The Pennsylvania State University
The Graduate School

CAUSAL INFERENCE BY SEMIPARAMETRIC IMPUTATION

A Thesis in
Statistics
by
Doh Yung Kang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2007

The thesis of Doh Yung Kang was reviewed and approved* by the following:

Joseph L. Schafer
Associate professor of Statistics
Thesis Advisor, Chair of Committee

Murali Haran
Assistant professor of Statistics

Steven F. Arnold
Professor of Statistics

Vernon M. Chinchilli
Professor of Biostatistics and Statistics
Chairman, Department of Health Evaluation Sciences

Bruce G. Lindsay
Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Causal effects are comparisons among the outcomes that a study subject would have under different treatment conditions. Because no subject can receive multiple treatments at the same time, causal inference may be regarded as a missing-data problem. Inferences about causal effects are challenging in the presence of confounders, which may distort estimated effects due to their mutual associations with the treatment assignment and with the outcomes. To deal with this problem, we propose a marginal causal model (MCM), a regression that allows average causal effects to vary with respect to one or more variables of interest. We estimate the parameters of the MCM by constructing imputation models and replacing the missing potential outcomes with predicted values. The causal effects are then estimated by solving a set of estimating equations based on the observed and imputed outcomes. To mitigate the biases that may result when the imputation models are misspecified, we augment the imputation models with covariates derived from estimated propensity scores. We apply these methods to data from real and simulated observational studies of the causal effects of dieting on body weight among adolescent girls.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am most grateful and indebted to my thesis advisor, Dr. Joseph Schafer, for the large doses of guidance and encouragement he has shown me during my graduate study. He directed me to explore the world of causal inference and imputation using both Bayesian and frequentist ideas. I am also grateful and indebted to him for his financial support and enlightening discussions on the topics of my research. I thank my other committee members, Dr. Steven F. Arnold, Dr. Murali Haran, and Dr. Vernon Chinchilli, for their encouraging commentary on my work. I thank Dr. Bruce Lindsay for his extremely valuable comments on estimating equation approaches. I thank my wife Deborah for her love and invaluable support during my graduate study. I appreciate prayer support from Missionary Daniel and Deborah Kim. I thank my parents who provided me with all necessary education that I received in my home country. I thank Mother Sarah Barry who has always encouraged me with the word of God. I thank all of my church members for their prayers for me. I thank my Lord Jesus for helping me to study this field. This research was supported by National Institute on Drug Abuse grant P50-DA-10075.

Chapter 1

Introduction

1.1 Goal

The purpose of causal inference is to study the marginal effect of an assignable treatment on a subsequent outcome. Causal effects are defined as comparisons among the responses under different assignments of treatments to the same set of units. Because the outcome can be realized or observed under only one treatment for each unit, causal inference becomes a problem of missing data. A well-designed randomized experiment is a powerful tool for causal inference, because the randomization ensures that the the study units' pretreatment characteristics under the different treatments are, on average, homogeneous. In an observational study, however, the treatment assignment is not under the experimenters' control. In that case, causal inference becomes challenging because the estimated causal effects may be distorted by pretreatment variables that could be related to both the treatment assignment and the outcome. Such variables are known as *confounders*. Furthermore, causal effects may vary across subdomains of the target population defined by other variables. The latter we call *analytic variables*. Our goal is to estimate average causal effects in the population and to assess how they vary with respect to the analytic variables.

Table 1.1. Rubin Causal Model with analytic variables

Unit	x_i	z_i	T_i	$Y_i(0)$	$Y_i(1)$
i=1	x_1^T	z_1^T	*0*	$Y_1(0)$	$Y_1^*(1)$
i=2	x_2^T	z_2^T	*0*	$Y_2(0)$	$Y_2^*(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i=m	x_m^T	z_m^T	*0*	$Y_m(0)$	$Y_m^*(1)$
i=m+1	x_{m+1}^T	z_{m+1}^T	*1*	$Y_{m+1}^*(0)$	$Y_{m+1}(1)$
i=m+2	x_{m+2}^T	z_{m+2}^T	*1*	$Y_{m+2}^*(0)$	$Y_{m+2}(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i=n	x_n^T	z_n^T	*1*	$Y_{in}^*(0)$	$Y_{in}(1)$

1.2 Average causal effect (ACE)

Potential outcomes are the responses that could be realized under different assignments of treatments to a study unit. (Neyman, 1923; Rubin, 1978). The statistical analysis of causal inference based on potential outcomes is called the Rubin Causal Model (RCM) (Holland, 1986). In the RCM, causal effects are typically defined as the differences among two or more potential outcomes on the same study unit.

Consider the simple case of a binary treatment indicator, T_i, which takes the values 1 ("treatment") and 0 ("control") for units $i = 1, ..., n$. In this case, there are two potential outcomes for any subject, $Y_i(1)$ and $Y_i(0)$, which would be realized if the subject received the treatment or control, respectively. Suppose that each subject has additional covariates. These covariates consist of a vector of *confounders*, x_i, which are considered a nuisance, and a vector of *analytic variables*, z_i, which are of interest and are thought to be related to the causal effects. We will be assuming that all confounding effects have been accounted for, in the sense that the potential outcomes $Y_i(1)$ and $Y_i(0)$ are conditionally independent of the treatment indicator T_i given (x_i, z_i). Confounding effects may change the degree and direction of estimated causal effects, and we will need to adjust our estimated effects to account for the confounders x_i. We will also want to allow causal effects to vary with respect to the analytic variables z_i.

Within the RCM framework, the causal inference problem can be envisioned as in Table 1-1. In Table 1-1, the missing potential outcome for individual i under treatment $T_i = i$ is denoted by $Y_i^*(t)$. The causal effect of the treatment on unit i, defined as $Y_i(1) - Y_i(0)$, is unobservable because one of the two potential outcomes

will be missing. This is often called the "fundamental problem" of causal inference (Holland, 1986). The average causal effect (ACE) for the entire population is defined as

$$ACE = E\left(Y_i(1) - Y_i(0)\right). \tag{1.1}$$

It may also be of interest to define ACE's among those who actually receive the treatment and among those who do not. ACE's for the treated and the untreated will be denoted by

$$ACE_1 = E\left(Y_i(1) - Y_i(0)|T_i = 1\right), \tag{1.2}$$

$$ACE_0 = E\left(Y_i(1) - Y_i(0)|T_i = 0\right). \tag{1.3}$$

More generally, we may allow the average causal effect to vary with respect to the analytic variables, in which case we can make the ACE a function of z_i,

$$ACE(z) = E\left(Y_i(1) - Y_i(0) \,|\, z_i = z\right). \tag{1.4}$$

In the same way, we could generalize ACE_0 and ACE_1 to vary with z_i as well.

1.3 Assumptions

If both potential outcomes were observed for each unit, we could, with minimal assumptions, estimate $E\left(Y_i(1)|T_i = 1\right)$ by the mean

$$\left(\textstyle\sum_i t_i\right)^{-1} \textstyle\sum_i t_i y_i(1).$$

In the same way, we could estimate $E\left(Y_i(0)|T_i = 0\right)$ by

$$\left(\textstyle\sum_i \left(1 - t_i\right)\right)^{-1} \textstyle\sum_i \left(1 - t_i\right) y_i(1).$$

Because half of the potential outcomes are missing, however, we need to make assumptions about the mechanism by which the treatment was assigned to the

units. Adapting Rubin's (1978) assumption of unconfoundedness, we will assume

$$Y_i(t) \perp T_i \,|\, v_i, \tag{1.5}$$

where $v_i = (x_i, z_i)$ includes both the confounders and the analytic variables. In other words, we suppose that there are no confounding variables that are not measured and not recorded either in x_i or in z_i. Assumptions of unconfoundedness are standard practice when making causal inferences. In real applications, this assumption is not always reasonable. Nevertheless, it is an important starting point, and even under this assumption, statisticians do not agree on what the best methods are for estimating ACE's.

Unconfoundedness allows us to partition the joint distribution of a potential outcome $Y_i(t)$ and the treatment indicator T_i as

$$P\left(Y_i(t), T_i|v_i\right) = P\left(Y_i(t)|v_i\right) P\left(T_i|v_i\right). \tag{1.6}$$

The quantity $P\left(T_i = 1|v_i\right) \stackrel{def}{=} \pi_i$ is called the *propensity score* (Rosenbaum and Rubin, 1983). The other factor, $P\left(Y_i(t)|v_i\right)$, describes relationships between the potential outcome, the confounders and the analytic variables. In the absence of randomization, estimating $E(Y_i(t))$ or $E(Y_i(t) \,|\, z_i)$ requires assumptions about one or both of these factors. Much of the literature on causal inference (Rosenbaum, 2002) emphasizes the role of the propensity score, making few assumptions about the potential outcomes. With a correctly specified propensity model, it is possible to obtain consistent estimates of causal effects without imposing any model on $Y_i(0)$ or $Y_i(1)$. Methods based on propensity scores include matching and stratification (Rosenbaum and Rubin, 1983, 1984; Rosenbaum, 2002) and inverse-propensity weighting (Hirano and Imbens, 2002; Hahn, 1998).

Alternatively, we can obtain estimates of ACE's by making assumptions about the potential outcomes. Imposing a fully parametric joint model on $Y_i(0)$ and $Y_i(1)$, the unconfounded assumption implies that the missing potential outcomes are missing at random (MAR), and likelihood-based or Bayesian inference are then possible without specifying any model for T_i. Parametric estimates of ACE's will be highly efficient if the modeling assumptions are correct. But because of potentially high rates of missing information, these estimates may be sensitive to

model failure. In addition, a fully parametric approach will require us to make assumptions about the partial correlation between $Y_i(0)$ and $Y_i(1)$ given v_i, and this correlation cannot be estimated from the observed data.

Whether we make assumptions about the potential outcomes or the propensity scores, it is important to build these models well. In applications involving real observational studies, these models will be, at best, only approximately true. Over the last decade, new methods have been developed that combine models for both the propensities and the potential outcomes (Rotnitzky and Robins 1997; Robins, 2000; Robins and Rotnitzky, 2001; Bang and Robins, 2005). Estimates based on dual modeling may have an interesting property known as double robustness, which means that the resulting estimates remain consistent if either of the two models (but not both) is incorrect.

1.4 Motivating example

In this section, we describe a motivating example of an observational study to assess the effects of dieting on body weight. The national concern about overweight and obesity in the United States has made finding effective strategies for weight control a high priority. Sales of dieting aids and supplements now exceed $40 billion annually. A widely shared notion about dieting is that it is not effective and may, in the long run, lead to weight gain. This view was advanced in a book *Dieting Makes You Fat* (Cannon & Einzig, 1983). Indeed, is not uncommon to hear remarks such as, "I went on a diet last year; now I weigh ten pounds more." Rather than comparing the weight of a person before and after dieting, the more relevant causal question is, "How much would the dieter weigh now if she or she had not dieted?" Upon reviewing the literature, Hill (2004) concluded that scientific evidence for the effect of dieting on weight is weak. All we really know at present is that being fat leads people to diet.

Data used in this dissertation were drawn from the National Longitudinal Study of Adolescent Health (Add Health) (Udry, 2003). Add Health is a nationally representative sample of American middle and high school students measuring a broad array of health-related characteristics and behaviors. Using these data, we will estimate the average causal effect of dieting in girls measured at the first wave

(1994-95) on body weight recorded at the second wave (1995-96) one year later.

Dieting in Add Health is determined by a self-report which indicates whether the participant dieted in order to maintain or lose weight within the last seven days. We use this measure of self-reported dieting in the week prior to Wave 1 as a proxy for dieting during the following year. An item measuring intention to diet in the next year, even if it were available, may not necessarily give better information on dieting behavior, because intentions to diet are not always actualized. No definition of dieting was provided; each participant was free to interpret the question in her own way.

For adolescents, changes in body weight are accompanied by growth in height. Following common practice, we express weight in terms of body mass index (BMI), defined as the participant's weight divided by her squared height (kg/m^2). According to guidelines issued by the Centers for Disease Control and Prevention (CDC), an adult is considered to be overweight or obese if his or her BMI exceeds 25.0 or 30.0, respectively. For children and teenagers, BMI varies with the natural growth cycle. A child or teenager is considered to be at risk for overweight or actually overweight if his or her BMI exceeds the 85th or 95th percentile, respectively, of the age and sex-specific BMI distributions published by the CDC.

Participants' decisions to diet are clearly related to BMI at Wave 1, and the strong correlation between BMI at the two waves ($R \approx 0.87$) makes it essential to control for the baseline BMI as a confounder and, at the same time, a very good predictor for the BMI at Wave 2. Other potential confounders that are known to influence dieting behavior include age, race and ethnicity, self-perception of weight relative to peers, self-perceived physical fitness, acceptance by peers and personality characteristics such as self-esteem, self-efficacy and locus of control. Items from Wave 1 related to these characteristics were identified and included in the analysis.

Because the true causal effect of dieting on BMI in this population is unknown, the Add Health data themselves provide little objective basis for evaluating the performance of any method. For evaluation purposes, we used information from the Add Health sample to generate an artificial but realistic population of one million adolescent girls. Variables recorded for this population include body weight, dieting status, and variables related to dieting. Distributions for these variables

and relationships among them closely mimic those in the Add Health sample.

Using artificially simulated data has the following advantages. First, the potential outcomes under dieting and no dieting are present for each individual, so the true causal effects are known. Second, the actual Add Health study has many complications that make a proper analysis difficult. The data were collected by a complex multistage procedure with unequal probabilities of selection at multiple stages. The sample has missing items and dropout between the waves. These complications are found in many observational studies, but attempting to deal with them all in this dissertation would distract attention from the key issues of causal inference that we want to address. The samples that we will draw from the artificial population, on the other hand, will be simple random samples of N=1,000 subjects with no missing data.

When creating the population variables, we intentionally avoided the use of common parametric models (e.g. normal distributions, linear and logistic regressions) that might be assumed by a data analyst. Rather, we chose a combination of generalized additive modeling (Hastie & Tibshirani, 1990) and kernel density estimation (Silverman, 1992). We do, however, apply standard parametric models in our analysis procedures. This adds a touch of realism, because populations encountered in practice always depart in some ways from the assumptions made by analysts. A list of the variables in the population with descriptions and summary statistics is provided in Table 1-2.

The average effect of dieting in this population is small. For the majority of girls who are not overweight, there is no compelling need to control or reduce their weight, and dieting for these girls could be detrimental. Rather than estimating the average causal effect in the whole population, we will focus on the average effect only among those who dieted, which is also very small. Through exploratory analyses of the population data, we found that the causal effects appear to vary with WORKHARD: "When you get what you want, it's usually because you worked hard for it" (1=strongly agree, ..., 5=strongly disagree). This variable is intended to measure locus of control. Dieting is most effective for girls with values near 1 and counterproductive for those with values near 5. In addition to estimating the ACE among those who dieted, we will also be estimating the ACE within classes of WORKHARD.

Figure 1-1 shows boxplots for the actual causal effects of dieting on BMI for all one million girls in the population, classified by the five levels of z_i=WORKHARD. The causal effects of dieting are defined as $Y_i(1) - Y_i(0)$, where $Y_i(1)$ is the value of BMI at Wave 2 if the girl diets, and $Y_i(1)$ is her BMI if she does not. Figure 1-1 also shows the naive estimates of the ACE's in each group based on a standard t-test. These naive estimates are simply the average BMI observed at Wave 2 among those who dieted, minus the average observed BMI at Wave 2 among those who did not, and they estimate

$$E\left(Y_i(1)|T_i = 1, z_i\right) - E\left(Y_i(0)|T_i = 0, z_i\right). \tag{1.7}$$

The large discrepancies between these naive estimates and the true average causal effects are an indication of selection bias, mainly with respect to BMI. Girls who diet are, on average, heavier than girls who do not. They have larger values of BMI at Wave 1 and, because of the strong correlation between values of BMI at the two waves, this confounding effect makes the naive estimates very large and positive.

1.5 Scope of the thesis

In Chapter 2, we review some modern statistical techniques for estimating average causal effects, most of which are based on propensity scores. Chapter 3 introduces the concept of a marginal causal model (MCM) and presents a new method for estimating causal parameters based on semiparametric imputation. This semiparametric imputation method does not require us to specify a correlation between the potential outcomes. By introducing propensity-related covariates into the imputation models, we help to protect ourselves against bias that could arise if the imputation models are incorrect. The estimation methods in Chapter 3 are appropriate for simple random samples. The actual Add Health study, however, uses a complex survey design, so in Chapter 4 we extend the techniques to allow for such designs. In Chapter 5, we present results from applying the techniques of Chapters 3–4 to simple random samples from the artificial population and to the actual data from Add Health. Major conclusions and comments on these findings

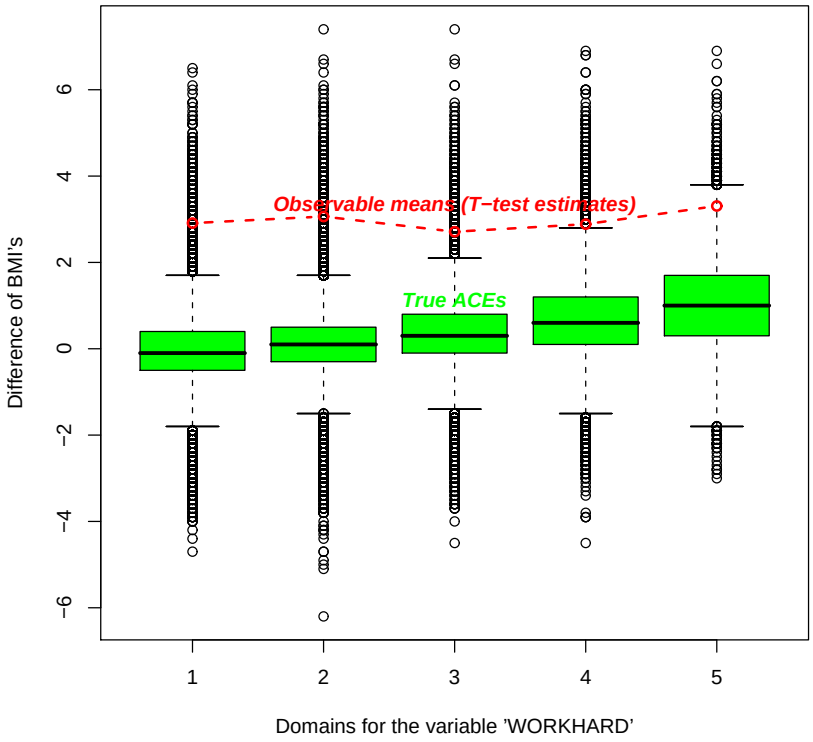

Figure 1.1. Causal effects in the entire population indicated by boxplots, with dotted lines representing t-test estimates from the observable potential outcomes

are given in Chapter 6.

Table 1.2. Variables used in the artificial data

Name	Description	Mean	SD
DIET	1=dieted, 0=did not diet	0.19	0.39
BMI.1	Body mass index at Wave I	22	4.38
BMI.2	Body mass index at Wave II	22.37	4.8
BLACK	1=Black, 0=otherwise	0.23	0.42
NBHISP	1=non-Black Hispanic, 0=otherwise	0.15	0.36
GRADE	Grade in school at Wave I (7, . . . , 11)	9.2	1.39
SLFHLTH	Self-rating of overall health(1=excellent, 2=very good,3=good, 4=fair, 5=poor)	2.18	0.91
SLFWGHT	Self-rating of weight (1=very underweight, 2=slightly under,3=about right, 4=slightly over, 5=very over)	3.3	0.78
WORKHARD	"When you get what you want, it's usually because you workedhard for it" (1=strongly agree, . . . , 5=strongly disagree)	2.14	0.9
GOODQUAL	"You have lots of good qualities"(1=strongly agree, . . . ,5=strongly disagree)	1.8	0.68
PHYSFIT	"You are physically fit" (1=strongly agree,. . . , 5=strongly agree)	2.26	0.93
PROUD	"You have a lot to be proud of" (1=strongly agree, . . . ,5=strongly disagree)	1.77	0.75
LIKESLF	"You like yourself just the way you are "(1=strongly agree, . . . , 5=strongly disagree)	2.16	1
ACCEPTED	"You feel socially accepted" (1=strongly agree, . . . ,5=strongly disagree)	1.97	0.79
FEELLOVD	"You feel loved and wanted" (1=strongly agree, . . . ,5=strongly disagree)	1.73	0.76

Chapter 2

Overview of causal inference through propensity scores and regression

2.1 Confounding effects and regression methods

If the treatment is randomly assigned, inferences about causal effects are straightforward; an average causal effect may be estimated by the difference between the mean outcome among the treated and the mean outcome among the untreated. However, in the case of an observational study where the treatment assignment is beyond the control of the experimenters, inferences about causal effects become challenging, because the confounders, which may be associated with outcomes and/or treatments can distort the degree and direction of causal effects.

A widely used strategy for causal inference in observational studies is to measure as many confounders as possible, and then compare the mean of the treated and the untreated conditionally given the confounders. This basic idea of regression, which underlies classical analysis of covariance and structural equation modeling, has and continues to dominate the social and behavioral sciences.

Over the last two decades, however, another set of methods has been widely accepted in the fields of economics and medical research. These methods are based on the propensity score (Rosenbaum & Rubin, 1983).

2.2 Matching and stratification with propensity scores

The propensity score is the conditional probability that the subject receives a treatment given the observed pretreatment covariates. The propensity score is a concise summary of the subject' pretreatment characteristics. Given the subjects' propensities of receiving the treatment, we can match or stratify their outcomes using the estimated propensity scores in order to control for differences in the distributions of the covariates and reduce selection bias. For example, Rosenbaum and Rubin (1984) showed that stratification based on the quintiles of the propensity scores will reduce bias by at least 90%. Rosenbaum (2002) provides detailed examples of propensity-based matching and stratification.

Using propensity scores for matching and stratification in observational studies has its roots in Fisher's randomization-based inference for the classical design of experiments (Fisher, 1935;1945). In randomization-based inference—i.e. Fisher's hypothetical example of the lady testing tea—the experimenter has complete control over the treatment assignment. Probability enters the experiment only through the assignment by the experimenter. An observational study can be viewed as a kind of 'broken' experiment with an unknown treatment mechanism that lies beyond the experimenter's control. If the treatment mechanism were revealed through actual or estimated propensities, the experiment could be restored so that causal inferences could be drawn.

Relying on estimated propensity scores in observational studies may provide a degree of robustness in drawing causal inferences. It is because propensity-based methods such as matching and stratification do not require a strictly correct model for the potential outcomes. If, however, we ignore the relationships between the potential outcomes and covariates, our estimates of causal effects may have low efficiency.

2.3 Propensity scores in inverse propensity weighting

Propensity scores define the treatment mechanism and summarize the subjects' characteristics. Weighting the responses of the treated individuals by the inverses of their propensities can mimic the distribution of responses that would be seen if the treatment were applied to the full population. This is what we call the inverse propensity weighting (IPW) method. When the potential outcomes are assumed to be realized values from a very large population, weighting the observed potential outcomes can produce sufficient statistics for estimating the parameters of interest in the population. One of the important features of the infinite population is that the response variables are assumed to be random and to be realized. For example, consider $E(Y_i(t))$, the infinite population mean of a single potential outcome under a realized treatment at $T_i = t$. Then it can be shown that, under unconfoundedness,

$$E(Y_i(t)) = E(\pi_i^{-1} T_i Y_i(t)), \tag{2.1}$$

where π_i is the known propensity score. Thus $E(Y_i(t))$ can be simply estimated by

$$n^{-1} \sum_i \widehat{\pi}_i^{-1} t_i y_i(t). \tag{2.2}$$

Although we assume that the potential outcome $Y_i(t)$ is a random variable in the infinite population, the distributional assumptions on $Y_i(t)$ are very minimal; the propensity score π_i, however, must be modeled correctly.

The methods of weighting, stratification and matching all require a correct propensity model. However, weighting methods may depend on correct model for the propensity scores more than matching or stratification. It is because weighting methods try to represent an infinite population quantity by inflating (extremely inflating, in many cases) the observed quantity by the inversed propensity weight.

By viewing T_i as a sampling inclusion indicator, the inverse of propensity weighting (IPW) methods are seen to be essentially same as the Horvitz-Thompson

(HT) estimator (1952) in the classical survey literature. In the survey literature, the instability of an HT estimator is sometimes corrected by a method known as 'generalized regression' estimation (Cassel et al, 1977), also called model-assisted survey estimation (Särndel et al 1989, 2003). This method increases the efficiency of the HT estimator by introducing a regression model.

In survey sampling, the inclusion probability is known, and the power of the model assisted estimators come from the regression model. Similar to this method, a class of estimators called doubly robust (DR) estimators were independently developed in the missing-data and causal-inference literatures (Robins, Rotnitzky and Zhao, 1995; Robins and Rotnitzky, 1995; Rotnitzky, Robins and Scharfstein, 1998; Bang and Robins 2005).

DR methods improve upon IPW methods by introducing additional information through a regression model in the same way that the model assisted estimator does. The theoretically distinguishing feature of DR estimates is that they remain asymptotically unbiased if either the propensity model or the regression model is misspecified. Consider a simple DR estimator

$$n^{-1}\sum_i \widehat{\pi}_i^{-1} t_i y_i(t) + n^{-1}\sum_i (1 - \widehat{\pi}_i^{-1} t_i)\widehat{y}_i(t), \tag{2.3}$$

where $\widehat{y}_i(t)$ is a predicted value from a regression of $y_i(t)$ on covariates. This is just an IPW estimator augmented by a regression-based correction, the second term of the DR estimator above. It can be shown that when the propensity model $\widehat{\pi}_i$ is correctly specified, the second term has an expectation of zero. And when the regression prediction $\widehat{y}_i(t)$ is correct, $\widehat{\pi}_i^{-1} t_i$ will vanish on average in large samples, making the DR estimator asymptotically unbiased.

However, Kang and Schafer (2006) show that the DR methods of Robins et al. are not the only way to achieve double robustness. Kang and Schafer (2006) also show that, in certain circumstances, even DR methods by Robins et al. may not significantly improve upon a simple regression method. Unless the propensity score is very stable, bounded well away from zero and one, the IPW method and IPW related DR methods may not effectively estimate average causal effects.

2.4 Regression methods with propensity-related covariates

Regression methods view the potential outcomes as random variables from an infinite population and require them to be modeled. Unlike matching, stratification or weighting, regression-based methods rely on the relationships between potential outcomes and covariates rather than relationships between the treatment indicator and covariates.

Modeling the potential outcomes given covariates can be challenging, especially when the covariates are high dimensional. Little and An (2004) proposed regression model that allows the mean of an outcome to vary with the estimated propensities through a spline basis. The underlying rationale of this method is that, because the propensity score is a concise scalar summary of the multidimensional covariates, a non-parametric regression of the outcome on the propensity score would estimate average causal effects without bias. In other words, under unconfoundedness,

$$Y_i(t) \perp T_i | \pi_i. \tag{2.4}$$

Little and An's (2004) estimator of $E(Y_i(t))$ is just $n^{-1}\sum_i \hat{m}(x_i)$ where $\hat{m}(x_i)$ is estimated from the smooth regression of $Y_i(t)$ on π_i among the cases for which $Y_i(t)$ is observed.

Putting propensity related covariates into the regression model can be a good idea because these propensity related covariates can correct the bias due to the failure of the regression model. Kang and Schafer (2006) showed that, in some cases, regression methods with propensity-related covariates can perform better than the DR methods of Robins et al. Adjusting for confounders using both a propensity model and regression model is a powerful idea. However, how we combine the information from these two models is critical. Propensity-related covariates in a regression model for potential outcomes help us to correct for bias arising from the possible misspecification of that regression model.

Chapter 3

Marginal causal model

3.1 Model for the completely observed potential outcomes

3.1.1 ACE's with analytic variables

Suppose we assume that the average potential outcome under a treatment $T_i = t$ at a particular value of the analytic variables z_i is

$$E\left(Y_i(t)|z_i\right) = g\left(z_i^T \phi_{(t)}\right), \tag{3.1}$$

where $g\left(\cdot\right)$ is a known invertible link function. Then the causal parameters of interest expressed as

$$\theta = \phi_{(1)} - \phi_{(0)}, \tag{3.2}$$

so that

$$\begin{aligned} z_i^T \theta &= z_i^T \phi_{(1)} - z_i^T \phi_{(0)} & (3.3) \\ &= g^{-1}\left(E\left(Y_i(1)|z_i\right)\right) - g^{-1}\left(E\left(Y_i(0)|z_i\right)\right). & (3.4) \end{aligned}$$

We set up the problem in this way in order to give θ a causal interpretation under the potential-outcomes framework (Robins 2000). Equation 3.1 will be called the

Marginal Causal Model (MCM). This is a *marginal* model because it describes the marginal distribution of the outcome under treatment t rather than joint distribution of potential outcomes under all treatments. This is a *causal* model, because it describes potential outcomes rather than the observed outcome.

Suppose that for a numeric outcome, we choose the link function $g(\cdot)$ to be the identity link and take $z_i^T = 1$; then θ is simply an average causal effect (ACE) for the entire population,

$$E\left(Y_i(1) - Y_i(0)\right). \tag{3.5}$$

In the case of a binary outcome, $Y_i(t) = 0$ or 1, if $g(\cdot)$ is chosen to be the inverse logistic function $g(a) = e^a/(1+e^a)$, then the elements of θ are log-odds ratios. If $g(a) = e^a$ then the elements of θ are log relative risks.

For example, consider $z_i^T\theta = (1, sex_i)(\theta_1, \theta_2)^T$ where $sex_i = 1$ for male and $sex_i = 0$ for female, and let $g(a) = e^a/(1+e^a)$; then θ_1 will be log-odds ratio for female and θ_2 will be log-odds ratio for male minus the log-odds ratio for female. That is,

$$\begin{aligned}
\theta_1 &= \log\left(\frac{P\left(Y_i(1)=1|sex_i=0\right)}{P\left(Y_i(1)=0|sex_i=0\right)}\right) - \log\left(\frac{P\left(Y_i(0)=1|sex_i=0\right)}{P\left(Y_i(0)=0|sex_i=0\right)}\right) \\
\theta_2 &= \log\left(\frac{P\left(Y_i(1)=1|sex_i=1\right)}{P\left(Y_i(1)=0|sex_i=1\right)}\right) - \log\left(\frac{P\left(Y_i(0)=1|sex_i=1\right)}{P\left(Y_i(0)=0|sex_i=1\right)}\right) \\
&\quad - \left\{\log\left(\frac{P\left(Y_i(1)=1|sex_i=0\right)}{P\left(Y_i(1)=0|sex_i=0\right)}\right) - \log\left(\frac{P\left(Y_i(0)=1|sex_i=0\right)}{P\left(Y_i(0)=0|sex_i=0\right)}\right)\right\}.
\end{aligned}$$

In the same way, if we take $g(a) = e^a$, then $\exp(\theta_1)$ is odds ratio for female and $\exp(\theta_2)$ is odds ratio for male divided by odds ratio for female. If $g(a) = a$, then θ_1 will be the ACE for female and θ_2 will be the ACE for male minus the ACE for female.

3.1.2 Estimation

With complete data, the vectors of estimands $\left(\phi_{(0)}^T, \phi_{(1)}^T\right)$ which determine the causal estimand $\theta = \phi_{(1)} - \phi_{(0)}$ could be estimated by the solution to the estimating equation $[0] \stackrel{set}{=} \sum_{i=1}^n \Psi_i(\phi)$ which has the following property.

Lemma 1. *Suppose that we apply a working assumption of constant variance to* $Y_i(0)$ *and* $Y_i(1)$ *and a working assumption of no covariance between them. Then, if the potential outcomes were all seen, an estimating function for the* $2p \times 1$ *dimensional distinct parameters* $\left(\phi_{(0)}^T, \phi_{(1)}^T\right)^T \stackrel{def}{=} \phi$ *with the identity, logit or log link would be*

$$\begin{aligned}\sum_{i=1}^{n} \Psi_i(\phi) &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \phi} V(\mu_i)^{-1} (y_i - \mu_i) \right) \\ &= \sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \mu_i(0)}{\partial \phi_{(0)}} \upsilon_0^{-1} (y_i(0) - \mu_i(0)) \\ \frac{\partial \mu_i(1)}{\partial \phi_{(1)}} \upsilon_1^{-1} (y_i(1) - \mu_i(1)) \end{pmatrix},\end{aligned}$$

where $y_i = (y_i(0), y_i(1))^T$, $\mu_i = E(Y_i|z_i)$, $\partial \mu_i / \partial \phi = \left(\partial \mu / \partial \phi_{(0)i}, \partial \mu / \partial \phi_{(1)i}\right)^T$ *and* $\upsilon_t = Var(Y_i(t))$. *The solutions to the estimating equation* $\sum_{i=1}^{n} \Psi_i(\phi) \stackrel{set}{=} [0]$, *where* $[0] = (0, ..., 0)^T$ *are consistent estimators for* ϕ.

The proof of this lemma is in the Appendix. Because there are no joint observations of $(Y_i(0), Y_i(1))$, it is impossible to estimate the covariance between $Y_i(0)$ and $Y_i(1)$. However, the estimating equation above suggests that a consistent estimator for ϕ can be obtained regardless of the correlation structure for the two potential outcomes for the identity, logit or log link function. Let $\Psi_i(\phi)$ be two independent estimating equations for $t = 0, 1$,

$$\sum_{i=1}^{n} U_i^{(t)} \left(\phi_{(t)}\right) \stackrel{set}{=} [0], \tag{3.6}$$

where $U_i^{(t)}(\phi_{(t)})$ is an estimating function for $\phi_{(t)}$. Let $\widehat{\phi}_{(t)}^G$ be solution to estimating equation (3.6). We call $\widehat{\theta}^G = \widehat{\phi}_{(1)}^G - \widehat{\phi}_{(0)}^G$ the 'gold-standard' estimator for the following reasons. First, no confounders need to be considered in estimation, because all potential outcomes are completely observed. Second, no assumptions about the treatment mechanism are needed. Thus $\widehat{\theta}^G$, if it could be calculated, would be a $\sqrt{n}-$consistent estimator for the true causal estimand $\theta = \phi_{(1)} - \phi_{(0)}$. Under standard regularity conditions, the asymptotic variance for $\widehat{\theta}^G$ based on the

estimating function $U_i^{(t)}\left(\phi_{(t)}\right)$ is

$$\sum_{t=0,1}\left(H_{\phi_{(t)}}\right)^{-1}E\left(U_i^{(t)}U_i^{(t)T}\right)\left(\left(H_{\phi_{(t)}}\right)^{-1}\right)^T,$$

where $U_i^{(t)}=U_i^{(t)}\left(\phi_{(t)}^G\right)$ and $H_{\phi_{(t)}}=E\left(\partial U_i^{(t)T}/\partial\phi_{(t)}\right)$. As shown by Liang and Zeger (1986), an appropriate estimator for this variance is

$$\sum_{t=0,1}n\left(\widehat{H}_{\phi_{(t)}}\right)^{-1}\sum_i\widehat{U}_i^{(t)}\widehat{U}_i^{(t)T}\left(\left(\widehat{H}_{\phi_{(t)}}\right)^{-1}\right)^T,$$

where $\widehat{U}_i^{(t)}=U_i^{(t)}\left(\widehat{\phi}_{(t)}^G\right)$ and $\widehat{H}_{\phi_{(t)}}=\sum_i\left(\partial U_i^{(t)T}/\partial\phi_{(t)}\right)|_{\phi_{(t)}=\widehat{\phi}_{(t)}^G}$.

3.1.3 Estimation among the treated

Causal treatment effects can be studied in at least three populations: the entire population, the treated group, and the untreated group. Researchers who study a new treatment may like to assess the causal effects for the treated group rather than the entire population or the control group. Suppose that we redefine $\phi_{(0)}$, $\phi_{(1)}$ and θ to be the analogous versions of (3.1)-(3.2) among the treated. Then $\phi_{(t)}$ is such that $E\left(Y_i(t)|T_i=1,z_i\right)=g\left(z_i^T\phi_{(t)}\right)$. The corresponding estimation function becomes

$$\textstyle\sum_{i=1}^n t_iU_i^{(t)}\left(\phi_{(t)}\right)\stackrel{set}{=}[0],\tag{3.7}$$

where $U_i^{(t)}(\phi_{(t)})$ is the same estimating function as before. Solving this equation leads to the 'gold-standard' estimator for the treated.

3.2 Estimation with partially observed potential outcomes using predictive mean imputation

3.2.1 Estimating equations for imputation model

Until now we have supposed that the potential outcomes were all observed. But in reality, they cannot be all seen at the same time. This is the fundamental problem of causal inference (Holland 1986). To solve this problem, we propose a predictive mean imputation approach. Unlike the marginal structural model of Robins, Hernan and Brumback (2000), which is based on inverse-propensity weighting, we replace missing values of $Y_i(0)$ and $Y_i(1)$ by predicted mean values obtained from regressions of the potential outcomes on v_i. Note that v_i includes a vector of confounders x_i and a vector of analytic variables z_i. Once these predictive means fill in the missing outcomes, the average causal estimand can be simply estimated from the combined set of observed potential outcomes and imputed missing potential outcomes.

Suppose that we impute the predictive mean $E\left(Y_i(t)|v_i, T_i = 1-t\right)$ for the missing potential outcome $Y_i(t)$ for $i \in \{i : T_i = 1-t\}$. Then we replace

$$U_i^{(t)}(\phi_{(t)}) = c_i\left(t\right)\left(y_i\left(t\right) - g\left(z_i^T\phi_{(t)}\right)\right) \tag{3.8}$$

by

$$U_i^{(t)*}\left(\phi_{(t)}\right) = c_i\left(t\right)\left(y_i^*\left(t\right) - g\left(z_i^T\phi_{(t)}\right)\right), \tag{3.9}$$

where $y_i^*\left(t\right)$ is $y_i\left(t\right)$ if $t_i = t$ and $E\left(Y_i\left(t\right)|v_i\right)$ if $t_i = 1-t$. Here $c_i\left(t\right) = z_i\left\{\partial g\left(\eta_i\left(t\right)\right)/\partial\eta_i\left(t\right)\right\} var(y_i\left(t\right))^{-1}$ and $\eta_i\left(t\right) = z_i^T\phi\left(t\right)$.

Our estimating function $U_i^{(t)*}\left(\phi_{(t)}\right)$ differs from that of Bang and Robins (2005), who defined

$$U_i^{(t)R}\left(\phi_{(t)}\right) = c_i\left(t\right)\left(y_i\left(t\right) - g\left(z_i^T\phi_{(t)}\right)\right)\frac{I\left(T_i = t\right)}{P\left(T_i = t|v_i^T\right)} + \Lambda_i, \tag{3.10}$$

where $P\left(T_i = 1|v_i^T\right)$ is propensity score (Rosenbaum and Rubin, 1983) and Λ_i is an augmenting function that makes the procedure 'doubly robust' (Robins and

Rotnitzky, 2001). To get a consistent estimator of $\phi_{(t)}$, they must assume that $P\left(T_i = 1|v_i^T\right)$ is bounded away from zero and one (Robins and Rotnitzky, 1997). Their estimating function may be sensitive to the specification of the propensity model. Our estimating function, on the other hand, requires a correctly specified mean structure for the imputation model $E\left(Y_i(t)|v_i, \beta_{(t)}\right)$.

$U_i^{(t)*}\left(\phi_{(t)}\right)$ may be viewed as an estimating function averaged over the predictive distribution of the missing potential outcomes, conditional on observed potential outcomes and covariates. This is similar to the E-step of an EM algorithm (Dempster et al, 1977) in the sense that both methods impute the predictive (expected) mean for missing part. However, there is difference: in case of EM, expectation is applied to the log-likelihood, but in our case, to the estimating function. Also, because $E_Y\left(U_i^{(t)}\left(\phi_{(t)}\right)\right) = E_V\left(E_{Y|V}(U_i^{(t)}\left(\phi_{(t)}\right))\right)$, the expectation of $U_i^{(t)}\left(\phi_{(t)}\right)$ is same as that of $U_i^{(t)*}\left(\phi_{(t)}\right)$. Imputation of predictive means is natural under the unconfounded condition, which makes it possible for $E\left(Y_i(t)|v_i, \beta_{(t)}\right) = E\left(Y_i(t)|v_i, T_i = 1-t, \beta_{(t)}\right)$. (Note that $E\left(Y_i(t)|v_i, T_i = 1-t, \beta_{(t)}\right)$ is the mean of the missing potential outcome of $Y_i(t)$.)

In summary, under the unconfounded condition, we build a model for the mean of the missing potential outcomes based on the observed potential outcomes and the corresponding covariates. Then the missing parts are predicted by $E\left(Y_i(t)|v_i, T_i = 1-t, \widehat{\beta}_{(t)}\right)$. If the mean structure $E\left(Y_i\left(t\right)|v_i, \beta_{(t)}\right)$ is correctly specified, then the unbiased estimating equation $\sum_{i=1}^{n} U_i^*\left(\phi_{(t)}, \beta_{(t)}\right) \stackrel{set}{=} [0]$ provides a $\sqrt{n}-$consistent estimator $\widehat{\phi}_{(t)}^{PMI}$ which is asymptotically equivalent to our gold standard estimator $\widehat{\phi}_{(t)}^{G}$.

3.2.2 Point estimation

3.2.2.1 Point estimation for the entire population

To account for the uncertainty of imputing $E\left(Y_i\left(t\right)|v_i\right)$ for the missing parts, we introduce additional estimating equations for $\beta_{(t)}$, the vector of parameters for the predictive mean of linear regression $Y_i\left(t\right)$ on v_i,

$$E(Y_i(t)|v_i, \beta_{(t)}) = E(Y_i(t)|v_i, T_i = 1-t, \beta_{(t)}) = m\left(v_i^T\beta_{(t)}\right). \tag{3.11}$$

Suppose that $\widehat{\beta}_{(1)}$ and $\widehat{\beta}_{(0)}$ are solutions to a pair of independent estimating equations,

$$\sum_{i=1}^{n}(1-t_i)S_i^{(0)}(\beta_{(0)}) \overset{set}{=} [0] \text{ and } \sum_{i=1}^{n}(t_i)S_i^{(1)}(\beta_{(1)}) \overset{set}{=} [0], \tag{3.12}$$

where

$$S_i^{(t)}(\beta(t)) = c_i^*(t)\left\{y_i(t) - m\left(v_i^T\beta_{(t)}\right)\right\}, \tag{3.13}$$

$c_i^*(t) = v_i\left\{\partial m\left(\eta_i^*(t)\right)/\partial\eta_i^*(t)\right\} var(y_i(t))^{-1}$, and $\eta_i^*(t) = v_i^T\beta_{(t)}$. The simultaneous estimation function for φ are

$$\Psi^*(\varphi) = \sum_{i=1}^{n}\begin{pmatrix}(1-t_i)S_i^{(0)}\left(\beta_{(0)}\right)\\ (t_i)S_i^{(1)}\left(\beta_{(1)}\right)\\ U_i^{(0)*}\left(\phi_{(0)},\beta_{(0)}\right)\\ U_i^{(1)*}\left(\phi_{(1)},\beta_{(1)}\right)\end{pmatrix}, \tag{3.14}$$

where $\varphi = \left(\beta_{(0)},\beta_{(1)},\phi_{(0)},\phi_{(1)}\right)^T$. To get $\sqrt{n}-$consistent estimates for all parameters, we can first estimate $\beta_{(t)}$ because the estimation of $\beta_{(t)}$ is not influenced by estimating $\phi_{(0)}$ and $\phi_{(1)}$. Suppose that $\widehat{\beta}_{(t)}$ is a solution vector to $\sum_{i=1}^{n}S_i^{(t)}\left(\beta_{(t)}\right) \overset{set}{=} [0]$ and thus $\sqrt{n}\left(\widehat{\beta}_{(t)} - \beta_{(t)}\right) = O_p(1)$. Then we can substitute $\widehat{\beta}_{(t)}$ for $\beta_{(t)}$ to get $\widehat{\phi}_{(t)}$ which is the solution to estimating equation $\sum_{i=1}^{n}U_i^{(t)*}\left(\phi_{(t)},\widehat{\beta}_{(t)}\right) \overset{set}{=} [0]$ where

$$U_i^{(t)*}\left(\phi(t),\widehat{\beta}_{(t)}\right) = \left(c_i(t)\left(y_i^*(t) - z_i^T\phi(t)\right)\right). \tag{3.15}$$

Also it is interesting to see that, for estimating ACE, the resultant estimate $\widehat{\phi}(t)$ is equivalent to the likelihood estimator from a multivariate normal distribution on $(Y_i(0), Y_i(1), V_i)$.

Lemma 2. *Suppose that $g(\cdot)$ is the identity link and $z_i = 1$ so that the estimand of interest is ACE. Then the solution $\widehat{\phi}(t)$ to $\Psi^*(\varphi) \overset{set}{=} [0]$ is equivalent to the maximum likelihood estimator for ACE from a multivariate normal distribution on $(Y_i(0), Y_i(1), V_i)$ under the MAR (missing at random) assumption.*

This lemma suggests that there exists a coherence between the estimating

equation-based imputation approach and observed-data likelihood approach in the case of ACE. If the parametric model is correct, the likelihood estimator will be more efficient. But it is often the case that a multivariate parametric model for $(Y_i(0),\ Y_i(1),\ V_i)$ is difficult to specify. Once we build a correct imputation model and analytic model in a semi-parametric fashion, it is straightforward to get $\widehat{\theta}_{PMI} = \widehat{\phi}_{(1)} - \widehat{\phi}_{(0)}$ which acts like $\widehat{\theta}^G$, our 'gold-standard' estimator for the true causal estimand θ.

3.2.2.2 Point estimation for the the treated group

As we discussed in Section 3.1.3, estimating the causal effects among the treated may be of particular interest. In this case, the estimating equations for the treated group have completely observed potential outcomes while the control group has the missing potential outcomes. Therefore, we need only one imputation model for the control group ($T_i = 0$). The modified estimating function are then

$$\Psi_i(\varphi) = \sum_{i=1}^{n} \begin{pmatrix} (1-t_i) S_i^{(0)}\left(\beta_{(0)}\right) \\ (t_i)\, U_i^{(0)*}\left(\phi_{(0)}, \beta_{(0)}\right) \\ (t_i)\, U_i^{(1)}\left(\phi_{(1)}\right) \end{pmatrix}. \tag{3.16}$$

Solving this equation leads to the consistent estimator for the causal estimand $\theta = \phi_{(1)} - \phi_{(0)}$ as defined in Section 2.3.

3.2.3 Variance estimation

3.2.3.1 Variance estimation for $\widehat{\theta}_{PMI}$ in the entire population

Under standard regularity conditions, the multivariate limiting distribution for the estimator based on simultaneous estimating functions is

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_{(0)} - \beta_{(0)} \\ \widehat{\beta}_{(1)} - \beta_{(1)} \\ \widehat{\phi}_{(0)} - \phi_{(0)} \\ \widehat{\phi}_{(1)} - \phi_{(1)} \end{pmatrix} \longrightarrow N(0, \Sigma), \tag{3.17}$$

where Σ is the general sandwich variance

$$E\left(\partial\Psi_i^*/\partial\varphi\right)^{-1} E(\Psi_i^*\Psi_i^{*T}) \left(E\left(\partial\Psi_i^{*T}/\partial\varphi\right)^{-1}\right)^T$$

for the estimand $\varphi = \left(\beta_{(0)}, \beta_{(1)}, \phi_{(0)}, \phi_{(1)}\right)$. Σ can be estimated by

$$\widehat{\Sigma} = n\left(\widehat{H}\right)^{-1} \sum_{i=1}^{n} (\widehat{\Psi}_i^* \widehat{\Psi}_i^{*T}) \left(\widehat{H}\right)^T, \tag{3.18}$$

where $\widehat{H}$ is a 4×4 partitioned matrix $\sum_{i=1}^{n} \left(\partial\Psi_i^*/\partial\varphi|_{\varphi=\widehat{\varphi}}\right)$. Here $\widehat{\Psi}_i^*$ denotes the estimating function Ψ_i^* which replaces the parameters φ by the estimate $\widehat{\varphi}$. By standard matrix algebra, it can be shown that the marginal asymptotic distribution for the causal estimand is

$$\sqrt{n}\left(\widehat{\phi}_{(t)} - \phi_{(t)}\right) \longrightarrow N(0, V(\phi_{(t)}) + V(\phi_{(t)}, \beta_{(t)})), \tag{3.19}$$

where $V(\phi_{(t)})$ and $V(\phi_{(t)}, \beta_{(t)})$ are the variance components which are the functions of $\phi_{(t)}$ and $(\phi_{(t)}, \beta_{(t)})$ respectively. The formulas for these components are given in the Appendix, and $U_i^{*(t)} = U_i^{*(t)}\left(\phi_{(t)}, \beta_{(t)}\right)$,

$$var\left(\widehat{\beta}_{(t)}\right) = H_{\beta_{(t)}}^{-1} E\left(S_i\left(\beta_{(t)}\right) S_i\left(\beta_{(t)}\right)^T\right) \left(H_{\beta_{(t)}}^{-1}\right)^T. \tag{3.20}$$

We can see that $V(\phi_{(t)}\ )$ is the variance component for $\phi_{(t)}$ when $\beta_{(t)}$ is known. Because $\beta_{(t)}$ needs to be estimated, the variance component $V\left(\phi_{(t)}, \beta_{(t)}\right)$ conveys the uncertainty of estimating these nuisance parameters. This leads to the marginal asymptotic distribution for the causal estimand θ,

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \longrightarrow N(0, \Sigma_\theta), \tag{3.21}$$

where

$$\begin{aligned} \Sigma_\theta &= var\left(\widehat{\phi}_{(1)} - \widehat{\phi}_{(0)}\right) \\ &= var\left(\widehat{\phi}_{(1)}\right) + var\left(\widehat{\phi}_{(0)}\right) - 2cov\left(\widehat{\phi}_{(0)}, \widehat{\phi}_{(1)}\right) \\ &= \sum_{t=0,1} \left(V(\phi_{(t)}) + V(\phi_{(t)}, \beta_{(t)})\right). \end{aligned} \tag{3.22}$$

3.2.3.2 Variance estimation for $\widehat{\theta}^{\tau}_{PMI}$ in the treated group

Applying the procedure of simultaneous estimating functions to the treated group, the solution to the estimating equations (3.17) has the asymptotic normal distribution

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_{(0)}-\beta_{(0)}\\ \widehat{\phi}_{(0)}-\phi_{(0)}\\ \widehat{\phi}_{(1)}-\phi_{(1)}\end{pmatrix}\longrightarrow N(0,\Sigma), \tag{3.23}$$

where

$$\Sigma = E\left(\partial\Psi_i/\partial\varphi\right)^{-1}E((\Psi_i)\left(\Psi_i\right)^T)\left(E\left(\partial\Psi_i/\partial\varphi\right)^{-1}\right)^T. \tag{3.24}$$

It can be shown that the causal estimator for the treated has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)\longrightarrow N(0,\Sigma_\theta), \tag{3.25}$$

where $\Sigma_\theta = \sum_{t=0,1}V(\phi_{(t)}) + V(\phi_{(0)},\beta_{(0)})$ as before. In the actual variance computation, the explicit form of each variance component is not needed because the estimated variances for the respective estimators can be calculated by partitioning the variance estimator for Σ with respect to the parameter sub-vectors.

3.3 Methods for augmenting the imputation model by estimated propensity scores

3.3.1 The role of propensities in the imputation model

A regression imputation model for potential outcomes gives predictions for the missing potential outcomes based on the covariates v_i. When the covariate distributions in the treatment and control groups are similar, the estimated causal effects may not be overly sensitive to the imputation model. If the covariate distributions are very different in the two groups, further adjustments may be needed. For this, we incorporate propensity related covariates into the imputation model

to protect against the possible misspecification of this model.

Under the unconfounded condition, the propensity score $\pi_i = P\left(T_i = 1|v_i\right)$ has the property

$$Y_i(t_i) \perp T_i|k\left(\pi_i\right), \tag{3.26}$$

where $k\left(\cdot\right)$ is any 1:1 function. This property implies

$$E(Y_i(t)|T_i, k\left(\pi_i\right)) = E(Y_i(t)|k\left(\pi_i\right)),$$

and hence the average outcome $E\left(Y_i(t)\right)$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i(t)|k\left(\pi_i\right)). \tag{3.27}$$

Functions of the propensity scores can play a role in the imputation model to reduce bias in estimating the mean of a potential outcome. This idea was originally proposed by Little and An (2004) who demonstrated a doubly robust property in regression estimation using propensity related covariates.

3.3.2 Imputation model with functions of the propensity scores

Suppose that the vector of covariates v_i is now enlarged to include a function of propensity scores, along with the nuisance confounders x_i and analytic covariates z_i. The mean structure for the imputation model is now

$$E(Y_i(t)|v_i, T_i = 1 - t; \beta^*_{(t)}) = k\left(e\left(\pi_i\right); \alpha_{(t)}, \beta^{(A)}_{(t)}\right) + m\left(\left(x_i, z_i\right)^T; \beta^{(B)}_{(t)}\right), \tag{3.28}$$

where $k\left(e\left(\pi_i\right); \alpha, \beta^{(A)}_{(\alpha)(t)}\right)$ is a function of the logit propensity scores

$$e_i = e\left(\pi_i\right) = \mathit{logit}\left(P(T_i = 1|v_i, \alpha)\right) \tag{3.29}$$

which depends on a vector of parameters α. In practice, we often choose

$$k\left(e_i;\alpha,\beta_{(t)}^{(A)}\right)=(1,I_{(\alpha)i}^{(1)},...,I_{(\alpha)i}^{(S)})^T\beta_{(t)}^{(A)},$$

where $I_{(\alpha)i}^{(s)}$ is indicator for s^{th} propensity stratum based on the quantiles of e_i's: $I_{(\alpha)i}^{(1)}=1$ if e_i is less than the first quantile and 0 otherwise; $I_{(\alpha)i}^{(2)}=1$ if e_i is bigger than or equal to the first quantile and less than the second quantile, and $I_{(\alpha)i}^{(2)}=0$, otherwise; and so on. Now suppose that we choose $m\left(v_i;\beta_{(t)}^{(B)}\right)=v_i^T\beta_{(t)}^{(B)}$. Then the estimating equation for $\beta_{(t)}^{*}=\left(\beta_{(t)}^{(A)},\beta_{(t)}^{(B)}\right)$ becomes

$$\sum_{i\in\{i:t_i=t\}}S_i(\beta_{(t)}^{*},\alpha)\overset{set}{=}[0]\,, \tag{3.30}$$

where $S_i(\beta_{(t)}^{*},\alpha)=\left\{y_i\left(t\right)-v_{(\alpha)i}^T\beta_{(t)}^{*}\right\}q_i\left(t\right)^{-1}$, $q_i\left(t\right)^{-1}=v_{(\alpha)i}/var(y_i\left(t\right))$ and $v_{(\alpha)i}$ is chosen to be $(1,I_{(\alpha)i}^{(1)},...,I_{(\alpha)i}^{(S)},x_i,z_i)$. A feature of this mean structure is that when $(x_i,z_i)^T$ is not balanced for two groups within a single potential outcome, or $m\left((x_i,z_i)^T;\beta_{(t)}^{(B)}\right)$ is specified not well enough to predict the missing potential outcomes correctly, its bias can be adjusted for by introducing correctly specified function of propensity scores $k\left(e_i;\beta_{(t)}^{(A)}\right)$. However, if $(x_i,z_i)^T$ itself is balanced for the two groups and $m\left((x_i,z_i)^T;\beta_{(t)}^{(B)}\right)$ is correctly specified, then $k\left(\pi_i;\beta_{(t)}^{(A)}\right)$ will become insignificant. By choosing propensity stratum indicators as augmented covariates, the selection bias is expected to be reduced by approximately 90% (Rubin and Rosenbaum, 1984). One can also apply a spline basis such as

$$\beta_1e_i+\beta_2e_i^2+\beta_3\left(e_i-k_1\right)_+^2+\beta_4\left(e_i-k_2\right)_+^2+\beta_5\left(e_i-k_3\right)_+^2,$$

where k_j is the j^{th} quartile for a function of propensity scores e_i and $(e_i-k_1)_+$ means (e_i-k_1) if $e_i\geq k_1$ and zero otherwise. In this case, the regression model for the potential outcomes with respect to the logit propensities is a quadratic regression spline with knots at k_1, k_2 and k_3. The spline function is simply a way to allow the means of the potential outcomes to vary with the propensity scores. If the spline model is correctly specified, it will correct the possible non-linear bias from the chosen linear model $(x_i,z_i)^T\beta_{(t)}^{(B)}$. On the other hand, if $(x_i,z_i)^T\beta_{(t)}^{(B)}$ is

correctly specified, any functions of propensities will be redundant, merely causing the model to be overfitted.

Because we do not know the true propensity scores, we need to estimate them, e.g. by using a score function for the vector of parameters from a logistic regression,

$$A(\alpha) = \sum_{i=1}^{n} \left(A_i(\alpha)\right) = \sum_{i=1}^{n} \left(\left(x_i, z_i\right)\left(t_i - \pi_i\right)\right). \tag{3.31}$$

Putting these parameters together into a simultaneous estimating function, we have

$$\Psi^{+} = \sum_{i=1}^{n} \begin{pmatrix} A_i\left(\alpha\right) \\ (1-t_i)S_i\left(\beta_{(0)}, \alpha\right) \\ (t_i)S_i\left(\beta_{(1)}, \alpha\right) \\ U_i^{*(0)}\left(\phi_{(0)}, \beta_{(0)}, \alpha\right) \\ U_i^{*(1)}\left(\phi_{(1)}, \beta_{(1)}, \alpha\right) \end{pmatrix}. \tag{3.32}$$

In $\Psi^{+} \stackrel{set}{=} [0]$, we solve the estimating equations sequentially. First, we can independently solve $\sum_{i=1}^{n} A_i\left(\alpha\right) \stackrel{set}{=} [0]$ because the estimation of α is not influenced by the rest of the parameters. Suppose $\widehat{\alpha}$ is a consistent solution to $\sum_{i=1}^{n} A_i\left(\alpha\right) \stackrel{set}{=} [0]$. Then $\sqrt{n}\left(\widehat{\alpha} - \alpha\right)$ is $O_p\left(1\right)$ so we can substitute $\widehat{\alpha}$ for α to get $\widehat{\beta}_{(0)}$ and $\widehat{\beta}_{(1)}$, which are the solution to estimating equations $\sum_{i=1}^{n}(1-t_i)S_i\left(\beta_{(0)}, \widehat{\alpha}\right) \stackrel{set}{=} [0]$ and $\sum_{i=1}^{n}(t_i)S_i\left(\beta_{(1)}, \widehat{\alpha}\right) \stackrel{set}{=} [0]$, respectively. The rest of the procedure is same as the case for the estimating equations (3.15). Detailed descriptions of this matrix and asymptotic distribution for its marginalized ACE estimator are in the Appendix.

The estimating equations for the causal estimands among the treated are simpler than Ψ_i^{+},

$$\Psi^{\tau+} = \sum_{i=1}^{n} \begin{pmatrix} A_i\left(\alpha\right) \\ (1-t_i)S_i\left(\beta_{(0)}, \alpha\right) \\ (t_i)U_i^{*(0)}\left(\phi_{(0)}, \beta_{(0)}, \alpha\right) \\ (t_i)U_i^{*(1)}\left(\phi_{(1)}\right) \end{pmatrix}. \tag{3.33}$$

The estimation procedure is same as that for $\Psi^{+} \stackrel{set}{=} [0]$ except that now there is no need to estimate an imputation model for the treated.

3.4 Other methods for combining the estimated propensity scores with imputation-style modeling of the potential outcomes

3.4.1 Stabilization of IPW related methods

The class of IPW estimators in Chapter 2 inflates the observed outcomes by the inverse propensity scores. When the propensity get close to zero or one, these simple IPW estimators become unstable in terms of bias and efficiency. One way to stabilize the estimator is to apply a denominator equal to the sum of the inflation factors,

$$\widehat{\phi}_{(t)}^{IPW2} = \left(\sum_i t_i \pi_i^{-1}\right)^{-1} \sum_i t_i \pi_i^{-1} y_i\left(t_i\right). \tag{3.34}$$

$\widehat{\phi}_{(t)}^{IPW2}$ can be understood as a ratio estimator from survey sampling, and it is an approximately unbiased estimator for $\phi_{(t)}$. In the same way, a stabilized DR estimator is

$$\widehat{\phi}_{(t)}^{IPDR2} = \left(\sum_i t_i \pi_i^{-1}\right)^{-1} \sum_i \left\{ v_i^T \widehat{\beta}_{(t_i)} + t_i \pi_i^{-1} \left(y_i\left(t_i\right) - v_i^T \widehat{\beta}_{(t_i)} \right) \right\}. \tag{3.35}$$

3.4.2 Imputation model with weighted sufficient statistics

In this section, we provide an estimating function for a weighted least square estimator discussed by Kang and Schafer (2006). Suppose $E(Y_i(t_i)|v_i, 1-t_i; \beta_{(t)}^*)$ has the simple form

$$E\left(Y_i(T_i = t)|v_i, T_i = 1-t; \beta_{(t)}^w\right) = v_i^T \beta_{(t)}^w. \tag{3.36}$$

In this linear model setting, we can also establish a doubly robust estimator for coefficients $\beta_{(t)}^w$. Kang and Schafer (2006) showed that the estimator

$$\widehat{\beta}_{(t)}^w = \left(\sum_{i=1}^n t_i \widehat{\pi}_i^{-1} v_i v_i^T \right)^{-1} \left(\sum_{i=1}^n t_i \widehat{\pi}_i^{-1} v_i y(t_i) \right) \tag{3.37}$$

is a doubly robust estimator for $\beta^w_{(t)}$. The proof of this result is shown in Appendix. In Equation (3.39), we can see that each sufficient statistic is weighted by the estimated inverse propensity score. Ones can show that $\widehat{\beta}^w_{(t)}$ is the solution to the estimating equations

$$\begin{aligned} S(\beta^w_{(t)}, \alpha) &= \sum_{i \in \{i: t_i = t\}} S_i(\beta^w_{(t)}, \alpha) \\ &= \sum_{i \in \{i: t_i = t\}} \left(\frac{v_i^* \left(y_i^*(t) - v_i^{*T} \beta^w_{(t)} \right)}{var(y_i^*(t))} \right), \end{aligned} \tag{3.38}$$

where $v_i^* = \widehat{\pi}_i^{-1/2} v_i$, $y_i^* = \widehat{\pi}_i^{-1/2} y_i$. If we put these parameters together into one set of simultaneous estimating equations,

$$\Psi_i^w = \sum_{i=1}^{n} \begin{pmatrix} A_i(\alpha) \\ (1 - t_i) S_i \left(\beta^w_{(0)}, \alpha \right) \\ (t_i) S_i \left(\beta^w_{(1)}, \alpha \right) \\ U_i \left(\phi^w_{(0)}, \beta^w_{(1)}, \alpha \right) \\ U_i \left(\phi^w_{(0)}, \beta^w_{(1)}, \alpha \right) \end{pmatrix}, \tag{3.39}$$

these simultaneous estimating equations give a marginal variance for the average causal effects that incorporate uncertainty due to the imputation and propensity models. Because the nuisance estimating functions for imputation models are weighted, the variance should be expected to be larger than for the unweighted version (3.33). The estimation procedure for this setting is similar to that given by Equation (3.33).

Chapter 4

Marginal causal modeling for complex survey data

4.1 Estimating functions with survey weights

In this chapter, we extend the MCM procedures of Chapter 3 to complex survey data. Survey inference can be viewed as missing data problem: selected study subjects from the population are considered to be an observed sample, and the rest of the population is missing. However, unlike an ordinary missing data problem where the missing data mechanism is not known, the probability of sampling (inclusion probability) is known in a survey data set. Inferences that do not properly take into consideration the inclusion probabilities may lead to biased estimates for the population of interest.

One way of incorporating the inclusion probability into estimation is to weight the estimating function of interest to establish it as a population representative function. Consider the estimation of a vector parameter φ from a model applied to the population. That is, we define φ to be the solution to $\sum \psi_i = 0$, where ψ_i is the corresponding estimating function for subject i, and the sum is taken over the entire population. If all subjects in the population were selected—i.e., a census was taken—then averaging the function ψ_i over the all study subjects would reproduce φ. In practice, however, the target parameter φ must be estimated from the selected sample. In this case, the estimating function ψ_i may be weighted by the

inverse of the inclusion probability; this weighted form of ψ_i can be then averaged over the chosen sample to estimate φ.

Weighting estimating functions in this way is a well accepted survey inference methodology (Binder, 1983), and it is already implemented in both free and commercial statistical software. Adopting this methodology, our stacked estimating function for the MCM may simply be weighted by the inverse of the inclusion probability to estimate causal effects in the population.

We could estimate causal effects by weighting responses according to the inverse probability of sample inclusion and treatment assignment. The main reason why we may want to avoid this kind of inverse propensity weighting in survey data analysis is that, if the propensity is used as an additional factor in survey weight, a selected subject's response may be in danger of being weighted too much. For example, consider a respondent with a survey weight of 1000. By weighting, she will represent 1000 similar people like her. If we have to weight this person additionally by a propensity weight of 1000, the pseudo sample for this one subject will replicate one million copies of her. If the propensity model were extremely accurate, this inflation might make sense; however, we cannot often suppose that the propensity model is that trustworthy.

4.2 Estimation procedure

We propose a procedure in which the treatment mechanism is estimated by a propensity score whose corresponding score function is weighted by the inverse of the inclusion probability. After estimating the unknown propensity score, we use propensity related covariates in the imputation models for the potential outcomes. The parameters of the imputation model are also estimated by weighted estimating functions. The predictive mean values are, therefore, representative of the population. Once the population representative imputation model is fit, estimating functions for the causal parameters are created, with missing potential outcomes replaced by regression predictions. These functions are also weighted so that the resultant causal estimators account for the complex survey design. In other words, the whole vector of stacked estimating functions ψ_i is weighted by survey weights; this is the basis for the asymptotic sandwich variance proposed by Binder (1983).

Specifically, Equation 3.3 is modified to become

$$\Psi^* = \sum_{i=1}^{n} \begin{pmatrix} w_i A_i(\alpha) \\ w_i(1-t_i) S_i\left(\beta_{(0)}, \alpha\right) \\ w_i(t_i) S_i\left(\beta_{(1)}, \alpha\right) \\ w_i U_i^{*(0)}\left(\phi_{(0)}, \beta_{(0)}, \alpha\right) \\ w_i U_i^{*(1)}\left(\phi_{(1)}, \beta_{(1)}, \alpha\right) \end{pmatrix}, \tag{4.1}$$

where w_i is the inverse of the ith respondent's inclusion probability.

Chapter 5

Application

5.1 Simulated case study

In this section, we present results from a simulated case study to evaluate the performance of the estimators discussed in the previous chapters (Chapters 3-4). Researchers are often interested in the causal effects among the treated group. Table 5-1 shows results for estimators of the average causal effect θ among the treated,

$$\theta_1 = E\left(Y_i\left(1\right) - Y_i\left(0\right)|T_i = 1\right). \tag{5.1}$$

From our artificial population of size 1,000,000, we repeatedly drew 1,000 simple random samples of size 1,000 without replacement. The log odds of the propensity score for diet is estimated under a model that includes the effects of all covariates from Table 1-1. The model is

$$logit\left(\pi_i\right) = v_i^T \widehat{\alpha}, \tag{5.2}$$

where $v_i = (1,\ BLACKC_i,\ NBHISP_i, GRADE_i,\ SLFHLTH_i,\ \ SLFWGHT_i,$ $WORKHARD_i,\ GOODQUAL_i,\ PHYFIT_i,\ PROUD_i,\ LIKESLF_i,\ ACCEPTED_i,$ $FEELOVD_i)$. Note that $E(Y_i\left(1\right)|T_i = 1)$ can be estimated by the sample mean of the observed BMI values among the dieters. The non-trivial part of this problem

is to estimate $E(Y_i(0)\,|T_i=1)$, which equates to

$$P(T_i=1)^{-1}\left\{E\left(Y_i(0)\right)-E(Y_i(0)\,|T_i=0)P(T_i=0)\right\}, \tag{5.3}$$

because

$$E\left(Y_i(0)\right)=E(Y_i(0)\,|T_i=1)P(T_i=1)+E(Y_i(0)\,|T_i=0)P(T_i=0). \tag{5.4}$$

Thus, $E(Y_i(0)\,|T_i=1)$ can be estimated by

$$(n_1/n)^{-1}\left(\widehat{E}\left(Y_i(0)\right)-\overline{y}_0 n_0/n\right), \tag{5.5}$$

where $n_0=\sum_i(1-t_i)$, $n_1=\sum_i t_i$, and $\overline{y}_0$ is the sample mean of the observed $y_i(0)$'s. The imputation estimator $\hat{E}(Y_i(0)\,|T_i=1)$ is

$$\begin{aligned}
& (n_1/n)^{-1}\left(\widehat{E\left(Y_i(0)\right)}-\overline{y}_0 n_0/n\right)\\
= \;& (n_1/n)^{-1}\left(n^{-1}\textstyle\sum_i v_i^T\widehat{\beta}_{(0)}-\overline{y}_0 n_0/n\right)\\
= \;& (n_1/n)^{-1}\left(n^{-1}\textstyle\sum_i v_i^T\widehat{\beta}_{(0)}-\left(n_0^{-1}\textstyle\sum_i(1-t_i)\,y_i(0)\right)n_0/n\right)\\
= \;& n_1^{-1}\left(\textstyle\sum_i t_i v_i^T\widehat{\beta}_{(0)}+\sum_i(1-t_i)\left(v_i^T\widehat{\beta}_{(0)}-y_i(0)\right)\right)\\
= \;& n_1^{-1}\textstyle\sum_i t_i v_i^T\widehat{\beta}_{(0)},
\end{aligned} \tag{5.6}$$

where

$$\widehat{\beta}_{(0)}=\left(\textstyle\sum_i t_i v_i v_i^T\right)^{-1}\left(\sum_i t_i v_i y(t_i)\right). \tag{5.7}$$

The mean structures that used in the simulation study are as follows.
1. Mean structure for the imputation model for the estimator *ace*: $E(Y_i(0)\,|v_i,T_i=1,\beta_{(0)})=v_i^T\beta_{(0)}$ where v_i contains both the analytic variables and confounders.
2. Mean structure for the imputation model for the estimator *ace.I*: $E(Y_i(0)\,|v_i,T_i=1,\beta^I_{(0)})=v_i^{*T}\beta^I_{(0)}$, where

$$v_i^*=\left(1,I_i^{(1)},...,I_i^{(4)},v_i\right), \tag{5.8}$$

and $I_i^{(j)}$ denotes a dummy indicator for propensity stratum j.

3. Mean structure for the imputation model for the estimator *ace.g*: $E(Y_i(0)|v_i, T_i = 1, \beta_{(0)}) = v_i^{*T}\beta_{(0)}^{g}$, where

$$v_i^* = \left(1, e_i, e_i^2, (e_i - k_1)_+^2, (e_i - k_2)_+^2, (e_i - k_3)_+^2, x_i, z_i\right), \tag{5.9}$$

e_i is a logit propensity, and (k_1, k_2, k_3) is quartiles of the estimated logit-propensities.

4. Logit propensity for (*ace.smi*): $logit(\widehat{\pi}_i) = v_i^{*T}\beta_{(0)}$, where

$$v_i^* = \left(1, I_i^{(1)}, ..., I_i^{(4)}, \gamma_{(0)i}^T v_i\right) \text{ for } i \in \{i : T_i = 0\}. \tag{5.10}$$

6. Mean structure for the *ace.oipw*, which is an IPW estimator :

$$\widehat{E}(Y_i(0)) = n^{-1}\textstyle\sum_i \dfrac{(1 - t_i)\, y_i(0)}{(1 - \widehat{\pi}_i)}. \tag{5.11}$$

7. Mean structure for the *ace.oipdr*, a doubly robust inverse-propensity estimator whose form is similar to that of Equation (2.2). Let $\widehat{y}_i(0) = v_i^{*T}\widehat{\beta}^{ipdr}$, and $v_i^* = (1, v_i)$, then

$$\widehat{E}(Y_i(0)) = n^{-1}\textstyle\sum_i \left(\widehat{y}_i(0) + \dfrac{t_i}{\widehat{\pi}_i}(y_i(0) - \widehat{y}_i(0)\right). \tag{5.12}$$

8. Mean structure for the *ace.ipw*, a stabilized version of an IPW estimate whose denominator is replaced by weighted sample total:

$$\widehat{E}(Y_i(0)) = \left(\sum\nolimits_i \frac{(1 - t_i)\, y_i(0)}{(1 - \widehat{\pi}_i)}\right)\left(\sum\nolimits_i \frac{(1 - t_i)}{(1 - \widehat{\pi}_i)}\right)^{-1}. \tag{5.13}$$

9. Mean structure for the *ace.ipdr*, a stabilized doubly robust estimator whose denominator is replaced by weighted sample total. Let $\widehat{y}_i(0) = v_i^{*T}\widehat{\beta}^{ipdr}$, and $v_i^* = (1, v_i)$, then

$$\widehat{E(Y_i(0))} = \left(\sum\nolimits_i \left(\widehat{y}_i(0) + \frac{t_i}{\widehat{\pi}_i}(y_i(0) - \widehat{y}_i(0))\right)\right)\left(\sum\nolimits_i \frac{(1 - t_i)}{(1 - \widehat{\pi}_i)}\right)^{-1}. \tag{5.14}$$

Table 5.1. Results of simulation study among the treated

Estimators	*Bias*	*RMSE*	*%Bias*	*MAE*
ace	0.08	0.2	40.1	0.14
ace.I	0.04	0.2	20	0.14
ace.g	**0**	**0.2**	**1.1**	**0.14**
ace.smi	0.04	0.2	18.4	0.14
ace.oipw	-0.17	0.64	-27.5	0.41
ace.oipdr	0	0.2	-1.6	0.14
ace.ipw	-0.11	0.35	-34.9	0.22
ace.ipdr	0	0.2	-1.6	0.14
ace.bripdr	2.82	2.85	643.2	2.81

10. Mean structure for the *ace.ipbr* (Bang and Robins, 2005). Let

$$v_i^* = \left(1, t_i, v_i, \left\{t_i\widehat{\pi}_i + (1-t_i)\left(1-\widehat{\pi}_i\right)\right\}^{-1}\right),$$

then

$$\widehat{E}\left(Y_i\left(0\right)\right) = n^{-1}\textstyle\sum_i\left(v_i^{*T}\widehat{\beta}^{ipbr}\right), \tag{5.15}$$

where

$$\widehat{\beta}^{ipbr} = \left(\textstyle\sum_i t_i v_i^* v_i^{*T}\right)^{-1}\left(\textstyle\sum_i t_i v_i^* y(t_i)\right). \tag{5.16}$$

In Table 5.1, "*Bias*" is the average difference between the estimates and the true average causal effect, among the treated (0.1411) from our artificial population, and "*%Bias*" is the bias as a percentage of the estimate's standard deviation. (A useful rule-of-thumb is that the performance of interval estimates and test statistics begins to deteriorate when the bias of the point estimate exceeds about 40% of its standard deviation.) "*RMSE*" is square root of the average value of difference between the estimates and the true average causal effect. "*MAE*" reports the median of the absolute differences between the estimates and the true average causal effect which discards the worst 50% of the estimates. Among all estimators, *ace.g* has the best performance with respect to all criteria of measures. This result suggests that a spline form of the logit propensity related covariates effectively corrects the bias induced by a misspecified imputation model. The

Table 5.2. Results of simulation study for the entire population

Estimators	*Bias*	*RMSE*	*%Bias*	*MAE*
ace.I	0.02	0.2	10.9	0.13
ace.g	**0.01**	**0.19**	**3.6**	**0.14**
ace.oipdr	0.03	0.24	11.5	0.16
ace.ipdr	0.02	0.23	10	0.16

IPW estimators are biased. The IPDR estimator, which introduces an augmented function to make IPW more robust, corrects the bias of IPW remarkably well; it is as good as the imputation estimators. The IPBR estimator is a class of IPDR estimator suggested by Bang and Robins (2005), who wrote:

> "We see that we must add to the regression the inverse probability of treatment weighted covariate..., which is the (estimated) inverse of the PS(propensity score) for treated subjects (T_i= 1) and the inverse of "1 minus the PS" for untreated subjects (T_i= 0). Other choices can result in inconsistent estimation of the average treatment effect."

This simulation study, however, suggests that the estimator of Bang and Robins (2005) may perform more poorly than the other methods, even with respect to MAE. The doubly robust estimators (*ace.ipdr* and *ace.oipdr*) and predictive imputation with propensity scores (*ace.I* and *ace.g*) perform similarly. But for estimating the ACE in the entire population (Table 5.2), *ace.g* performs better than other estimates.

We now turn to the motivating example of Chapter 1 in which we assess the causal effects within categories defined by the variable WORKHARD. Using the estimator *ace.g*, we construct a confidence interval for the ACE as in Figure 1.1. The result is shown in Figure 5.1. Because the population is large (10^6), the confidence intervals (lines within the boxes) have shrunk.

Although Figure 5.1 was drawn based on the entire population, it is worthwhile to examine a sample randomly drawn from this population (Figure 5.2). The boxplots show the causal effects for individuals based on the known potential outcomes in a random sample of size 1000. The arrow lines indicate the confidence intervals for the estimated average causal effects from predictive mean imputation using a quadratic spline basis defined by the logit propensity score (*ace.g*). Recall that the

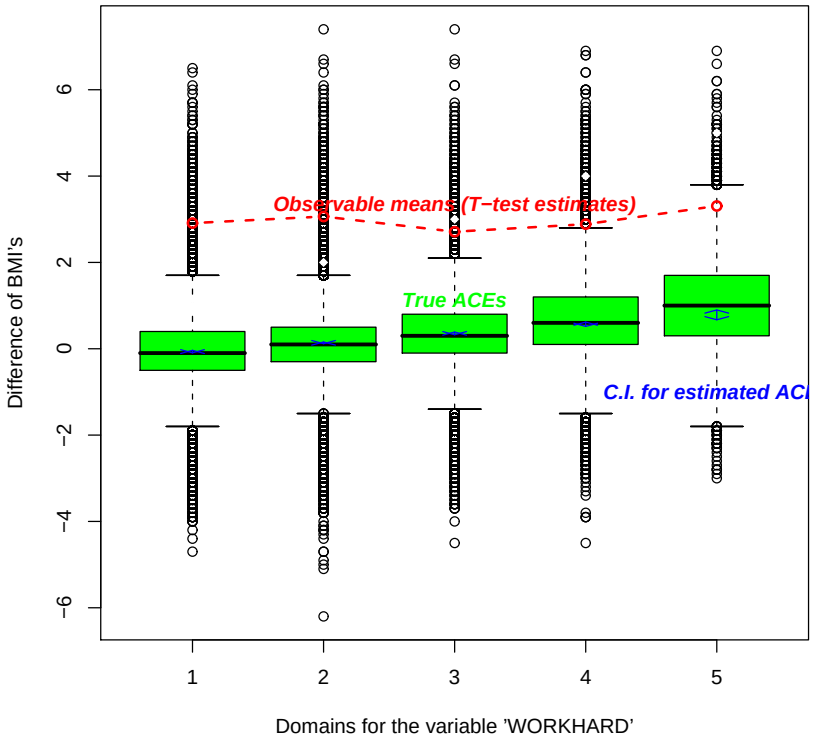

Figure 5.1. True ACEs and estimated ACEs for the entire population

domain is the variable WORKHARD. In this arbitrary sample, the number of the girls with WORKHARD = 5 is only 16, which may explain why the estimation at this domain is inefficient. The 95% confidence intervals all cover the true ACE's.

Figure 5.2 illustrates that in a random sample of adolescent girls, the ACE's vary across the observed levels of the item "WORKHARD". Those who believe that if they work hard their situation may improve tend to experience beneficial effects of dieting. On the other hand, those who do not believe so rarely diet, but when they do, the effects of dieting are less beneficial.

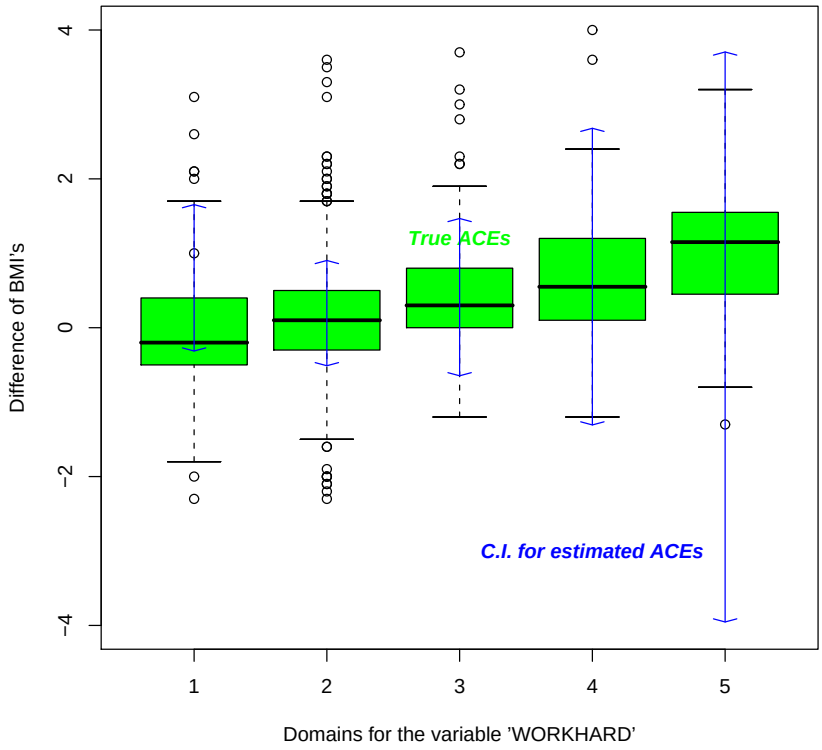

Figure 5.2. True ACEs and estimated ACEs for a random sample of size 1000

5.2 Case study: causal inference with a complex survey

In Section 5.1, we evaluated the performance of our causal estimators and found that the imputation methods perform well. In this section, we apply the MCM to the actual data from the Add Health study, which was drawn by a complex multistage design (stratified cluster design). We will use the methods of Chapter 4 to estimate the average causal effect of dieting in the overall population.

This analysis uses 5,140 girl students from the original Add Health sample with complete data responses for the necessary variables; that is, girls with incomplete responses were not included in the estimation. Figure 5.3 shows that dieters and

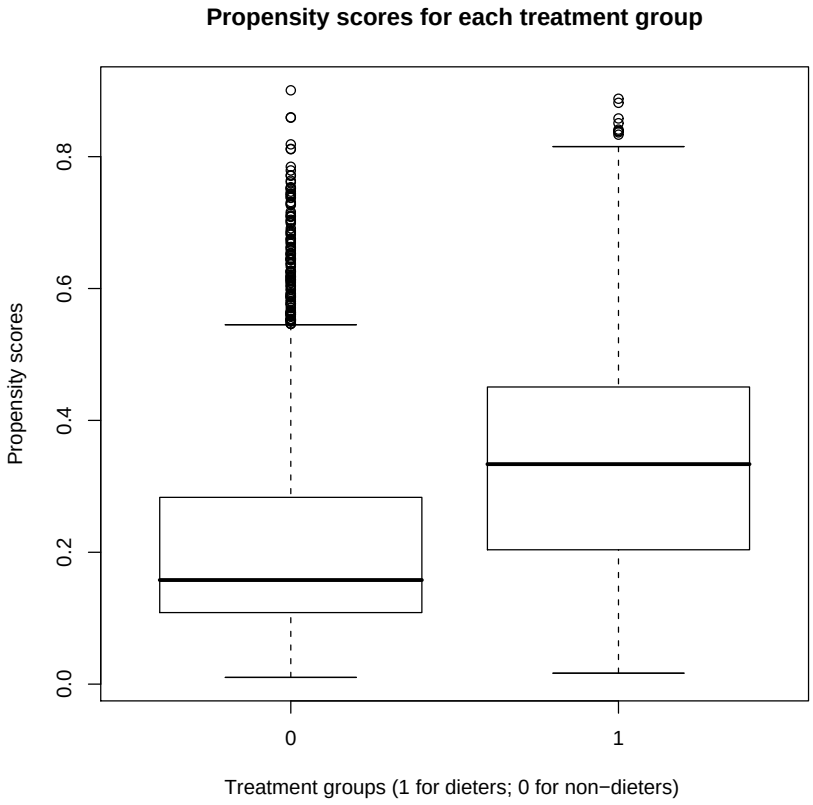

Figure 5.3. The estimated propensity scores of the original population

non-dieters differ with respect to their estimated propensity scores. This propensity score model is same as in the previous section, except that the survey weight was included to weight the score function in order to estimate population representative propensity scores.

Figure 5.3 shows that the two groups—dieters and non-dieters— overlap in their tendency to diet. Based on the performance of causal estimators for the artificial population in Section 5.1, we chose an imputation model that included a quadratic spline basis for the propensity scores. We used the survey weight to weight the estimating function for imputation parameters, so that the predictive means represent values from population. The R-squared measures for both imputation models were about 80%.

Figure 5.4 shows the estimated causal effects within classes defined by the

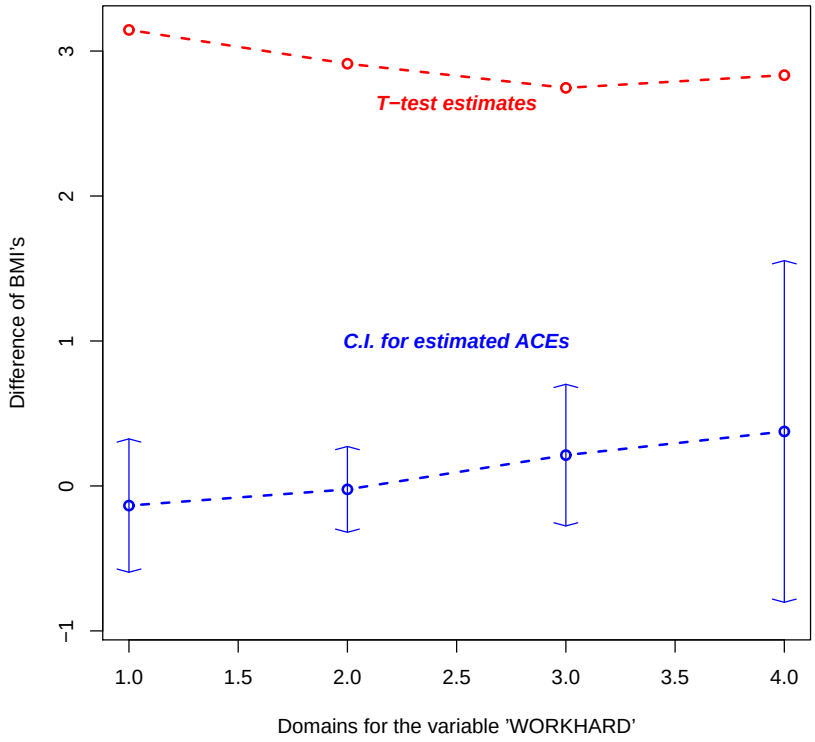

Figure 5.4. ACE's among the survey sample indicated by the boxplot with the dotted lines representing t-test estimates from the observable potential outcomes

analytic variable "WORKHARD." In Firgure 5.4, the last two categories were combined into one because of the small number of girls who answered "strongly disagree" (category 5). This result is similar to what we found in Section 5.1; the t-test estimate shows consistently positive ACE's across the analytic variable, whereas the MCM shows no significant ACE's. MCM does show a tendency for the causal effects to increase with respect to the analytic variable, but overall the average effects of diet on BMI are practically negligible.

Chapter 6

Conclusions

6.1 Imputation: a solution to the problem of confounding effects

The modern trend in causal inference is to compare potential outcomes for the same set of units. We have proposed to control for nuisance confounders by establishing a regression model to predict the missing potential outcomes based on relationships between potential outcomes, covariates of interest and confounders. The prediction and imputation are based on a semi-parametric imputation model with propensity-related covariates which conveys information about the treatment mechanism. Following imputation, ACE's can be directly estimated from observed and imputed potential outcomes. In this way, imputation provides an intuitively appealing method for estimating ACE's, and it may have substantial advantages over other propensity related methods.

6.2 Advantages of an estimating-equation approach to imputation

Multiple imputation (Rubin, 1987; 2005) for causal inference would require a fully parametric model for the potential outcomes, including specification of the inestimable correlations among the potential outcomes. The necessity of correctly

specifying the correlation between potential outcomes which can never be jointly observed is a significant drawback of fully parametric approaches. Because the observed data provide no information for estimating this correlation, the assumptions about this correlation can never be tested. Our predictive mean imputation approach, however, does not require assumptions about the correlation if we are using an identity, log or logit link function. The uncertainty introduced by predicting the missing potential outcomes is accounted for through simultaneous estimating equations. This estimation could possibly be made more efficient by recasting the simultaneous estimating functions as a quadratic inference functions (Qu et al., 2000).

6.3 A different way to construct doubly robust estimators

Scharfstein et al. (1999) and Bang and Robins (2005) use the inverse propensity as a covariate in a regression model. Inverse propensity related methods may work fairly well for the case where the propensity scores are bounded away from extreme values (0 or 1). But the practical limits for these bounds are not clear. Instead of using inverse propensities, we have found that accounting for the propensity scores through strata (dummy indicators) or spline functions in the regression models helps to reduce bias.

6.4 Measurement of diet

Either indirectly through simulation or directly through real data analysis, we found no significant causal effects of dieting on BMI. However the measurement of dieting in this survey is not of high quality; it is a self-report measure pertaining to a one-week period, whereas the outcome BMI was measured one year later. Ideally one would like a more reliable and valid way to measure dieting behaviors perhaps formulating it as a latent variable. Causal inference in which the treatment indicator is a latent variable is an important topic for future research, but it is beyond the scope of this thesis.

Appendix A

Appendix

A.1 Proof of Lemma 3-1

Denote the scalar components of the working variance functions by $Var\left(y_i\left(t\right)\right)=\upsilon_t$ and $Cov\left(y_i\left(0\right),y_i\left(1\right)\right)=\upsilon_{01}$. Also let $h_{ti}=\partial\mu_i\left(t\right)/\partial\phi\left(t\right)$ and $s_{ti}=y_i\left(t\right)-\mu_i\left(t\right)$. Then the contribution of unit i to the estimating function,

$$\Psi_i(\phi)=\left(\partial\mu_i/\partial\phi\right)V\left(\mu_i\right)^{-1}\left(y_i-\mu_i\right),$$

under the identity link can be specified as

$$\begin{aligned}\Psi_i(\phi)&=\begin{pmatrix}h_{0i} & \frac{\partial\mu_i(0)}{\partial\phi(1)}\\ \frac{\partial\mu_i(1)}{\partial\phi(0)} & h_{1i}\end{pmatrix}\begin{pmatrix}\upsilon_0 & \upsilon_{01}\\ \upsilon_{01} & \upsilon_1\end{pmatrix}^{-1}\begin{pmatrix}s_{0i}\\ s_{1i}\end{pmatrix}\\&=C_\upsilon\begin{pmatrix}h_{0i}\upsilon_1 s_{0i}-h_{0i}\upsilon_{01}s_{1i}\\ -h_1\upsilon_{01}s_{0i}+h_1\upsilon_0 s_{1i}\end{pmatrix},\end{aligned}$$

where $C_\upsilon=\left(\upsilon_0\upsilon_1-\upsilon_{01}^2\right)^{-1}$, because $\partial\mu_i\left(0\right)/\partial\phi\left(1\right)=\partial\mu_i\left(1\right)/\partial\phi\left(0\right)=0$. If we assume the same mean structure for $\mu_i\left(1\right)$ and $\mu_i\left(0\right)$, with the identity link $g(a)=a$, then $h_{0i}=h_{1i}\overset{def}{=}h_i$. Now let $\sum_{i=1}^{n}h_i s_{ti}=s_t^*$, then the estimating equations

$\sum_{i=1}^{n}\Psi_i(\phi) \overset{set}{=} [0]$ are equivalent to

$$\begin{pmatrix} C_{\upsilon}(\upsilon_1 s_0^* - \upsilon_{01} s_1^*) \\ C_{\upsilon}(\upsilon_0 s_1^* - \upsilon_{01} s_0^*) \end{pmatrix} \overset{set}{=} [0],$$

which is same as

$$\begin{pmatrix} \upsilon_1^{-1} s_1^* \\ \upsilon_0^{-1} s_0^* \end{pmatrix} \overset{set}{=} [0].$$

This ends the proof for the case of the identity link. Note that $h_{ti} = \upsilon_t x_i^T$ for the case of logit link and $h_{ti} = \mu_i(t) x_i^T$ with $\upsilon_t = \mu_i(t)$ for the log link. The proofs for these two cases proceed as for the identity link. ■

A.2 Proof of Lemma 3-2

Consider the identity link *g(a)=a* and let $\widehat{\phi}(0) = n^{-1}\sum_{i=1}^{n} y_i^*(0)$ be the solution to

$$\sum_i c(0)\left(y_i^*(0) - z_i^T \phi_{(0)}\right) = 0,$$

where (0) is a constant vector. Then

$$n^{-1}\sum_{i=1}^{n} y_i^*(0) = n^{-1}\sum_{i=1}^{n}\left(\widehat{\beta}_0(0) + \upsilon_i^T \widehat{\beta}_1(0)\right),$$

because $\sum_{i=1}^{n} E(Y_i(0)|t_i = 0, \upsilon_i) = \sum_{i=1}^{n} Y_i(0)(1 - t_i)$.

Now suppose that

$$\begin{pmatrix} Y(0) \\ Y(1) \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_0(\sigma_0^2) & \Sigma_{01}(\rho\sigma_0\sigma_1) & \Sigma_{0X} \\ \Sigma_{10}(\rho\sigma_1\sigma_0) & \Sigma_1(\sigma_1^2) & \Sigma_{1X} \\ \Sigma_{X0} & \Sigma_{X1} & \Sigma_X \end{pmatrix}\right),$$

where $\Sigma_t(\cdot)$ and $\Sigma_{01}(\cdot)$ are diagonal matrices with the argument appearing on the diagonal. Then, assuming MAR, let $\varphi(0) = \left(\mu_X, \Sigma_X, \beta_0(0), \beta_1(0), \Sigma_{0|X}\right)$, $\beta_0(0)$ $= \mu_0 - \Sigma_{0X}\Sigma_X^{-1}\mu_X$, and $\beta_1(0) = \Sigma_{0X}\Sigma_X^{-1}$, $\Sigma_{0|X} = \Sigma_0 - \Sigma_{0X}\Sigma_X^{-1}\Sigma_{X0}$ so that the

likelihood becomes

$$\begin{aligned} L\left(\varphi\left(0\right);y\left(0\right)_{obs},X\right) &\propto P\left(y\left(0\right)_{obs},X;\varphi\left(0\right)\right) \\ &= \int P\left(y\left(0\right)_{obs},y\left(0\right)_{mis},X;\varphi\left(0\right)\right)d\left(y\left(0\right)_{mis}\right) \\ &= \prod_{i=1}^{n} f\left(X_i;\mu_X,\sigma_X^2\right)\prod_{i\in obs} f\left(y_i\left(0\right)|X_i;\beta_0\left(0\right),\beta_1\left(0\right),\Sigma_{0|X}\right). \end{aligned}$$

The MLE for $\varphi\left(0\right)$ is $\widehat{\varphi}\left(0\right)=\left(\widehat{\mu}_X,\widehat{\sigma}_X^2,\widehat{\beta}_0\left(0\right),\widehat{\beta}_1\left(0\right),\widehat{\sigma}_{0|X}^2\right)$ which is the solution to $\partial logL\left(\varphi\left(0\right);y\left(0\right)_{obs},X\right)/\partial\varphi=0$. $\widehat{\mu}_X$ is the sample mean of X, and $\widehat{\Sigma}_X$ is the sample variance of X. From the regression $f\left(y_i\left(0\right)|X_i;\beta_0\left(0\right),\beta_1\left(0\right),\Sigma_{0|X}\right)$, we obtain the consistent OLS estimators $\widehat{\beta}_0\left(0\right)$ and $\widehat{\beta}_1\left(0\right)$. Then we set up the two equations as

$$\widehat{\beta}_0\left(0\right)=\mu_0-\Sigma_{0X}\Sigma_X^{-1}\widehat{\mu}_X,\widehat{\beta}_1\left(0\right)=\Sigma_{0X}\Sigma_X^{-1}.$$

Solving the two sets of equation above, we get the estimator for μ_0

$$\widehat{\mu}_0=\widehat{\beta}_0\left(0\right)+\widehat{\mu}_X^T\widehat{\beta}_1\left(0\right)=n^{-1}\sum_{i=1}^{n}\left(\widehat{\beta}_0\left(0\right)+X_i^T\widehat{\beta}_1\left(0\right)\right),$$

which ends the proof. ■

A.3 Variance estimation with simultaneous estimating equations

The multivariate asymptotic distribution for a vector of parameter estimates based on estimating equations (3.33) is

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}_{(0)}-\beta_{(0)}\\ \widehat{\beta}_{(1)}-\beta_{(1)}\\ \widehat{\phi}_{(0)}-\phi_{(0)}\\ \widehat{\phi}_{(1)}-\phi_{(1)}\end{pmatrix}\longrightarrow N(0,\Sigma),$$

where $\Sigma = H^{-1}E\left(\Psi_i^+\Psi_i^{+T}\right)\left(H^{-1}\right)^T$,

$$\Psi_i^+ = \begin{pmatrix} A_i = A_i\left(\alpha\right) \\ S_{0i} = S_i\left(\beta_{(0)}, \alpha\right) \\ S_{1i} = S_i\left(\beta_{(1)}, \alpha\right) \\ U_{0i} = U_i\left(\phi_{(0)}, \beta_{(0)}, \alpha\right) \\ U_{1i} = U_i\left(\phi_{(1)}, \beta_{(1)}, \alpha\right) \end{pmatrix},$$

and

$$\Psi_i^+\Psi_i^{+T} = \begin{pmatrix} A_iA_i^T & A_iS_{0i}^T & A_iS_{1i}^T & A_iU_{0i}^T & A_iU_{1i}^T \\ S_{0i}A_i^T & S_{0i}S_{0i}^T & 0 & S_{0i}U_{0i}^T & 0 \\ S_{1i}A_i^T & 0 & S_{1i}S_{1i}^T & 0 & S_{1i}U_{1i}^T \\ U_{0i}A_i^T & U_{0i}S_{0i}^T & 0 & S_{0i}^TU_{0i}^T & 0 \\ U_{1i}A_i^T & 0 & U_{1i}S_{1i}^T & 0 & U_{1i}U_{1i}^T \end{pmatrix}.$$

Also,

$$H = \begin{pmatrix} H_\alpha & 0 & 0 & 0 & 0 \\ H_{\beta_0\alpha} & H_{\beta_0} & 0 & 0 & 0 \\ H_{\beta_1\alpha} & 0 & H_{\beta_1} & 0 & 0 \\ H_{\phi_{(0)}\alpha} & H_{\phi_{(0)}\beta_0} & 0 & H_{\phi_{(0)}} & 0 \\ H_{\phi_{(1)}\alpha} & 0 & H_{\phi_{(0)}\beta_0} & 0 & H_{\phi_{(1)}} \end{pmatrix}$$

$$= E\begin{pmatrix} \frac{\partial A_i(\alpha)}{\partial\alpha} & 0 & 0 & 0 & 0 \\ \frac{\partial S_i(\beta_0,\alpha)}{\partial\alpha} & \frac{\partial S_i(\beta_0,\alpha)}{\partial\beta_0} & 0 & 0 & 0 \\ \frac{\partial S_i(\beta_1,\alpha)}{\partial\alpha} & 0 & \frac{\partial S_i(\beta_1,\alpha)}{\partial\beta_1} & 0 & 0 \\ \frac{\partial U_i\left(\phi_{(0)},\beta_0,\alpha\right)}{\partial\alpha} & \frac{\partial U_i\left(\phi_{(0)},\beta_0,\alpha\right)}{\partial\beta_0} & 0 & \frac{\partial U_i\left(\phi_{(0)},\beta_0,\alpha\right)}{\partial\phi_{(0)}} & 0 \\ \frac{\partial U_i\left(\phi_{(1)},\beta_1,\alpha\right)}{\partial\alpha} & 0 & \frac{\partial U_i\left(\phi_{(1)},\beta_1,\alpha\right)}{\partial\beta_1} & 0 & \frac{\partial U_i\left(\phi_{(1)},\beta_1,\alpha\right)}{\partial\phi_{(1)}} \end{pmatrix}.$$

The marginal asymptotic distribution for the average causal estimand $\phi_{(t)}$ for $t = 0, 1$ is

$$\sqrt{n}\left(\widehat{\phi}_{(t)} - \phi_{(t)}\right) \longrightarrow N\left(0, \sigma^2_{\phi_{(t)}}\right).$$

Then, by matrix algebra, it can be shown that the marginal limiting variance for the estimator $\widehat{\phi}_{(t)}$ is a sum of variance components,

$$\sigma^2_{\phi_{(t)}} = V(\phi_{(t)}) + V(\phi_{(t)}, \alpha) + V(\phi_{(t)}, \beta_{(t)}) + V(\phi_{(t)}, \beta_{(t)}, \alpha),$$

where the components are

$$\begin{aligned}
V(\phi_{(t)}) &= H^{-1}_{\phi_{(t)}} E\left(UU^T\right)\left(H^{-1}_{\phi_{(t)}}\right)^T, \\
V(\phi_{(t)}, \alpha) &= H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\alpha} H^{-1}_{\alpha} E\left(AA^T\right)\left(H^{-1}_{\alpha}\right)^T \left(H_{\phi_{(t)}\alpha}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T \\
&\quad - H^{-1}_{\phi_{(t)}} E\left(UA^T\right)\left(H^{-1}_{\alpha}\right)^T \left(H_{\phi_{(t)}\alpha}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T \\
&\quad - H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\alpha} H^{-1}_{\alpha} E\left(AU^T\right)\left(H^{-1}_{\phi_{(t)}}\right)^T, \\
V(\phi_{(t)}, \beta_{(t)}) &= H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\beta_{(t)}} H^{-1}_{\beta_{(t)}} E\left(S_0 S_0^T\right)\left(H^{-1}_{\beta_{(t)}}\right)^T \left(H_{\phi_{(t)}\beta_{(t)}}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T \\
&\quad - H^{-1}_{\phi_{(t)}} E\left(US_0^T\right)\left(H^{-1}_{\beta_{(t)}}\right)^T \left(H_{\phi_{(t)}\beta_{(t)}}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T \\
&\quad - H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\beta_{(t)}} H^{-1}_{\beta_{(t)}} E\left(S_0 U^T\right)\left(H^{-1}_{\phi_{(t)}}\right)^T, \\
V(\phi_{(t)}, \beta_{(t)}, \alpha) &= H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\beta_{(t)}} H^{-1}_{\beta_{(t)}} E\left(S_0 A^T\right)\left(H^{-1}_{\alpha}\right)^T \left(H_{\phi_{(t)}\alpha}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T \\
&\quad + H^{-1}_{\phi_{(t)}} H_{\phi_{(t)}\alpha} H^{-1}_{\alpha} E\left(AS_0^T\right)\left(H^{-1}_{\beta_{(t)}}\right)^T \left(H_{\phi_{(t)}\beta_{(t)}}\right)^T \left(H^{-1}_{\phi_{(t)}}\right)^T.
\end{aligned}$$

It can also be shown that the marginal asymptotic variance for the estimator $\widehat{\theta}$ is

$$\sigma^2_{\theta} = \sigma^2_{\phi_{(t)}} + cov\left(\widehat{\phi}_{(0)}, \widehat{\phi}_{(1)}\right),$$

where $cov\left(\widehat{\phi}_{(0)}, \widehat{\phi}_{(1)}\right)$ can be found in the covariance matrix Σ. Specifically,

$$\begin{aligned}
cov\left(\widehat{\phi}_{(0)}, \widehat{\phi}_{(1)}\right) &= H^{-1}_{\phi_{(1)}} H_{\phi_{(1)}\alpha} H^{-1}_{\alpha} E\left(AA^T\right)\left(H^{-1}_{\phi_{(1)}} H_{\phi_{(1)}\alpha} H^{-1}_{\alpha}\right)^T + \\
&\quad H^{-1}_{\phi_{(1)}} H_{\phi_{(1)}\beta_1} H^{-1}_{\beta_1} E\left(S_{\beta_1} A^T\right)\left(H^{-1}_{\phi_{(0)}} H_{\phi_{(0)}\alpha} H^{-1}_{\alpha}\right)^T - \\
&\quad H^{-1}_{\phi_{(1)}} E\left(U_1 A^T\right)\left(H^{-1}_{\phi_{(0)}} H_{\phi_{(0)}\alpha} H^{-1}_{\alpha}\right)^T + \\
&\quad H^{-1}_{\phi_{(1)}} H_{\phi_{(1)}\alpha} H^{-1}_{\alpha} E\left(AS_0^T\right)\left(H^{-1}_{\phi_{(0)}} H_{\phi_{(0)}\beta_0} H^{-1}_{\beta_0}\right)^T -
\end{aligned}$$

$$H_{\phi_{(1)}}^{-1} H_{\phi_{(1)}\alpha} H_{\alpha}^{-1} E\left(AU_0^T\right)\left(H_{\beta_0}^{-1}\right)^T.$$

Uncertainty due to estimation of the nuisance parameters needed to estimate the ACE is reflected in the formula above.

A.4 Proof of double robustness of the WLS regression estimator (3.39)

Under the linear model $y_i(1) = y_i = v_i^T\beta + \epsilon_i$, β can be estimated by WLS estimators defined as

$$\widehat{\beta} = \left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i v_i^T\right)^{-1} \left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i y_i\right).$$

We now show that $\widehat{\mu}_{WLS} = \frac{1}{n}\sum_{i=1}^{n} v_i^T\widehat{\beta}$ is an asymptotically doubly robust estimator. First, suppose that the propensity model is not correctly specified but the regression model is correctly specified. In that case, $\hat{\mu}_{WLS}$ converges to

$$\begin{aligned}
E\left(v_i^T\widehat{\beta}\right) &= E\left(E\left(v_i^T\left(\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}v_iv_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}v_iy_i\right)|v_i\right)\right)\\
&= E\left(v_i^T\left(\sum_{i=1}^{n} \pi_i\widehat{\pi}_i^{-1}v_iv_i^T\right)^{-1}\left(\sum_{i=1}^{n} \pi_i\widehat{\pi}_i^{-1}v_iE(y_i|v_i)\right)\right)\\
&= E\left(v_i^T\left(\sum_{i=1}^{n} \pi_i\widehat{\pi}_i^{-1}v_iv_i^T\right)^{-1}\left(\sum_{i=1}^{n} \pi_i\widehat{\pi}_i^{-1}v_iv_i^T\beta\right)\right)\\
&= E\left(v_i^T\beta\right) = E\left(E(y_i|v_i)\right) = E(y_i) = \mu.
\end{aligned}$$

Thus, $\hat{\mu}_{WLS}$ is asymptotically unbiased estimator of the average causal effect μ.

Now suppose the regression model is not correctly specified but the propensity model is correctly specified. In that case,

$$E\left(\widehat{\beta}\right) = E\left(\left(\sum_{i=1}^{n} t_i\widehat{\pi}^{-1}v_iv_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i\widehat{\pi}^{-1}v_iy_i\right)\right)$$

$$
\begin{aligned}
&= E\left(\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i v_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i \left(v_i^T\beta + \epsilon_i\right)\right)\right) \\
&= \beta + E\left(\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i v_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i \epsilon_i\right)\right) \\
&= \beta + E\left(E\left(\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i v_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i \widehat{\pi}^{-1} v_i \epsilon_i\right)|v_i\right)\right) \\
&= \beta + E\left(E\left(\left(\sum_{i=1}^{n} v_i v_i^T\right)^{-1}\left(\sum_{i=1}^{n} v_i E(\epsilon_i|v_i)\right)\right)\right).
\end{aligned}
$$

The term after β in last line converges to zero because, under ignorability,

$$E(\epsilon_i|v_i)=E(\epsilon_i|v_i, t_{i=1})$$

and $E(\epsilon_i|v_i, t_{i=1}) \approx \frac{1}{m}\sum_{i=1}^{m}\left(y_i - v_i^T\widehat{\beta}\right) = 0$, where $\widehat{\beta} = \left(\sum_{i=1}^{n} t_i v_i v_i^T\right)^{-1}\left(\sum_{i=1}^{n} t_i v_i y_i\right)$. Now,

$$
\begin{aligned}
\left(E\left(\widehat{\beta}\right) \approx \beta\right) &\iff E\left(\widehat{\beta} - \beta\right) \approx 0 \\
&\iff E\left(E\left(\widehat{\beta} - \beta|v_i\right)\right) \approx 0 \\
&\implies E\left(E\left(v_i^T\widehat{\beta} - v_i^T\beta|v_i\right)\right) \approx 0 \\
&\iff E\left(v_i^T\widehat{\beta}\right) \approx E\left(v_i^T\beta\right) \\
&\iff \frac{1}{n}\sum_{i=1}^{n} v_i^T\widehat{\beta} \approx \frac{1}{n}\sum_{i=1}^{n} v_i^T\beta \\
&\iff \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\widehat{\beta}\right) \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\beta\right).
\end{aligned}
$$

Also, because $E(\epsilon_i|v_i) \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\beta\right) \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\widehat{\beta}\right)$ by the weak ignorability condition,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\widehat{\beta}\right) \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\widehat{\beta}\right)\right)$$

$$\begin{aligned}
&\iff \left(0 \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_i - v_i^T\widehat{\beta}\right) \approx E\left(y_i - v_i^T\widehat{\beta}\right)\right) \\
&\iff E\left(v_i^T\widehat{\beta}\right) \approx E\left(y_i\right) = \mu \\
&\iff E\left(\frac{1}{n}\sum_{i=1}^{n}\left(v_i^T\widehat{\beta}\right)\right) = E\left(\widehat{\mu}_{WLS}\right) \approx \mu,
\end{aligned}$$

which ends the proof. ■

A.5 Proof of double robustness of the improved DR estimator (3.37)

Define

$$\widehat{\mu}_{IPDR2} = \frac{1}{n}\sum_{i=1}^{n} v_i^T\widehat{\beta}^{IPDR} + \frac{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}\left(y_i - v_i^T\widehat{\beta}^{IPDR}\right)}{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}}.$$

We now show that $\widehat{\mu}_{IPDR2}$ is an asymptotically doubly robust estimator for the ACE μ. First, suppose that the propensity model is not correctly specified but the regression model is correctly specified. In that case, let

$$R = \frac{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}\left(y_i - v_i^T\widehat{\beta}^{IPDR}\right)}{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}}.$$

Then,

$$\begin{aligned}
&COV\left(\sum_i t_i\widehat{\pi}_i^{-1}, R\right) \\
&= E\left(\sum_i t_i\widehat{\pi}_i^{-1}R\right) - E\left(\sum_i t_i\widehat{\pi}_i^{-1}\right)E\left(R\right) \\
&= E\left(\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}\left(y_i - v_i^T\widehat{\beta}^{IPDR}\right)\right) - E\left(\sum_i t_i\widehat{\pi}_i^{-1}\right)E\left(R\right)
\end{aligned}$$

$$= E\left(E\left(\sum_i t_i\widehat{\pi}_i^{-1}\left(y_i - v_i^T\widehat{\beta}^{IPDR}\right)|v_i\right)\right) - E\left(E\left(\sum_i t_i\widehat{\pi}_i^{-1}|v_i\right)\right)E\left(R\right)$$
$$= C(0 - E\left(R\right)),$$

Where $C = \sum_i C_i = E\left(\sum_i t_i\widehat{\pi}_i^{-1}|v_i\right) = \sum_i E(t_i|v_i)\widehat{\pi}_i^{-1}$. It follows that

$$C(0 - E\left(R\right)) = \frac{COV\left(\sum_i t_i\widehat{\pi}_i^{-1}, R\right)}{\sqrt{Var(\sum_i t_i\widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}}$$
$$\leq \frac{1}{\sqrt{Var(\sum_i t_i\widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}}.$$

and hence

$$0 - E\left(R\right) \leq \frac{1}{\sum_i C_i\sqrt{Var(\sum_i t_i\widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}} \longrightarrow 0$$

as $n \longrightarrow 0$ if C_i is consistently bounded away from 0. Since the regression model is correctly specified,

$$E\left(\frac{1}{n}\sum_{i=1}^{n} v_i^T\widehat{\beta}^{IPDR}\right) = \mu.$$

Therefore, $\widehat{\mu}_{IPDR2}$ is an asymptotically unbiased estimator for μ.

Now suppose the regression model is not correctly specified but the propensity model is correctly specified. In that case, let

$$R = \frac{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}\left(y_i - v_i^T\widehat{\beta}^{IPDR}\right)}{\sum_{i=1}^{n} t_i\widehat{\pi}_i^{-1}}.$$

Then

$$COV\left(\sum_i t_i\widehat{\pi}_i^{-1}, R\right)$$
$$= E\left(\sum_i t_i\widehat{\pi}_i^{-1}R\right) - E\left(\sum_i t_i\widehat{\pi}_i^{-1}\right)E\left(R\right)$$

$$
\begin{aligned}
&= E\left(\sum_{i=1}^{n} t_i \widehat{\pi}_i^{-1}\left(y_i - v_i^T \widehat{\beta}^{IPDR}\right)\right) - E\left(\sum_i t_i \widehat{\pi}_i^{-1}\right) E\left(R\right) \\
&= E\left(E\left(\sum_i t_i \widehat{\pi}_i^{-1}\left(y_i - v_i^T \widehat{\beta}^{IPDR}\right)|v_i\right)\right) - E\left(E\left(\sum_i t_i \widehat{\pi}_i^{-1}|v_i\right)\right) E\left(R\right) \\
&= n\left(\mu - E(v_i^T\beta)\right) - nE\left(R\right) \\
&= n(\mu - E(v_i^T\beta) - E\left(R\right)).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
C(\mu - E(v_i^T\beta) - E\left(R\right)) &= \frac{COV\left(\sum_i t_i \widehat{\pi}_i^{-1}, R\right)}{\sqrt{Var(\sum_i t_i \widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}} \\
&\leq \frac{1}{\sqrt{Var(\sum_i t_i \widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}},
\end{aligned}
$$

and thus

$$
\left(\mu - E(v_i^T\beta)\right) - E\left(R\right) \leq \frac{1}{n\sqrt{Var(\sum_i t_i \widehat{\pi}_i^{-1})}\sqrt{Var\left(R\right)}} \longrightarrow 0
$$

as $n \longrightarrow 0$. Note that $E\left(\frac{1}{n}\sum_{i=1}^{n} v_i^T \widehat{\beta}^{IPDR}\right) = E(v_i^T\beta)$. Therefore, $\widehat{\mu}_{IPDR2}$ is an asymptotically unbiased estimator for μ. ■

Bibliography

[1] Bang H. and Robins J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**:692-972.

[2] Baker, F. and Kim, S-H. (2004) *Item Response Theory, 2nd ed.* Marcel Dekker, New York.

[3] Binder, D.A. (1983) On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

[4] Cannon G. & Einzig H. (1983) *Dieting Makes You Fat.* London: Century Publishing.

[5] Cassel, C.M., Särndal, C.E. and Wretman, J. (1977) *Foundations of Inference in Survey Sampling.* New York: Wiley.

[6] Clayton D., Spiegalhalter D., Dunn G., and Picles A. (1998) Analysis of longitudinal binary data from multiphase sampling *Journal of Royal Statistical Society B*, 60, 71-87

[7] Fisher R.A. (1935,1949) *The Design of Experiments.* Oliver & Boyd.

[8] Gelman, A., Carlin J.B., Stern, H.S., and Rubin D.B. (2004) *Bayesian Data Analysis.* London: Chapman and Hall.

[9] Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models.* New York: Chapman and Hall.

[10] Hill A.J. (2004) Does dieting make you fat? *British Journal of Nutrition*, 92, Suppl. 1, S15–S18

[11] Hirano, K. & Imbens, G.W.(2004) Continuous propensity scores. *In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Gelman, A. & Meng, X.L. (Eds.), 227–238, New York: Wiley.

[12] Holland, P.W. (1993) Statistics and causal inference *Journal of the American Statistical Association*, 81, 945-970

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

[14] Kang, J.D.Y. and Schafer, J.L.(2006) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, accepted by *Statistical Science*.

[15] Li, B. (2004) A lecture note for statistical inference (unpublished), the department of statistics, The Pennsylvania State Univeristy.

[16] Little, R.J.A. & An, H. (2004) Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14, 949–968.

[17] Liu, C. (2004) Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, A. Gelman and X.L. Meng (Eds.), 227-238, New York: Wiley.

[18] Neyman, J. (1923) On the application of probability theory to agricultural experiments. Translated in *Statistical Science*, 5, 465-480 (1990)

[19] Pearl,J. Causal diagrams for empirical research *Biometrika*, 82, 669-710.

[20] Qu, A., Lindsay, B.G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823-836.

[21] Robins JM. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6-10.

[22] Robins, J.M., Hernan, M. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560.

[23] Robins J.M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.

[24] Rotnitzky A, Robins JM. (1997) Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine*, 16, 81-102.

[25] Robins, J.M. and Rotnitzky, A. (2001) Comment on "Inference for semiparametric models: some questions and an answer," by P.J. Bickel and J. Kwon, *Statistica Sinica*, 11, 920–936.

[26] Rotnitzky A., Robins J.M., Scharfstein D. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statstical Association*, 93,1321-1339.

[27] Robins J.M., Rotnitzky A, Zhao L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

[28] Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

[29] Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

[30] Rosenbaum, P.R. (2002) *Observational Studies*. Springer-Verlag, New York.

[31] Rubin, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6, 34-58.

[32] Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

[33] Rubin, D.B. (2005) Causal inference using potential outcomes: Design, modeling, decisions, *Journal of the American Statistical Association,* 469, 322-331.

[34] Schafer, J.L. and Kang, J.D.Y. (2006) Average causal effects: A practical guide and simulated case study. Submitted to *P*sycological Methods.

[35] Särndal, C.E., Swensson, B. and Wretman, J. (1989) The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

[36] Särndal, C.E., Swensson, B. and Wretman, J. (2003) *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[37] Scharfstein, D.O., Rotnitzky, A., and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096-1120 (with Rejoinder, 1135-1146).

[38] Silverman, B.W. (1992) *Density Estimation for Statistics and Data Analysis* London: Chapman & Hall.

[39] Udry, J. R. (2003) The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Vita

Doh Yung Kang

Joseph Doh Yung Kang was born in Pohang, Southeast province of South Korea. From 1994 to 1996 he served in the U.S. Army as a Korean Augmented Solider. In 2000, he received the B.A. degree in division of Life Science from Korea University. In 2000, he happily married Deborah Kang. In 2003, he earned an M.S. degree in Biostatistics at University of Illinois at Chicago. In 2003, he enrolled in the Ph.D. program in Statistics at The Pennsylvania State University. Since 2003, he has been employed in the Methodology Center of The Pennsylvania State University as a Research Assistant. He is also an active member of the University Bible Fellowship Church.