Y|Z=0,\mathbf{X}])] \\
&\quad = E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z(E[Y_0|\mathbf{X}]-E[Y_0|\mathbf{X}])] = 0
\end{aligned}
$$

Plugging these into the Equation (3.6), $\hat{\Delta}_{DR,ATO}$, we have demonstrated that when the outcome model is correctly specified, that is $m_1(\mathbf{X},\boldsymbol{\alpha}_1)$ and $m_0(\mathbf{X},\boldsymbol{\alpha}_0)$ are correctly specified $\hat{\Delta}_{DR,ATO} \to_p \Delta$

3.8 Appendix B2. Proof of the doubly robust property of the ATO doubly robust estimator when the propensity score model is correctly specified

We will now prove that when the propensity score model is correctly specified $\hat{\Delta}_{DR,ATO}$ converges to Δ.

We can rewrite the Equation (3.6) as

$$
\begin{aligned}
\hat{\Delta}_{DR,ATO} = &\left\{ \frac{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_iY_i}{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} - \frac{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)Y_i}{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right\} \\
&+ \left\{ \frac{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1) + \sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)}{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + \sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right. \\
&\qquad \left. - \frac{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)}{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i} \right\} \\
&+ \left\{ \frac{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_im_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0) + \sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)}{\sum_{i=1}^n (1-e(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))Z_i + \sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right. \\
&\qquad \left. - \frac{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)}{\sum_{i=1}^n e(\mathbf{X}_i,\hat{\boldsymbol{\beta}})(1-Z_i)} \right\}
\end{aligned}
$$

$e(\mathbf{X},\boldsymbol{\beta}) = e(\mathbf{X}) = E[Z|\mathbf{X}] = E[Z|Y_1,\mathbf{X}]$ by no unmeasured confounders.

The first term converges to

$$\frac{E[(1-e(\mathbf{X},\boldsymbol{\beta}))ZY]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z]} - \frac{E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)Y]}{E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)]}$$

This is a consistent estimator for Δ, as this is the estimator, $\hat{\tau}_{AT0}$ (Equation (3.1)) defined by Li et al. (Theorem 1 (Li, Morgan, and Zaslavsky 2016)).

We now need to show that the second and third term are 0.

The second term converges to

$$\frac{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)] + E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)m_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z] + E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)]} - \frac{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z]}$$

The $E[(1-e(\mathbf{X},\boldsymbol{\beta}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)] = E[(e(\mathbf{X},\boldsymbol{\beta}))(1-Z)m_1(\mathbf{X},\boldsymbol{\alpha}_1)]$ and $E[(1-e(\mathbf{X},\boldsymbol{\beta}))Z] = E[e(\mathbf{X},\boldsymbol{\beta})(1-Z)]$, therefore the this is equivalent to

$$\frac{E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)] + E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X}))Z] + E[(1-e(\mathbf{X}))Z]} - \frac{E[(1-e(\mathbf{X}))Zm_1(\mathbf{X},\boldsymbol{\alpha}_1)]}{E[(1-e(\mathbf{X}))Z]} = 0$$

Similarly, for the third term, $E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)] = E[(e(\mathbf{X}))(1-Z)m_0(\mathbf{X},\boldsymbol{\alpha}_0)]$ therefore

$$\frac{E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)] + E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)]}{E[(1-e(\mathbf{X}))Z] + E[(1-e(\mathbf{X}))Z]} - \frac{E[(1-e(\mathbf{X}))Zm_0(\mathbf{X},\boldsymbol{\alpha}_0)]}{E[(1-e(\mathbf{X}))Z]} = 0$$

Therefore, $\hat{\Delta}_{ATO,DR} \to_p \Delta$

3.9 Appendix C. R Code to calculate the large-sample variance for the ATO doubly robust estimator

The following functions will calculate the ATO doubly robust estimator (Equation (3.6)) along with the large-sample variance (Equation (3.16)) derived in Appendix A2.

The `sandwich_ato` function will output a `tibble` with the doubly robust estimator (`est`) and the appropriate standard deviation (`se`).

```
sandwich_ato <- function(data,
                         ps_form,
                         outcome_form,
                         ps_family,
                         outcome_family) {

  ps_form <- as.formula(ps_form)
  ps_model <- glm(ps_form, data = data, family = ps_family)
  x <- names(coef(ps_model))[-1]
  z <- data[, all.vars(ps_form)[1]]

  outcome_form <- gsub(glue::glue(all.vars(ps_form)[1], " \\+|",
                                  all.vars(ps_form)[1], "\\+") ,
                       "", outcome_form)
  outcome_form <- as.formula(outcome_form)
  out_model_y1 <- glm(outcome_form,
                      data = data[z == 1, ],
                      family = outcome_family)
  out_model_y0 <- glm(outcome_form,
                      data = data[z == 0, ],
                      family = outcome_family)
  v <- names(coef(out_model_y1))[-1]
  y <- data[, all.vars(outcome_form)[1]]

  n <- nrow(data)

  ps <- predict(ps_model, type = "response")
  y1 <- predict(out_model_y1, newdata = data, type = "response")
  y0 <- predict(out_model_y0, newdata = data, type = "response")

  delta1 <- (sum((1 - ps) * z * (y1 - y0))
             + sum(ps * (1 - z) * (y1 - y0))) /
    (sum((1 - ps) * z ) + sum(ps * (1 - z)))
  delta2 <- sum((1 - ps) * z * (y - y1)) / sum((1 - ps) * z)
```

```
  delta3 <- sum(ps * (1 - z) * (y - y0)) / sum(ps * (1 - z))

  data$z <- z
  data$y <- y
  data$ps <- ps
  data$y1 <- y1
  data$y0 <- y0

  u <- purrr::pmap(data, build_u,
                   x = x, v = v,
                   delta1 = delta1,
                   delta2 = delta2,
                   delta3 = delta3)
  l_b <- purrr::map(u, ~ outer(.x, .x))
  B <- purrr::reduce(l_b, `+`) / n
  l_a <- purrr::pmap(data, build_a,
                     x = x, v = v,
                     delta1 = delta1,
                     delta2 = delta2,
                     delta3 = delta3,
                     family = outcome_family)
  A <- purrr::reduce(l_a, `+`) / n
  yum <- solve(A) %*% B %*% t(solve(A)) / n
  var <-
    t(c(1, 1, -1, rep(0, nrow(A) - 3))) %*%
    yum %*%
    c(1, 1, -1, rep(0, nrow(A) - 3))
  se <- sqrt(var)
  tibble::tibble(est = delta1 + delta2 - delta3, se = as.numeric(se))
}

build_u <- function(z, ps,
                    y, y1, y0,
                    x, v,
                    delta1, delta2, delta3, ...) {
  dots <- list(...)
```

```
  x <- matrix(c(1, unlist(dots[x])), nrow = 1 + length(x))
  v <- matrix(c(1, unlist(dots[v])), nrow = 1 + length(v))
  u1 <- ((1 - ps) * z + ps * (1 - z)) * (y1 - y0 - delta1)
  u2 <- (1 - ps) * z * (y - y1 - delta2)
  u3 <- ps * (1 - z) * (y - y0 - delta3)
  u4 <- v %*% z * (y - y1)
  u5 <- v %*% (1 - z) * (y - y0)
  u6 <- x %*% (z - ps)
  c(u1, u2, u3, u4, u5, u6)
}

build_a <- function(z, ps,
                    y, y1, y0,
                    x, v,
                    delta1, delta2, delta3,
                    family, ...) {
  dots <- list(...)
  x <- matrix(c(1, unlist(dots[x])), nrow = 1 + length(x))
  v <- matrix(c(1, unlist(dots[v])), nrow = 1 + length(v))

  y1_deriv <- y_deriv(family, y1, v)
  y0_deriv <- y_deriv(family, y0, v)

  # derivative with respect to delta1
  a_delta1 <- c(((1 - ps) * z + ps * (1 - z)),
                rep(0, (2 + 2 * length(v) + length(x))))
  # derivative with respect to delta2
  a_delta2 <- c(0, (1 - ps) * z,
                rep(0, (1 + 2 * length(v) + length(x))))
  # derivative with respect to delta3
  a_delta3 <- c(0, 0, ps * (1 - z),
                rep(0, (2 * length(v) + length(x))))
  # derivative with respect to alpha1
  a_alpha1 <- matrix(
    c(-((1 - ps) * z + ps * (1 - z)) * y1_deriv,
      (1 - ps) * z * y1_deriv,
```

```
      matrix(0, ncol = length(v)),
      y1_deriv %*% t(v) * z,
      matrix(0, ncol = length(v), nrow = c(length(v) + length(x)))),
    ncol = length(v), byrow = TRUE)
  # derivative with respect to alpha0
  a_alpha0 <- matrix(
    c(((1 - ps) * z + ps * (1 - z)) * y0_deriv,
      matrix(0, ncol = length(v)),
      ps * (1 - z) * y0_deriv,
      matrix(0, ncol = length(v), nrow = length(v)),
      y0_deriv %*% t(v) *(1 - z),
      matrix(0, ncol = length(v), nrow = length(x))),
    ncol = length(v), byrow = TRUE)
  # derivative with respect to beta
  a_beta <- matrix(
    c(- x %*%
        ((- ps * (1 - ps)) * z + (ps * (1 - ps)) * (1 - z)) *
        (y1 - y0 - delta1),
      x %*% ps * (1 - ps) * z * (y - y1 - delta2),
      - x %*% ps * (1 - ps) * (1 - z) * (y - y0 - delta3),
      matrix(0, ncol = length(x), nrow = c(2 * length(v))),
      x %*% t(x) * ps * (1 - ps)),
    ncol = length(x), byrow = TRUE)
  cbind(a_delta1, a_delta2, a_delta3, a_alpha1, a_alpha0, a_beta)
}

y_deriv <- function(family, y, v) {
  if (family == "gaussian") {
    y_deriv <- v
  }

  if (family == "binomial") {
    y_deriv <- y * (1 - y) * v
  }

  if (family == "poisson") {
```

```
    y_deriv <- y * v
  }
  as.matrix(y_deriv)
}
```

Here is an example using the `sandwich_ato` function along with the data from Chapter 2 (available on GitHub (https://github.com/LucyMcGowan/dr-example-code)).

```
df_url <- "http://bit.ly/df_continuous"
load(url(df_url))
sandwich_ato(df_continuous,
             ps_form = "z ~ x_1 + x_2",
             outcome_form = "y ~ z + x_1",
             ps_family = binomial("probit"),
             outcome_family = "gaussian")
```

```
## # A tibble: 1 x 2
##     est    se
##   <dbl> <dbl>
## 1  1.01 0.101
```

CHAPTER 4

CONTEXTUALIZED TIPPING POINT SENSITIVITY ANALYSES FOR UNMEASURED CONFOUNDING

4.1 Background

The strength of the evidence provided by observational studies is inherently limited by the potential influence of unmeasured confounding variables. This limitation should neither be ignored nor used as a blanket dismissal of all observational studies' findings. Researchers should quantify the aspects of a hypothetical confounder that could change the size of their observed effect or make a true null effect appear statistically significant. Every observational study with a statistically significant finding should include a quantified sensitivity to unmeasured confounding analysis. However, a 2008 systematic review by Groenwold et al. showed such analyses were rarely done (Groenwold et al. 2008). They examined 174 observational studies in five general medical journals and five epidemiological journals published between January 2004 and April 2007. While the potential for unobserved confounding was reported in 102 (58.6%) of reviewed articles, 15 (8.6%) commented on the potential effect of such remaining confounding and only 4 (2.3%) conducted sensitivity analysis to estimate potential impact of unobserved confounding. To see if the landscape had improved since then, we performed a review of 90 observational studies with statistically significant findings published in 2015 in the Journal of the American Medical Association, the New England Journal of Medicine, and the American Journal of Epidemiology. We saw little improvement with 41 (45.6%) mentioning the issue of unmeasured confounding as a limitation and only 4 (4.4%) performing a quantitative sensitivity analysis. Even when sensitivity analyses are performed, they can remain difficult for clinically oriented readers to understand. These deficiencies reveal the need for practical guidance and simple tools to help both the medical research community incorporate sensitivity analyses into their papers and readers perform such analyses themselves when a paper has failed to provide one.

One challenge of translating these methods into common practice is finding the right level of simplification. Consider the rule of thumb, "In a study with binary outcomes and binary exposures the relative risk may be off by a factor of 2, but unlikely to

be off more than that." (Belle 2011). While under simplified assumptions, this rule applies widely to generalized regression settings yielding relative risks, odds ratios, and hazard ratios; the criteria may be too liberal for studies missing one or more variables known to be strong confounders and too conservative for studies that adjust for all major known confounders. It ignores the study design's quality; whereas, a sensitivity analysis should guide the reader through evaluating it. A sensitivity analysis should focus the discussion on the rigor of the design, the thoroughness of capturing known confounders, and the plausibility of an unmeasured confounder or confounders being sufficient to nullify the conclusions. A well designed study that has controlled for several important confounders via matching, weighting, and/or regression-based covariate adjustment can provide the context in which the hypothetical unmeasured confounder's properties should be viewed.

This article re-frames the work of Cornfield et al. (Cornfield et al. 1959), Schlesselman (Schlesselman 1978), Rosenbaum and Rubin (Rosenbaum and Rubin 1983), and Lin, Psaty, and Kronmal (Lin, Psaty, and Kronmal 1998) to focus on three quantities within the context of a study with a binary exposure showing a statistically significant effect. The quantities are:

1. The bound of the confidence interval for the exposure's observed effect that is closer to the null, i.e. the bound closer to 1 for an odds ratio, hazards ratio, or relative risk.
2. A strength of the association between an unmeasured binary confounder and the outcome based on clinical knowledge and/or the observed effects of the measured covariates.
3. Given 1 and 2, calculate the differential prevalence of the unmeasured binary confounder between the exposed and unexposed populations needed to nullify the statistically significant effect.

Focusing on these three quantities allows us to simplify the methods to a tipping point analysis, that can be referenced by researchers wanting to include a quantitative sensitivity analyses and by readers wishing to understand the sensitivity of studies that failed to include such an analysis.

Additionally, VanderWeele and Ding recently suggested a tipping point sensitivity analysis simplification referred to as the E-value (Ding and VanderWeele 2016; VanderWeele and Ding 2017). We extend this E-value to a setting where one can calculate an "observed E-value" for each measured confounding, contextualizing the sensitivity

analysis.

This paper will demonstrate best practices for calculating these tipping point analyses under a variety of scenarios, with a focus on contextualizing the sensitivity analysis using observed confounders, in the spirit of Hsu and Small (Hsu and Small 2013).

There has been a large amount of research in this area; a brief history is included in Appendix A.

4.2 Methods

4.2.1 Tipping point calculation

The main objective of a tipping point sensitivity analysis is to report the qualities of an unmeasured confounder needed to change the statistical significance of one's findings. For example, a hazard ratio of 1.25 with a 95% confidence interval (1.1, 1.5) would no longer be significant at the $\alpha = .05$ level if adjusting for a hypothetical unmeasured confounder caused the lower bound to cross 1. The "tipping point" analysis would find the smallest possible effect of an unmeasured confounder that would cause this to happen. The methods explained here apply to both binary outcomes, analyzed using logistic regression, as well as survival time outcomes with censoring, analyzed using proportional hazards models.

To determine whether an exposure, Z, is associated with an outcome, Y, one can observe whether the relative risk, odds ratio, or hazard ratio of Z is equal to 1. As a tipping point analysis, we are interested in which values of an unmeasured confounder would cause the lower or upper confidence interval of the association measure to cross the null; we refer to this bound closest to the null as the "limiting bound", or LB. Lin et al. (Lin, Psaty, and Kronmal 1998), show that the observed association between Z and Y can be adjusted based on the size and prevalence of an independent unmeasured confounder U, for a binary unmeasured confounder, and the size and mean difference between exposure groups for a continuous unmeasured confounder. Under the assumption that the sensitivity parameters are fixed, the variance of the observed effect is the same as the variance of the adjusted effect. This allows all adjustments to apply to confidence intervals the same way they would apply to point estimates. Lin et al. algebraically derive equations to update biased estimates in log-linear regression for unmeasured confounders. Simulations show that these sensitivity analyses can be extended to the logistic regression and censored survival time cases (Lin, Psaty, and

Kronmal 1998). The relationship for the binary unmeasured confounder is as shown in Equation (4.1).

$$LB_{adj} = LB_{obs}\frac{RR_{UD}p_0 + (1 - p_0)}{RR_{UD}p_1 + (1 - p_1)} \tag{4.1}$$

Where LB_{adj} is the limiting bound of the risk ratio, odds ratio, or hazard ratio for Z adjusting for the unmeasured confounding and known confounders, LB_{obs} is the observed limiting bound obtained from the model including known confounders but excluding the unmeasured confounder, p_1 and p_0 are the prevalences of the unmeasured confounder in the exposed and unexposed populations, respectively, and RR_{UD} is the association between the unmeasured confounder and the outcome both in the presence and absence of the exposure, i.e. with the assumption of no interaction. Notice here the unmeasured confounder is assumed to be binary, as we are estimating prevalences in the exposed and unexposed populations. Using a similar equation, Lin et al. derive the relationship between a continuous unmeasured confounder (normally distributed, $U \sim N(\mu_Z, 1)$) and an outcome.

$$LB_{adj} = \frac{LB_{obs}}{RR_{UD}^{\mu_1 - \mu_0}} \tag{4.2}$$

Where μ_1 is the mean of the unmeasured confounder in the exposed population, and μ_0 is the mean of the unmeasured confounder in the unexposed population. Notice here the variance is assumed to be 1. Any normally distributed confounder can fit this specification by scaling by the standard deviation within each exposure group.

In order to encourage widespread use of this methodology, we offer a rearranged version of these equations. We use the relationship shown in the equations above to solve for the minimum RR_{UD} with varying levels of p_0 and p_1 in the binary case, and μ_1 and μ_0 in the continuous case, such that the original association is no longer statistically significant, in the binary case setting LB_{adj} equal to 1 (Equation (4.3)).

$$RR_{UD}(LB_{obs}, p_0, p_1) = \frac{(1 - p_1) + LB_{obs}(p_0 - 1)}{LB_{obs}p_0 - p_1} \tag{4.3}$$

This would allow investigators to state, "A hypothetical unobserved binary confounder that is prevalent in p_1 of the exposed population and p_0 of the unexposed population would need to have an association with Y of RR_{UD} to tip this analysis at the 5%

level, rendering it inconclusive." Similarly we have rearranged this equation to solve for p_1 or p_0, given the remaining parameters. For example, solving for p_1 is shown in Equation (4.4).

$$p_1(LB_{obs}, RR_{UD}, p_0) = \frac{LB_{obs}(p_0(RR_{UD} - 1) + 1) - 1}{RR_{UD} - 1} \tag{4.4}$$

Similarly, solving for p_0 results in Equation (4.5).

$$p_0(LB_{obs}, RR_{UD}, p_1) = \frac{p_1(RR_{UD} - 1) - LB_{obs} + 1}{LB_{obs}(RR_{UD} - 1)} \tag{4.5}$$

Suppose that we are interested in a number of small unmeasured confounders that would tip the analysis. We can solve for n, the number of independent unmeasured confounders that would cause this analysis to tip as follows (Equation (4.6)).

$$n(LB_{obs}, RR_{UD}, p_1, p_0) = \frac{-\log(LB_{obs})}{\log\{RR_{UD}p_0 + (1 - p_0)\} - \log\{RR_{UD}p_1 + (1 - p_1)\}} \tag{4.6}$$

In the continuous case, the relationship is only dependent on the difference between the means, μ_1 and μ_0, rather than the means themselves, so we can simplify equation 2, replacing $\mu_1 - \mu_0$ with δ. Rearranging equation 2 as a tipping point analysis results in Equation (4.7).

$$RR_{UD}(LB_{obs}, \delta) = LB_{obs}^{1/\delta} \tag{4.7}$$

Solving for the unmeasured confounder's difference in means between exposure groups results in Equation (4.8).

$$\delta(LB_{obs}, RR_{UD}) = \frac{\log(LB_{obs})}{\log(RR_{UD})} \tag{4.8}$$

Similar to the binary case (Equation (4.6)), we can also solve for n the number of unmeasured confounders we would need to tip the analysis with a given δ and RR_{UD}.

$$n(LB_{obs}, RR_{UD}, \delta) = \frac{\log(LB_{obs})}{\delta \log(RR_{UD})} \tag{4.9}$$

Building on this methodology, Ding and VanderWeele offer an "assumption free" method that no longer requires that the unmeasured confounding be binary, but rather represent this relationship as relative risk, in the binary case represented as $RR_{EU} = p_1/p_0$. (Ding and VanderWeele 2016) They further recommend reporting the minimum RR_{UD} needed to tip under a particular RR_{EU}. In the binary case, this is equivalent setting p_1 to 1, varying p_0 from 0 to 1. Since we are interested in the tipping point such that the original association is no longer statistically significant, the adjusted limiting bound, LB_{adj} is set equal to 1 (Equation (4.10)).

$$1 = LB_{obs}\frac{RR_{UD}/RR_{EU} + (1 - 1/RR_{EU})}{RR_{UD}} \tag{4.10}$$

VanderWeele and Ding further suggest focusing on the point that minimizes the strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the exposure and outcome, conditional on the measured covariates, to explain away an observed exposure-outcome association (Ding and VanderWeele 2016; VanderWeele and Ding 2017). They call this value an "E-value" (Equation (4.11)).

$$\text{E-value} = LB_{obs} + \sqrt{LB_{obs} \times (LB_{obs} - 1)} \tag{4.11}$$

This E-value demonstrates the joint minimum strength of association with both the exposure and outcome needed to tip the analysis (allow the lower bound to cross one, i.e. $LB_{adj} = 1$). If one was interested in a different tipping point, we can reintroduce the LB_{adj} into the equation as shown in Equation (4.12).

$$\text{E-value}_{adj} = \frac{LB_{obs}}{LB_{adj}} + \sqrt{\frac{LB_{obs}}{LB_{adj}} \times \left(\frac{LB_{obs}}{LB_{adj}} - 1\right)} \tag{4.12}$$

We will demonstrate the utility of each of these equations below.

4.2.2 Software

We have created an R package (R Core Team 2017), `tipr`, that allows for the implementation of this method. It can be installed by running the following.

The `tip_with_binary()` function takes the following arguments:

- `p1`: estimated prevalence of the unmeasured confounder in the exposed population
- `p0`: estimated prevalence of the unmeasured confounder in the unexposed population
- `gamma`: estimated size of an unmeasured confounder
- `lb`: lower bound of your observed effect
- `ub`: upper bound of your observed effect

A user can supply this function with any two of the three sensitivity parameters, `p1`, `p0`, and `gamma`, as well as the upper confidence bound (`ub`) and lower confidence bound (`lb`) of the observed effect, and the size of the third parameter needed to tip the analysis will be calculated. If all three parameters (`p1`, `p0`, and `gamma`) are specified, the function will return *number* of independent unmeasured confounders of the size and prevalence specified will be needed to tip the analysis.

Similarly, the `tip_with_continuous()` function takes the following arguments:

- `mean_diff`: estimated mean difference of the unmeasured confounder between the exposure groups
- `gamma`: estimated size of an unmeasured confounder
- `lb`: lower bound of your observed effect
- `ub`: upper bound of your observed effect

Note that only the limiting bound, the bound closer to the null, is actually utilized in the calculation, but for ease of use we ask for both the upper and lower confidence bounds and determine the limiting bound for the user. The utility will be demonstrated in the examples below.

4.2.3 Tipping point contextualization

There are three main quantities that make an unmeasured confounder, or any covariate for that matter, meaningful:

1. How imbalanced is the unmeasured confounder between the exposure groups?
2. How predictive is the unmeasured confounder of the outcome?
3. How independent is the unmeasured confounder from the other covariates?

The first two quantities have a straight forward impact; the more imbalanced the unmeasured confounder is between exposure groups and the more predictive the

unmeasured confounder is of the outcome, the larger it's impact. Generally, the more "independent" the unmeasured confounder is from other covariates, given it is a confounder, the larger it's impact. In other words, if the covariates you have already measured account for the majority of the variability in the exposure or outcome that would be explained by the unmeasured confounder, then the impact of missing this confounder will be less pronounced. An assumption of the equations above is that the unmeasured confounder is independent from all measured covariates, making it a conservative way to assess sensitivity.

We propose that perhaps having a good understanding of the confounders that were *measured* will assist in conceptualizing and constructing plausible scenarios for sensitivity analyses for confounders that are *unmeasured*. We suggest examining the imbalance between exposure groups using standardized mean differences, visualized through a Love plot (Love 2002; Hansen and Fredrickson 2014), and examining the predictive power and independence of each covariate, using a new plot we have named the "observed bias" plot.

4.2.3.1 Love plots

Shifting our focus to the measured covariates, there are many tools suggested to examine this first quantity, the imbalance between the exposure groups (Austin 2009; Groenwold et al. 2011; Li, Morgan, and Zaslavsky 2016; Stuart, Lee, and Leacy 2013; Rubin 2001; Imai, King, and Stuart 2008). A common metric is standardized mean difference. For continuous covariates, standardized mean difference is calculated as seen in Equation (4.13).

$$d = \frac{(\bar{x}_{\text{exposed}} - \bar{x}_{\text{unexposed}})}{\sqrt{\frac{s^2_{\text{exposed}} + s^2_{\text{unexposed}}}{2}}} \tag{4.13}$$

Where $\bar{x}_{\text{exposed}}$ and s^2_{exposed} are the sample mean and variance in the exposed group, and, similarly, $\bar{x}_{\text{unexposed}}$ and $s^2_{\text{unexposed}}$ are the sample mean and variance in the unexposed group. For binary covariates, standardized mean difference is calculated as seen in Equation (4.14).

$$d = \frac{(\hat{p}_{\text{exposed}} - \hat{p}_{\text{unexposed}})}{\sqrt{\frac{\hat{p}_{\text{exposed}}(1-\hat{p}_{\text{exposed}}) + \hat{p}_{\text{unexposed}}(1-\hat{p}_{\text{unexposed}})}{2}}} \tag{4.14}$$

Where $\hat{p}_{\text{exposed}}$ and $\hat{p}_{\text{unexposed}}$ are the prevalence of the dichotomous variable in the exposed and unexposed subjects.

One can calculate the standardized mean difference for each covariate before and after propensity score adjustment; these quantities are often visualized by ordering by their magnitude in the unadjusted cohort and plotting the values in a figure known as a "Love plot" (Love 2002; Hansen and Fredrickson 2014). This metric and associated plot give important insight into the covariate balance before and after propensity score adjustment, however it only addresses one of the three crucial quantities.

In addition to Love plots, side-by-side boxplots and empirical cumulative distribution fuctions can be used to compare the distribution of continuous covariates between the exposed and unexposed subjects pre-and post-propensity score adjustment, as suggested by Austin and Stuart (Austin and Stuart 2015; Joffe et al. 2004). This can give more detailed insight into the full distribution of the covarite, ensuring that the propensity score adjustment balances the full distribution.

4.2.3.2 Observed bias plots

We propose an additional plot in addition to the Love plot, an "observed bias plot". This plot demonstrates how much leaving each single covariate out of the full modeling process "biases" the final result, the effect of the exposure on the outcome. The general idea is similar to the "omitted variable bias" discussed by Hosman et al. (Hosman, Hansen, and Holland 2010). Here, we are interested in how omitting each covariate shifts the point estimate and 95% confidence interval of the exposure-outcome effect. To create this plot, we first fit our model(s) as we normally would. In the case of an analysis that includes propensity score adjustment, for example, we fit the propensity score model and then the outcome model, estimating the exposure-outcome effect. We then repeat the entire process, leaving one covariate out at a time, and record the exposure effect and 95% confidence interval each time. We plot this exposure effect for every covariate, demonstrating how the effect of interest would change had we not observed the covariate at hand.

In addition to plotting the exposure effect for each covariate, we can also plot the adjusted E-value for each covariate. That is the E-value for moving the observed lower bound (the lower bound of the effect observed without the unmeasured confounder) to the adjusted lower bound (the lower bound of the effect with the unmeasured confounder), using Equation (4.12). This adds context to the E-value, allowing it to

be somewhat grounded in the observed covariates.

These observed bias plots need not be limited to the effect of leaving out each confounder one at a time. For example, it may be of interest to see the effect of leaving out a group of confounders. In the example here, we leave out all lab values to demonstrate how that would have changed our analysis. In addition, we can add a shifted effect for a hypothetical unmeasured confounder that would tip this analysis, i.e. bring the lower bound of the effect to 1, as well as a hypothetical unmeasured confounder that would bring the point estimate to 1.

4.3 Examples

We have constructed a series of scenarios we have seen prevalent in the medical literature where a contextualized sensitivity analysis may have been useful. In each of these scenarios, we are assuming that the result of interest, the association between the exposure and outcome, is significant.

1. You observe a particularly imbalanced covariate. If you missed an unmeasured confounder that has the same imbalance as this, that is independent of all the observed covariates, how predictive of the outcome would it need to be in order to tip your analysis?

2. You observe a covariate strongly associated with the outcome. If you missed an unmeasured confounder that has the same association as this, that is independent of all the observed covariates, how imbalanced between exposures would it need to be in order to tip your analysis?

3. You observe many covariates that are all slightly associated with the outcome. How many independent covariates of this magnitude would be needed to tip your analysis.

4. You calculate the E-value for your study and would like to ground this in your observed covariates.

To demonstrate each of these scenarios, we will use the Right Heart Catheterization dataset, originally used in Connors et al (Connors et al. 1996). This dataset was used to assess the effectiveness of right heart catheterization (RHC) in the initial care of critically ill patients. This cohort contains 5,735 patients, 2,184 in the treatment group (RHC) and 3,551 in the control group (no RHC). This is a particularly interesting

observational study, as it demonstrated a result counter to previously published recommendations for the use of RHC. The original analysis included 50 covariates used to estimate the propensity of being assigned to RHC. For demonstration purposes, we chose 20 to use here. We use demographics (age, sex), comorbidities (upper GI bleeding, renal disease, transfer status), physiological measurements taken on day 1 (bilirubin, hematocrit, white blood cell count, mean blood pressure, pH, PaO2/FiO2 ratio, albumin, respiratory rate, PaCO2, heart rate), diagnosis categories (Neurology and Hematology), APACHE score, SUPPORT model estimate of the probability of surviving 2 months, and DNR status on day 1. Please see Connors et al for the fully adjusted analysis and clinical interpretation of the RHC effect (Connors et al. 1996). We examine the balance using standardized mean differences and a Love plot. Additionally, we demonstrate the side-by-side boxplot and empirical cumulative distribution function for the continuous covariate APACHE score. We construct overlap weights (Li, Morgan, and Zaslavsky 2016) for each individual and perform a weighted survival analysis estimating the effect of right heart catheterization on 30 day survival, adjusting for all 20 covariates. We then fit the full analysis, leaving out one covariate at a time. Each time we estimate the effect of the exposure, right heart catheterization, on the outcome, 30 day survival, and compare it to the estimate with the fully specified analysis. Figure 4.1 displays the Love plot, Figure 4.2 displays the side-by-side boxplots and empirical cumulative distribution function for APACHE score, and Figure 4.3 displays the observed bias plot. The observed effect of RHC on 30 day survival is 1.24 (95% CI: 1.11, 1.37) (Table 4.1).

Table 4.1: The association with 30 day survival, adjusting for all other covariates.

	Hazard Ratio	95% LCL	95% UCL
RHC	1.24	1.11	1.37
APACHE score	1.00	1.00	1.01
WBC	1.00	1.00	1.00
Heart rate	1.00	1.00	1.00
PaO2/FIO2 ratio	1.00	1.00	1.00
Albumin	0.98	0.92	1.04
Hematocrit	1.00	0.99	1.01
Bilirubin	1.03	1.02	1.04
Mean blood pressure	1.00	1.00	1.00
PaCo2	0.99	0.99	1.00
PH	0.62	0.34	1.14

	Hazard Ratio	95% LCL	95% UCL
Respiratory rate	1.00	0.99	1.00
Age	1.00	1.00	1.01
Support prob. of surviving 2 months	0.08	0.06	0.11
Chronic Renal Disease	1.05	0.80	1.37
Upper GI Bleeding	1.58	1.23	2.03
Transfer Status	1.30	1.11	1.52
DNR status on day 1	2.59	2.22	3.02
Neurological Diagnosis	1.40	1.17	1.68
Hematologic Diagnosis	1.39	1.16	1.67
Sex	1.07	0.97	1.19

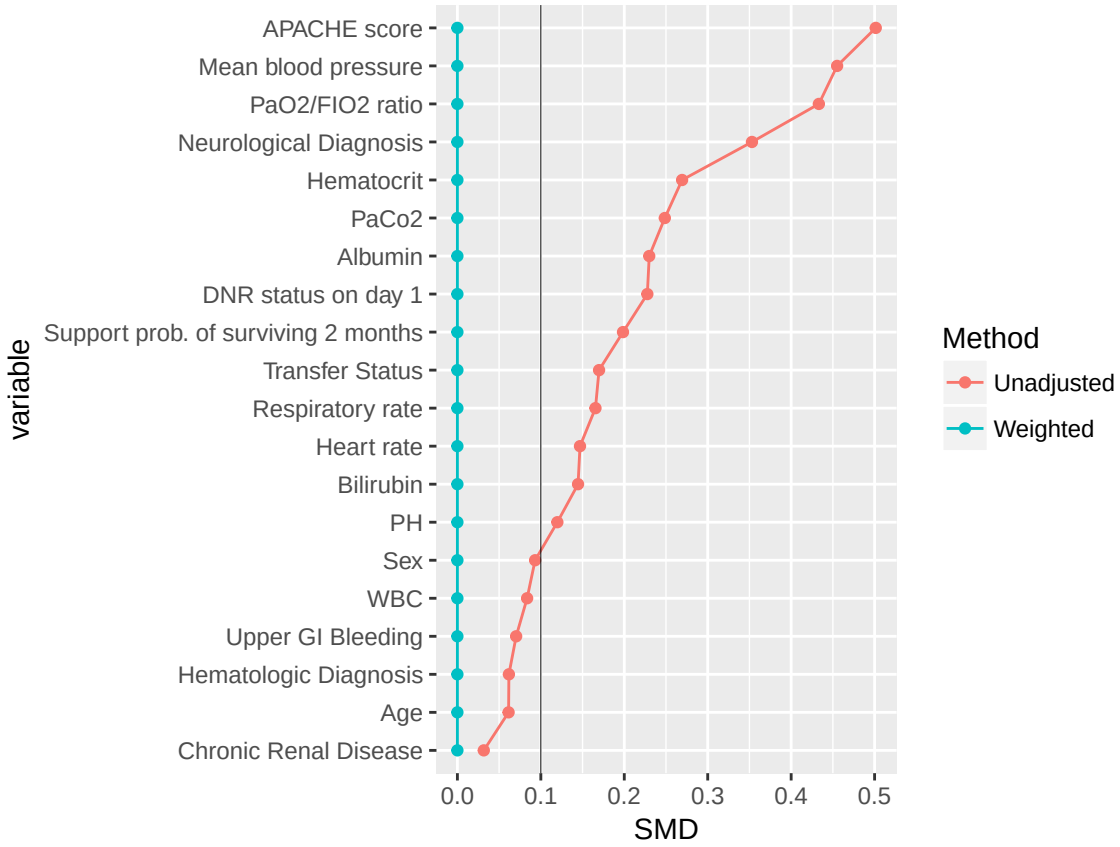

Figure 4.1: Love plot. This displays the standardized mean difference between the exposed and unexposed groups before (red) and after (blue) propensity score weighting. The vertical line at 0.1 represents the "rule of thumb" for an acceptable standardized mean difference.

Before diving into the scenarios, observe what we have learned from the two figures,

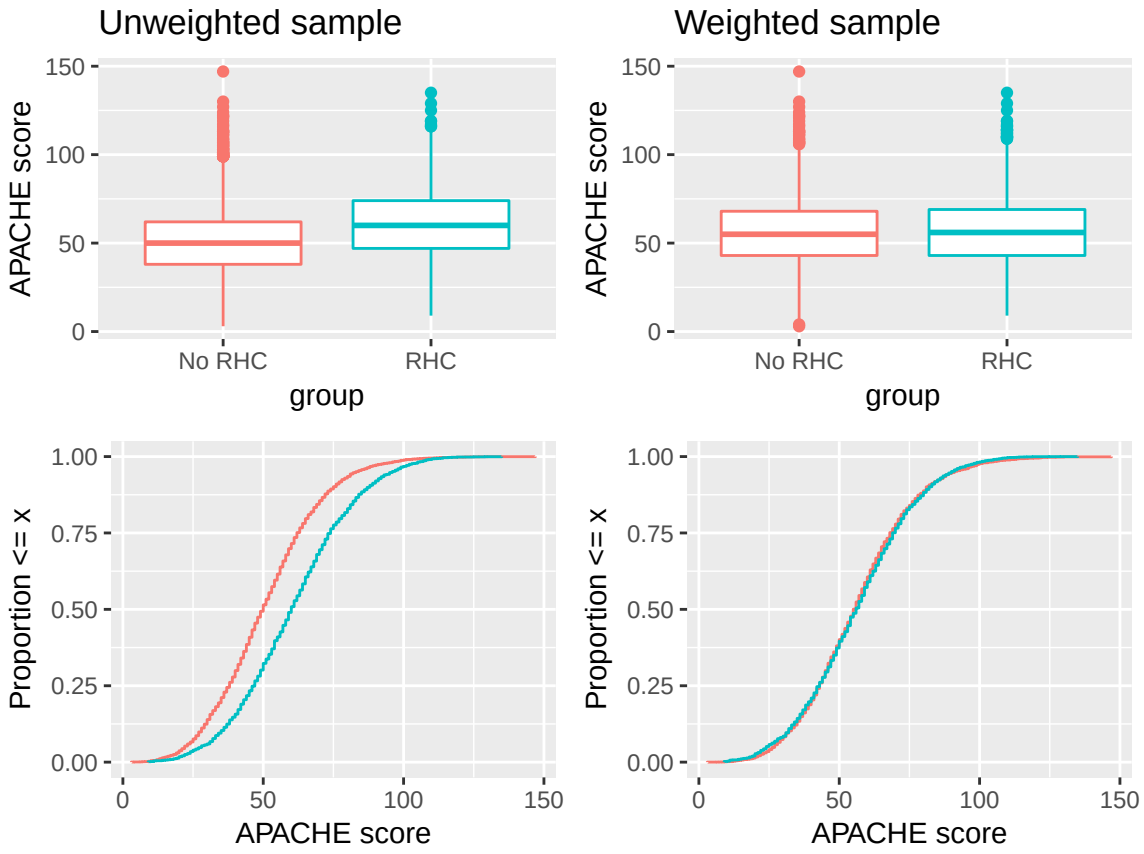

Figure 4.2: Distribution of APACHE score between exposed and unexposed subjects. The plots on the left represent the distribution among the unweighted sample, and the plots on the right represent the distribution among the propensity score weighted sample. The top plots are boxplots and the bottom plots are cumulative distributions. The blue represents the exposed, those with RHC, and the red represents the unexposed.

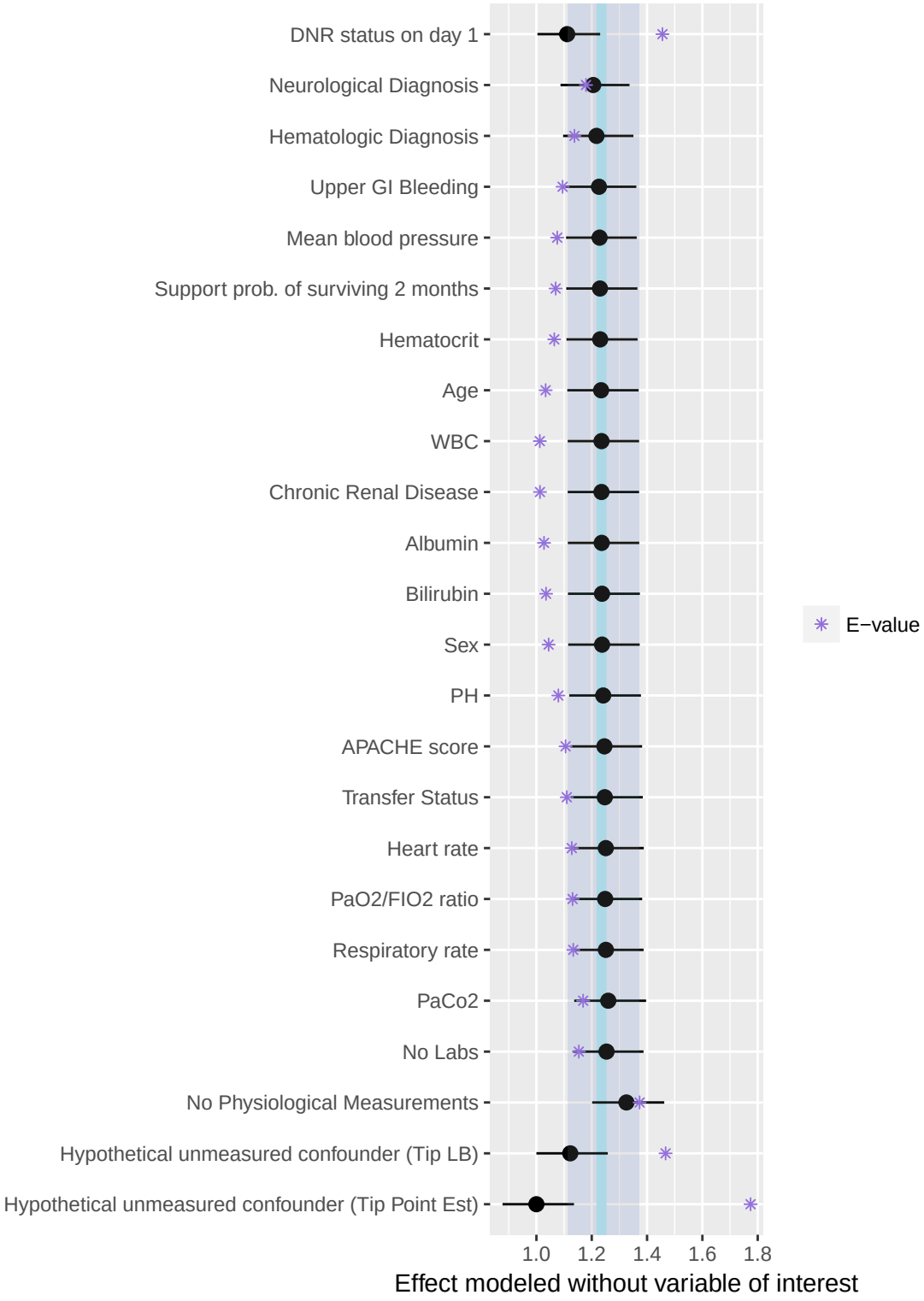

Figure 4.3: Observed bias plot. This displays how the hazard ratio and 95% confidence interfal of RHC on 30 day survival changes if each covariate were unobserved. The solid blue line is the hazard ratio for RHC in the full model (1.24). The blue shaded regin is the 95% confidence interval for the association between RHC and 30 day survival in the full model (1.11, 1.37).

the Love plot (Figure 4.1) and the observed bias plot (Figure 4.3). The Love plot demonstrates which covariates are most imbalanced prior to the propensity score adjustment (here the top 3 are APACHE Score, mean blood pressure, and PaO2/FiO2 ratio), and how balanced they are post-propensity score analysis. Although many covariates have large imbalance when unadjusted, after the propensity score adjustment, they all achieve acceptable balance, certainly below the 0.1 rule of thumb. The observed bias plot shows an important additional piece of information. For example, despite the large imbalance in these top three covariates, leaving them out of the entire process, as if they were unobserved, does not shift the overall observed exposure-treatment effect much. Table 4.1 sheds some light on partially why this is; while these covariates are strongly associated with the exposure, they are less so with the outcome. Additionally, their independence from the remaining covariates effects the observed bias plot. If, for example, APACHE score is highly correlated with the observed covariates, not including it in the study will not have as large of an impact. This careful study of the observed covariates will lend itself to a careful sensitivity analysis. Additionally, we can glean information about how sensitive our analysis was, for example what if we hadn't observed DNR status – it seems this would have almost tipped our analysis to cross 1, rendering it inconclusive.

4.3.1 Scenario 1

In this first scenario, we are concerned that there may be an unmeasured confounder with imbalance similar to that of "APACHE Score", as evidence by the magnitude of the standardized mean difference in Figure 4.1. Using the formula in Equation (4.8), we can calculate the size of the unmeasured confounder needed to tip an analysis with a mean difference the size of that observed for APACHE Score. Notice that Equation (4.8) relies on the unmeasured confounder to have a variance of 1 – in order to match this assumption using the observed covariate, we will scale the mean by dividing by it's standard deviation. The scaled mean APACHE score among the exposed is 3 and the scaled APACHE score among the unexposed is 2.71. We can calculate the size of an unmeasured confounder needed to tip this analysis given a mean difference of 0.29.

```
library(tipr)
tip_with_continuous(mean_diff = 0.29,
                    lb = 1.11,
                    ub = 1.37)
```

```
## [1] 1.433132
```

This allows us to state the following:

Examining one of our most imbalanced covariates, APACHE score, we observe a scaled mean difference of 0.29. A hypothetical unobserved binary confounder that has a mean difference of 0.29 between exposure groups would need to have an association with 30 day survival (HR) of 1.43 to tip this analysis at the 5% level, rendering it inconclusive.

4.3.2 Scenario 2

In this scenario, we have a covariate that is highly associated with the outcome and are interested in calculating how imbalanced an unmeasured confounder of the same magnitude would need to be in order to tip the analysis. Using the same example as above, we choose DNR status as our observed covariate of interest, since it has a large association with the outcome. Examining Table 4.1, one covariate has a larger association with our outcome of interest, the SUPPORT probability of surviving 2 months (which is to be expected since the outcome of interest is 30 day survival). This makes this covariate a poor choice for a sensitivity analysis, as it is unlikely that there is another covariate like it missing from the study.

In our outcome model, DNR status has an adjusted association with 30 day survival (HR) of 2.59 (Table 4.1). Since this is a binary confounder, the prevalence in the unexposed population is necessary in order to calculate the prevalence in the exposed population needed to tip this analysis. The prevalence of DNR status in the unexposed population is 0.14. Using Equation (4.4), we can calculate the how prevalent an unmeasured confounder with these specifications would need to be in the exposed population to tip this analysis.

```
tip_with_binary(p0 = 0.14,
                gamma = 2.59,
                lb = 1.11,
                ub = 1.37)
```

```
## [1] 0.2245824
```

This allows us to state the following:

Examining one of our covariates most highly associated with 30 day survival, we

observe an adjusted association of 2.59. A hypothetical unobserved binary confounder that is prevalent in 14% of the unexposed population with an association of 2.55 with 30 day survival would need to be prevalent in 22.46% of the exposed population to tip this analysis at the 5% level, rendering it inconclusive.

4.3.3 Scenario 3

In our third scenario, rather than being concerned with a single unmeasured confounder of the magnitude of one that was observed, we are interested in the effect of *multiple* independent unmeasured confounders. Using the RHC example, many of the physiological measurements, (bilirubin, hematocrit, white blood cell count, mean blood pressure, PaO2/FiO2 ratio, albumin, respiratory rate, PaCO2, heart rate), resulted in very small associations with the outcome, all less than 1.05 (or the inverse, 0.95). How many unmeasured confounders of this magnitude would it take to tip our analysis to render it inconclusive? In order to calculate this quantity for a continuous unmeasured confounder, we will need to specify the mean difference between the exposed and unexposed groups. Examining Figure 4.1, we can again choose the observed confounder that is the most imbalanced to ground this analysis, for example for mean blood pressure. The scaled mean difference in mean blood pressure between exposure groups is -0.19. This means that this is more prevalent in the unexposed group than the exposed, therefore the association between the hypothetical unmeasured confounder and outcome would need to be < 1 in order to tip this analysis. We will set it at 0.95. Alternatively, we will get the same answer if we flip the mean difference, and use 1/0.95, resulting in a mean difference of 0.19 and an association between the unmeasured confounder and outcome of 1.05.

```
tip_with_continuous(mean_diff = -0.19,
                    gamma = 1/1.05,
                    lb = 1.11,
                    ub = 1.37)
```

```
## [1] 11.25766
```

```
tip_with_continuous(mean_diff = 0.19,
                    gamma = 1.05,
                    lb = 1.11,
                    ub = 1.37)
```

```
## [1] 11.25766
```

It would take at least 11 more independent unmeasured confounders with a scaled mean difference between exposure groups of 0.19 to and an association with 30 day survival (HR) of 1.05 tip the observed analysis at the 5% level, rendering it inconclusive.

4.3.4 Scenario 4

Given the observed lower bound of 1.11, the associated E-value is 1.46. Examining Figure 4.3, we can add some context to this value. The only associated E-value close to this is that for DNR status on day 1. This implies that we would need to be missing an additional independent covariate akin to DNR status on day 1 in order to tip our analysis. Even dropping all physiological measurements would not reach an E-value great enough to tip this study to inconclusive.

4.4 Discussion

With the inclusion of quantified sensitivity to unmeasured confounding analyses being estimated at around 4%, the need for approaches that can gain traction is high. The goal of this paper is to encourage researchers to evaluate the potential impact of unmeasured confounders, using the straightforward methods we have presented here. It is our hope that the guidance and tools provided here ultimately lead to greater utilization of many of the methods available.

We want to emphasize that there has been extensive research in this area; please see additional background and references in Appendix A. The main method we build on was put forth by Lin et al. (Lin, Psaty, and Kronmal 1998). These derivations result in the same equations as Schlesselman (Schlesselman 1978) (Equation (4.1)). Setting LB_{adj} to 1, as we do, creates a rearranged version of Cornfield's original equation (Cornfield et al. 1959). These methods are also related to recent advancements the literature by Ding and VanderWeele (Ding and VanderWeele 2016; VanderWeele and Ding 2017). Ding and VanderWeele put forth a proposal for a sensitivity analysis without assumptions, which allows the researcher to set only two parameters, the relationship between the unmeasured confounder and the exposure (RR_{EU}) and the relationship between the unmeasured confounder and the outcome (RR_{UD}). This setting appears to be the same as our proposed Equation (4.3) with p_1 set to 1 and p_0

as $1/RR_{EU}$. They extend this idea to calculate an E-value, the minimum strength of association that an unmeasured confounder would need to have with both the exposure, RR_{EU}, and the outcome, RR_{UD}, to fully explain away the observed exposure-outcome association. The E-value method adds simplification, in that no sensitivity parameters need to be specified, however it may not generalize well. We contend that this may result in an ambiguous number, as it is not intrinsically grounded in the observed covariates and does not take into account plausible associations. For example this bounding factor, or E-value, may be unnecessarily conservative in many settings where a prevalence of 1 is not plausible. We therefore have updated these methods to allow them to be grounded in the observed covariates, via the adjusted E-value and observed bias plot.

This paper is most useful for the researcher who is concerned about the presence of an unmeasured confounder, but does not know the relationship of this confounder with the exposure and outcome, as well as the uncertainty involved. If all quantities were known, one could backwards engineer an unmeasured confounder that has specified prevalences in each exposure and a given association with the outcome while not changing your existing dataset's outcome, exposure, and covariates. This simulation would answer a slightly different question, which is how would the confidence intervals shift, probabilistically, if the unmeasured confounder were measured with some uncertainty. This is slightly different from our analysis, which is testing what if the effect of the unmeasured confounder was perfectly known and adjusted for. The latter allows for a simpler description of its impact.

Another frequent question that these methods evoke is "Can I do this for negative study results?". Theoretically yes, these methods could be used for negative study results, however we discourage it because it is unlikely to be illuminating. Consider the following three scenarios. 1) The setting most researchers have in mind is having a moderately wide confidence interval that just barely includes the null. Here a quantified sensitivity analysis would show that a fairly weak unmeasured confounder could shift the interval to exclude the null. So the negative result could easily be the result of an unmeasured confounder. However, a lack of statistical power is an equally, if not more, compelling argument. The sensitivity analysis seems unnecessary. 2) The confidence interval is centered on the null and is very wide. While it would take a strong unmeasured confounder to shift the interval enough to exclude the null, this is a reflection of the study's imprecision (the wide interval) not a reflection of the study's negative result being robust to unmeasured confounding. 3) The confidence interval is

centered on the null and is very narrow. While a weak unmeasured confounder could easily shift the interval to exclude the null, the plausible effect sizes would still be clinically meaningless. The sensitivity to unmeasured confounding analysis would not make a case that a clinically meaningful effect was potentially missed. That said, it is possible to estimate the strength of a confounder needed to shift the interval to exclude all clinically meaningless values, which could be informative.

4.5 Conclusion

This paper presents a useful, easily implemented, and intuitively understood approach to allow researchers to assess the potential impact of unmeasured confounders in observational research. The method can be applied to both past and future research, allowing readers to understand the sensitivity of studies that do not include such an analysis and allowing researchers to readily include such an analysis.

4.6 Appendix A. History of unmeasured confounding literature

In 1959, it was well known that there existed an association between smoking and lung cancer, but debate raged as to whether that was a causal relationship. Cornfield et al. engaged in a discussion about the association between smoking and lung cancer (Cornfield et al. 1959). They derived the association between smoking and lung cancer in the event that this association was due solely to a binary unmeasured confounder. In this capacity, Cornfield quantified the prevalence of a binary unmeasured confounder in the exposed and unexposed population that would be necessary to fully nullify the observed association between smoking and lung cancer. Cornfield demonstrated "if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers." (Cornfield et al. 1959)

In 1966, Bross coined the "Size Rule" (Bross 1966). Similar to Cornfield et al., Bross described the impact of a single unmeasured confounder on a given unadjusted effect by estimating what the relative risk of the exposure effect would be if there was really no exposure effect.

> "The Size Rule makes it plain that the counterhypothesis is incompatible with the facts concerning cigarette-cancer risks. Hence at this stage there are two choices: Cease to assert the counterhypothesis and continue to be a scientist, or continue to assert the counterhypothesis and cease to be a scientist. In either case there is no longer a scientific controversy." (Bross 1966)

In 1978, Schlesselman allowed the association between the exposure and outcome to vary (Schlesselman 1978). In 1983, Rosenbaum and Rubin moved the conversation forward by allowing categorical covariate adjustment for the exposure-outcome effect (Rosenbaum and Rubin 1983). In 1998, Lin Psaty, and Kronmal generalized the advancement of Rosenbaum and Rubin by framing the sensitivity analysis within a regression framework (Lin, Psaty, and Kronmal 1998). To this end, they demonstrated that in the case of a binary outcome (analyzed using logistic regression) and a censored time-to-event outcome (analyzed using a proportional hazards model), the "true" odds ratio or hazard ratio of an exposure can be estimated in the same manner that Schlesslman suggests. Using the Lin et al. method, the `R` (R Core Team 2017) package `obsSens` (Snow 2013) generates a tabular analysis with options for different outcome and confounder types.

This paper focuses on Lin, Psaty, and Kronmal's method, however several important advancements in this field have been made since then. Robins, Rotnitzky, and Scharfstein describe an approach that models the association of a counterfactual outcome with an exposure of interest within levels of the measured confounders (Robins, Rotnitzky, and Scharfstein 2000). Using this approach, the analyst no longer has to specify the type of unmeasured confounder (ie: whether it is discrete or continuous, whether there is a single confounder or multiple confounders, etc). They also briefly discuss sensitivity analyses in a Bayesian framework. Brumback et al. describe sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures (Brumback et al. 2004). This approach builds on that of Robins, Rotnitzky, and Scharfsein, essentially building a non-identifiable model that quantifies unmeasured confounding in terms of a sensitivity parameter and a user-specified function.

Greenland describes a Bayesian approach to sensitivity to unmeasured confounders analyses using Monte Carlo risk assessment (Greenland 2001). Greenland explains that the common method for approaching a sensitivity analysis, treating the unmeasured confounders as fixed values as if they are known, does not formally incorporate the

uncertainty about the sensitivity parameters and can be sensitive to the specification of the unmeasured confounder (Greenland 1998). He demonstrates that under certain circumstances, the output from a Monte Carlo risk adjustment with priors for the sensitivity parameters can approximate the posterior that would be obtained from a Bayesian analysis. Greenland further describes choosing priors for bias parameters and demonstrates how even in the case of a relatively low prior probability that an unmeasured confounder explains the association between an exposure and outcome, introducing unmeasured confounders in this manner can considerably increase the uncertainty of a causal relationship (Greenland 2003). These results are further summarized and described as Monte Carlo sensitivity analyses (MCSA) by Greenland (Greenland 2005). Similarly, McCandless, Gustafson, and Levy describe Bayesian sensitivity analyses for unmeasured confounding using MCMC (McCandless, Gustafson, and Levy 2007). They build on methods put forth by Lin, Psaty, and Kronmal (Lin, Psaty, and Kronmal 1998) to build Bayesian models with prior distributions used for the sensitivity analyses that approximate the sampling distribution of model parameters in a hypothetical sequence of observational studies. They demonstrate that credible intervals will on average have approximately nominal coverage probability under these circumstances. The authors further show that sensitivity analyses using information about measured confounders can improve the determination of the uncertainty of unmeasured confounders (McCandless, Gustafson, and Levy 2008). They assert that if the confounding effect of the unmeasured confounder is similar to that of the measured confounders, the Bayesian Sensitivity Analysis may give results that overstate the uncertainty about bias.

Sturmer et al. describe a method that utilizes propensity scores and regression calibration when a validated data set is available (Stürmer et al. 2005). This results in propensity score calibration to adjust for unmeasured confounding in cohort studies. They suggest estimating the propensity of the exposure as one normally would in the main study, then estimating the propensity score in a validation study twice - initially specified the same way the model in the main study is specified, then specified with additional covariates only available in the validation study. The propensity score for the main study is then calibrated using the two propensity score models in the validation study.

Schneeweiss recommends an array-based method that conducts sensitivity analyses on an array of parameters. This can then be visualized on a three dimensional plane with each parameter varied on an axis (Schneeweiss 2006). Schneeweiss also describes

a "Rule-out" method, in which one finds all combinations of the association between the unmeasured confounder and the outcome and the unmeasured confounder and the exposure that would move the point estimate to 1. The rule-out method is similar to the tipping point approach advocated here, with the latter focusing on the confidence bound closest to the null rather than the point estimate such robustness is influenced by the study's observed effect size and the estimate's precision.

VanderWeele has written extensively in this area with his coauthors (VanderWeele, Hernán, and Robins 2008; VanderWeele 2008b; VanderWeele 2008a; VanderWeele and Arah 2011; VanderWeele, Mukherjee, and Chen 2012; VanderWeele 2013; VanderWeele and Ding 2017; Ding and VanderWeele 2016). These important contributions include bringing sensitivity methods into the causal inference framework, extending them to mediation analyses, eliminating the assumptions regarding unmeasured confounders previously needed to create bounding inequalities, and proposing the E-value.

CHAPTER 5

CONCLUSION

In Chapter 2 we replicate the simulations set up by Freedman and Berk, and refute the broad claim that "weighting is likely to increase random error by a substantial amount". In particular, we recommend the more stable propensity score weights, the ATO and ATM weights. We hope that these results will encourage researchers to consider propensity score weighting for bias reduction.

In Chapter 3, we derive the doubly robust estimator for the ATO estimand, as well as the large-sample variance for both the ATO estimator and the doubly robust ATO estimator. Our Monte Carlo simulation comparing the large-sample variance for the doubly robust estimator to two other variance estimation techniques reveals that under our settings our doubly robust estimator and large-sample variance perform relatively well as long as at least one of the two models is correctly specified. Similarly, it seems that incorporating the propensity score estimation in the variance does generally improve the coverage properties when compared to the "naive robust standard errors" when the propensity score model is correctly specified, but the outcome model is incorrectly specified. When both models are correct, or the propensity score model is incorrectly specified but the outcome model is correctly specified, incorporating the propensity score estimation in the variance performs similarly to the naive model with robust standard errors in the continuous case, and slightly outperforms the naive model with robust standard errors in terms of coverage the binary case, due to a decrease in bias. Based on these results, we would recommend using the large-sample variance for the doubly robust estimator when intending to incorporate both a propensity score and outcome model in the estimation process.

In Chapter 4, we present an intuitive approach to allow researchers to assess the potential impact of unmeasured confounders in observational research. We extend the work of Rosenbaum and Rubin (1983) and Lin, Psaty, and Kronmal (1998) to create a contextualized "tipping point" analysis, as well as extend the Vanderweele and Ding (2017) E-value to be grounded in the observed covariates. In addition we provide guidance on best practices in assessing the impact of measured confounders, using tools such as Love plots and observed bias plots. Finally, we provide a number

of scenarios as well as an R package, `tipr`, to illustrate how these methods can be applied in practical settings.

Taken as a whole, these chapters provide guidance for a holistic causal inference process, from the choice of weighting scheme, to estimating the causal estimand and variance, to conducting a sensible sensitivity analysis. We provide both technical detail and derivation as well as applied examples and simulations to fully ground the reader in the ideas presented.

REFERENCES

Austin, Peter C. 2009. "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." *Statistics in Medicine* 28 (25): 3083–3107.

Austin, Peter C, and Elizabeth A Stuart. 2015. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies." *Statistics in Medicine* 34 (28): 3661–79.

Bang, Heejung, and James M Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73.

Belle, G. van. 2011. *Statistical Rules of Thumb*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com/books?id=UaEL-HI6v2YC.

Bross, IDJ. 1966. "Spurious effects from an extraneous variable." *Journal of Chronic Diseases*.

Brumback, Babette A, Miguel A Hernán, Sebastien J P A Haneuse, and James M Robins. 2004. "Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures." *Statistics in Medicine* 23 (5): 749–67.

Busso, Matias, John DiNardo, and Justin McCrary. 2009. "Finite sample properties of semiparametric estimators of average treatment effects." *Forthcoming in the Journal of Business and Economic Statistics*.

———. 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *Review of Economics and Statistics* 96 (5): 885–97.

Connors, A F, T Speroff, N V Dawson, and C Thomas. 1996. "The effectiveness of right heart catheterization in the initial care of critically III patients." *Jama*.

Cornfield, J, W Haenszel, E C Hammond, A M Lilienfeld, M B Shimkin, and E L Wynder. 1959. "Smoking and lung cancer: recent evidence and a discussion of some questions." *Journal of the National Cancer Institute* 22 (1): 173–203.

Crump, R K, V J Hotz, G W Imbens, and O A Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika* 96 (1): 187–99.

Ding, Peng, and Tyler J VanderWeele. 2016. "Sensitivity Analysis Without Assump-

tions." *Epidemiology (Cambridge, Mass.)* 27 (3): 368–77.

D'Agostino, R B. 1998. "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." *Statistics in Medicine* 17 (19): 2265–81.

Freedman, David A, and Richard A Berk. 2008. "Weighting regressions by propensity scores." *Evaluation Review* 32 (4): 392–409.

Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. "Doubly Robust Estimation of Causal Effects." *American Journal of Epidemiology* 173 (7): 761–67.

Greenland, Sander. 1998. "The sensitivity of a sensitivity analysis." *1997 Proceedings of the Biometrics Section American Statistica Association*, 19–21.

———. 2001. "Sensitivity Analysis, Monte Carlo Risk Analysis, and Bayesian Uncertainty Assessment." *Risk Analysis* 21 (4): 579–84.

———. 2003. "The Impact of Prior Distributions for Uncontrolled Confounding and Response Bias." *Journal of the American Statistical Association* 98 (461): 47–54.

———. 2005. "Multiple-bias modelling for analysis of observational data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168 (2): 267–306.

Groenwold, Rolf H H, Anna M M Van Deursen, Arno W Hoes, and Eelko Hak. 2008. "Poor Quality of Reporting Confounding Bias in Observational Intervention Studies: A Systematic Review." *Annals of Epidemiology* 18 (10): 746–51.

Groenwold, Rolf H H, Frank Vries, Anthonius Boer, Wiebe R Pestman, Frans H Rutten, Arno W Hoes, and Olaf H Klungel. 2011. "Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction." *Pharmacoepidemiology and Drug Safety* 20 (11): 1130–7.

Hansen, Ben B, and Mark M Fredrickson. 2014. "Omitted Variable Sensitivity Analysis with the Annotated Love Plot." *Society for Research on Educational Effectiveness*.

Hosman, C A, B B Hansen, and P W Holland. 2010. "The Sensitivity of Linear Regression Coefficients' Confidence Limits to the Omission of a Confounder." *The Annals of Applied Statistics*.

Hsu, Jesse Y, and Dylan S Small. 2013. "Calibrating Sensitivity Analyses to Observed

Covariates in Observational Studies." *Biometrics* 69 (4): 803–11.

Imai, Kosuke, Gary King, and Elizabeth A Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Joffe, Marshall M, Thomas R Ten Have, Harold I Feldman, and Stephen E Kimmel. 2004. "Model Selection, Confounder Control, and Marginal Structural Models." *The American Statistician* 58 (4): 272–79.

Kang, JDY, and J L Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical Science*.

Laan, Mark J van der, and James M Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. New York, NY: Springer New York.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky. 2016. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 80 (December): 0–0.

Li, Liang, and Tom Greene. 2013. "A Weighting Analogue to Pair Matching in Propensity Score Analysis." *The International Journal of Biostatistics*.

Lin, D Y, B M Psaty, and R A Kronmal. 1998. "Assessing the sensitivity of regression results to unmeasured confounders in observational studies." *Biometrics* 54 (3): 948–63.

Lipsitz, Stuart R, Joseph G Ibrahim, and Lue Ping Zhao. 1999. "A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood." *Journal of the American Statistical Association* 94 (448): 1147–60.

Love, T E. 2002. "Displaying covariate balance after adjustment for selection bias." Joint Statistical Meeting: Section on Health Policy Statistics.

Lumley, Thomas. 2011. *Complex Surveys: A Guide to Analysis Using R*. Vol. 565.

John Wiley & Sons.

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 (19): 2937–60.

Mao, Huzhang, and Liang Li. 2018. *PSW: Propensity Score Weighting Methods for Dichotomous Treatments*. https://CRAN.R-project.org/package=PSW.

McCandless, Lawrence C, Paul Gustafson, and Peter C Austin. 2009. "Bayesian propensity score analysis for observational data." *Statistics in Medicine* 28 (1): 94–112.

McCandless, Lawrence C, Paul Gustafson, and Adrian Levy. 2007. "Bayesian sensitivity analysis for unmeasured confounding in observational studies." *Statistics in Medicine* 26 (11): 2331–47.

McCandless, Lawrence C, Paul Gustafson, and Adrian R Levy. 2008. "A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding." *Journal of Clinical Epidemiology* 61 (3): 247–55.

Neugebauer, Romain, and Mark van der Laan. 2005. "Why prefer double robust estimators in causal inference?" *Journal of Statistical Planning and Inference* 129 (1-2): 405–26.

Neyman, Jersey. 1923. "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes." *Roczniki Nauk Rolniczych* 10: 1–51.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robins, James M. 2000. "Robust estimation in sequentially ignorable missing data and causal inference models." In *Proceedings of the American Statistical Association*, 6–10.

Robins, James M, Andrea Rotnitzky, and Mark van der Laan. 2000. "On Profile Likelihood: Comment." *Journal of the American Statistical Association* 95 (450): 477.

Robins, James M, Andrea Rotnitzky, and Daniel O Scharfstein. 2000. "Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models." In *Statistical Models in Epidemiology, the Environment, and Clinical*

Trials, 1–94. New York, NY: Springer New York.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 2012. "Estimation of Regression Coefficients When Some Regressors are not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable." *Statistical Science* 22 (4): 544–59.

Rosenbaum, P R, and D B Rubin. 1983. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society Series B*.

Rosenbaum, Paul R. 2012. "Optimal Matching of an Optimally Chosen Subset in Observational Studies." *Journal of Computational and Graphical Statistics* 21 (1): 57–71.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688.

———. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services and Outcomes Research Methodology* 2 (3-4): 169–88.

Samuels, Lauren Ruth. 2017. "Aspects of Causal Inference within the Evenly Matchable Population: The Average Treatment Effect on the Evenly Matchable Units, Visually Guided Cohort Selection, and Bagged One-to-One Matching." PhD thesis, Vanderbilt University.

Scharfstein, Daniel O, Andrea Rotnitzky, and James M Robins. 1999. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association* 94 (448): 1096.

Schlesselman, J J. 1978. "Assessing effects of confounding variables." *American Journal of Epidemiology* 108 (1): 3–8.

Schneeweiss, Sebastian. 2006. "Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics." *Pharmacoepidemiology and Drug Safety* 15 (5): 291–303.

Snow, Greg. 2013. *ObsSens: Sensitivity Analysis for Observational Studies*. https:

//CRAN.R-project.org/package=obsSens.

Stuart, Elizabeth A, Brian K Lee, and Finbarr P Leacy. 2013. "Prognostic scorebased balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research." *Journal of Clinical Epidemiology* 66 (8): S84–S90.e1.

Stürmer, Til, Sebastian Schneeweiss, Jerry Avorn, and Robert J Glynn. 2005. "Adjusting Effect Estimates for Unmeasured Confounding with Validation Data using Propensity Score Calibration." *American Journal of Epidemiology* 162 (3): 279–89.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, Tyler J. 2008a. "Sensitivity analysis: distributional assumptions and confounding assumptions." *Biometrics*.

———. 2008b. "The Sign of the Bias of Unmeasured Confounding." *Biometrics* 64 (3): 702–6.

———. 2013. "Unmeasured confounding and hazard scales: sensitivity analysis for total, direct, and indirect effects." *European Journal of Epidemiology* 28 (2): 113–17.

VanderWeele, Tyler J, and Onyebuchi A Arah. 2011. "Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders." *Epidemiology* 22 (1): 42–52.

VanderWeele, Tyler J, and Peng Ding. 2017. "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Ann Intern Med*, July.

VanderWeele, Tyler J, Miguel A Hernán, and James M Robins. 2008. "Causal directed acyclic graphs and the direction of unmeasured confounding bias." *Epidemiology (Cambridge, Mass.)* 19 (5): 720–28.

VanderWeele, Tyler J, Bhramar Mukherjee, and Jinbo Chen. 2012. "Sensitivity analysis for interactions under unmeasured confounding." *Statistics in Medicine* 31 (22): 2552–64.

Venables, W N, and B D Ripley. 2002. *Modern Applied Statistics with S*. 4th edition. Statistics and Computing. New York, NY: Springer New York.

Williamson, Elizabeth J, Andrew Forbes, and Ian R White. 2013. "Variance reduction in randomised trials by inverse probability weighting using the propensity score."

Statistics in Medicine 33 (5): 721–37.

Zeileis, Achim. 2004. "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software* 11 (10).

———. 2006. "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software* 16 (9).

Zigler, Corwin M, Krista Watts, Robert W Yeh, Yun Wang, Brent A Coull, and Francesca Dominici. 2013. "Model Feedback in Bayesian Propensity Score Estimation." *Biometrics* 69 (1): 263–73.

Zigler, Corwin Matthew. 2016. "The Central Role of Bayes Theorem for Joint Estimation of Causal Effects and Propensity Scores." *The American Statistician* 70 (1): 47–54.