ABSTRACT

Title of dissertation: CAUSAL INFERENCE WITH A CONTINUOUS TREATMENT AND OUTCOME: ALTERNATIVE ESTIMATORS FOR PARAMETRIC DOSE-RESPONSE FUNCTIONS WITH APPLICATIONS

Douglas Galagate, Doctor of Philosophy, 2016

Dissertation directed by: Dr. Joseph L. Schafer
Office of the Associate Director
for Research and Methodology
U.S. Census Bureau
and
Dr. Paul J. Smith
Mathematical Statistics Program
Department of Mathematics
University of Maryland

Causal inference with a continuous treatment is a relatively under-explored problem. In this dissertation, we adopt the potential outcomes framework. Potential outcomes are responses that would be seen for a unit under all possible treatments. In an observational study where the treatment is continuous, the potential outcomes are an uncountably infinite set indexed by treatment dose. We parameterize this unobservable set as a linear combination of a finite number of basis functions whose coefficients vary across units. This leads to new techniques for estimating the population average dose-response function (ADRF). Some techniques require a model for the treatment assignment given covariates, some require a model

for predicting the potential outcomes from covariates, and some require both. We develop these techniques using a framework of estimating functions, compare them to existing methods for continuous treatments, and simulate their performance in a population where the ADRF is linear and the models for the treatment and/or outcomes may be misspecified. We also extend the comparisons to a data set of lottery winners in Massachusetts. Next, we describe the methods and functions in our R package `causaldrf` using data from the National Medical Expenditure Survey (NMES) and Infant Health and Development Program (IHDP) as examples. Additionally, we analyze the National Growth and Health Study (NGHS) data set and deal with the issue of missing data. Lastly, we discuss future research goals and possible extensions.

CAUSAL INFERENCE WITH A CONTINUOUS TREATMENT AND OUTCOME: ALTERNATIVE ESTIMATORS FOR PARAMETRIC DOSE-RESPONSE FUNCTIONS WITH APPLICATIONS.

by

Douglas Galagate

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Dr. Joseph L. Schafer, Co-Advisor
Dr. Paul Smith, Co-Advisor
Dr. Eric Slud
Dr. Takumi Saegusa
Dr. Partha Lahiri
Dr. Benjamin Kedem

© Copyright by
Douglas Galagate
2016

Dedication

This work is dedicated to my parents, Rodolfo and Leticia,

and my siblings, Rolito and Mary Ann.

Acknowledgments

Many people have contributed in helping me complete this dissertation. I'm very lucky that Joe Schafer was my advisor. Joe was generous with his teaching, mentoring, and guidance. He was kind and encouraged me to be a better statistician, scientist, and person. Joe is someone I look up to.

I would also like to thank my co-advisor, Paul Smith for his advice throughout the years and guidance during the dissertation defense. Thanks to Eric Slud for helping me get internships at the FDA and U.S. Census Bureau. Without his help, I probably would not have finished the program. Thanks to Takumi Saegusa and Partha Lahiri for agreeing to serve on my thesis committee. Thank you to Benjamin Kedem for joining the committee on such short notice. Special thanks to Carlos Flores for collaborating with us on the lottery data set and for suggestions on Chapter 4 and also Keisuke Hirano and Guido Imbens for sharing the data. Thanks to Jennifer Hill for sharing data from the Infant Health and Development Program. I want to thank Donna Coffman for working with me on Chapter 6 and sharing the data on the NHLBI Growth and Health Study. Victor Chan, Tjalling Ypma, Edoh Amiran, Josh Delong, and others from WWU deserve a special mention.

My friends from the math department have made the experience fun and worthwhile: Neung Soo Ha, Ran Ji, Lucas Tcheuko, Nicholas Henderson, Kijoeng Nam, Jarvis Yu, David Shaw, Jong Jun Lee, Jiraphan Suntornchost, Shu Zhang, JoAnn Alvarez, Hisham Talukder, Sebastien Motsch, and Kwame Okrah. My friends from other departments reminded me of other research going on in the world: Gary

Paradee, Chaitanya Mudivarthi, Wonseok Hwang, and Adam Karcz. Friends from the home front deserve thanks: Rachel Sutcliffe, Tony Rogers, Cecilia Sequeira, Beth Kenyon, Sasan Ehsahee, and others. Thanks to Dorothea and Ed Mordan for letting me stay in their home during these years.

The mathematics department provided a good learning environment. Thanks to William Schildknecht for the teaching opportunities, Debbie Franklin for being a teaching mentor, and Esther David for sharing some fine Indian meals with me. Thanks to Celeste Regalado, Haydee Hidalgo, Linette Berry, Sharon Welton, Larry Washington, and Michael Laskowski for helping me.

I've learned a lot working at the U.S. Census Bureau with Tommy Wright, at the FDA with Antonio Paredes, and at USU with Tzu-Cheg (David) Kao. I would like to acknowledge friends and coworkers: Carolina Franco, Aaron Gilary, Jiashen You, James Livsey, Isaac Dompreh, Patrick Joyce, Darcy Morris, Ryan Janicki, Paul Massell, Derek Young, Josh Tokle, Ben Klemens, Marlow Lemons, Andrew Raim, Robert Ashmead, Chad Russell, Yves Thibaudeau, Osbert Pang, Gauri Datta, Thomas Mathew, Jerry Maples, Martin Klein, Chandra Erdman, Michael Leibert, Erica Magruder, Kelly Taylor, and especially Alisha Armas.

Most of all, I want to thank my family: Dad, Mom, Bro, Sis, Lisa, and Julia. Thanks for hanging out and keeping me positive. Talking to them almost daily kept us close. My cousins Lilibeth and Gerald Andrada provided fun family get-togethers and support on the east coast. Thanks to my family in California and the Philippines. I apologize to those I've inadvertently left out.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

$\phi(x)$	$(\sqrt{2\pi\sigma^2})^{-1}\exp[-(x-\mu)^2/(2\sigma^2)]$
$\Phi(x)$	$\mathcal{N}(0,1)$ cumulative distribution function
$\mathbb{1}(\cdot)$	Indicator function
$\mu(\cdot)$	$\mathrm{E}[Y_i(\cdot)]$ (average dose response function)
$\Delta\mathrm{PA}_{7,3}$	HAQ score change from year 3 to year 7
ADRF	Average dose-response function
APO	Average potential outcome
ACE	Average causal effect
ANCOVA	Analysis of covariance
BART	Bayesian additive regression trees
BMI	Body mass index
CART	Classification and regression trees
`causaldrf`	R package for ADRFs
CS	Common support subsample
DR	Double robust
GAM	Generalized additive model
GPS	Generalized propensity score
HAQ	Habitual activity questionnaire
HI	Hirano and Imbens (2004)
IPTW	Inverse probability of treatment weighting
ISMW	Inverse second moment weighting
MSE	Mean squared error
N	the number of elements in the data set
NGHS	National Growth and Health Study
OLS	Ordinary least squares
PA	Physical activity
PF	Propensity function
PFE	Prima facie estimator
PS	Propensity score
PO	Potential outcome
RHB	Robins et al. (2000)
SE	Standard error
SUTVA	Stable unit treatment value assumption
T_i	treatment value observed for unit i
$V(\cdot)$	Variance
$\mathbf{X}_i$	set of background covariates
$Y_i(t)$	potential outcome for unit i evaluated at t

Chapter 1: A framework for causal inference when the treatment is continuous

1.1 Introduction

Causal inference aims at the fundamental question of how changing the level of a cause or treatment can affect a subsequent outcome. Whether data analysts want to admit it or not, many analyses in behavioral, social, biomedical, and other fields of science are aimed at understanding causal relationships, even when the data or methods are not well suited to the task (Imbens and Rubin, 2015; Hernan and Robins, 2016; VanderWeele, 2015; Shadish et al., 2002).

Randomized experiments have long been regarded as the gold standard for inferring causal relationships. In many instances, however, conducting a randomized experiment is not feasible for reasons of timeliness, cost, or ethical constraints. In studying the health risks associated with tobacco use, for example, it is usually impractical to randomize cigarette smoking. For questions such as these, investigators must make do with observational data. Observational data, sometimes called quasi-experimental, or nonequivalent control group design data, arises when the treatment was not randomly assigned under the control of the investigator. Observational data

are available for many problems and sometimes are the only source of data for a given problem. The difficulty in using them for causal inference derives from the possibility of extraneous variables (confounders) that are correlated with both the the treatment and outcome, distorting the relationship between them.

There are different ways to formulate and address the problem of causal inference with observational data. In this dissertation, we will adopt the framework of potential outcomes (Neyman, 1923; Rubin, 1978). Thus far, most of the methods developed for potential outcomes suppose that the treatment variable is binary. In this dissertation, we focus on the relatively under-explored problem of estimating causal effects when the treatment is real-valued and continuous.

1.2 Potential outcomes notation

1.2.1 History

Potential outcomes are the responses that would be realized if different treatments were given to a unit. Some authors have called them counterfactuals (Greenland et al., 1999). An explicit notation for potential outcomes was invented by Neyman (1923) in the context of randomized experiments, but the idea seems to have been forgotten for more than half a century. Rubin (1978) reinvented the notation for observational studies and formulated causal inference as a problem of missing data (Little and Rubin, 2000; Rubin, 2005). Over time, the potential outcomes framework became known as the Rubin causal model or the Neyman-Rubin causal model (Holland, 1986).

Potential outcomes are not the only approach used in causal inference. Judea Pearl and his colleagues have developed an alternative system based on directed acyclic graphs (DAGs) and their notation called do-calculus. The do-calculus is a formalization of causal models that uses a do-operator called $do(x)$ which simulates physical interventions by removing certain functions from the model and replacing them with a constant, $X = x$ (Pearl, 2012). However, recent work makes steps to reconcile the different approaches (Richardson and Robins, 2013) and in some cases show their logical equivalence (Pearl, 2011). Another system for causal inference was suggested by Dawid (2000) but did not gain popularity.

1.2.2 Notation for a binary treatment

First we discuss the well studied problem of causal inference when the treatment is binary. Let T_i represent the value of the treatment applied to unit i. For example, $T_i = 1$ for the treatment group and $T_i = 0$ for control group. Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes for unit i under treatments 0 and 1, respectively. The observed outcome for unit i is denoted by $Y_i = Y_i(T_i)$, which can also be written as $T_i Y_i(1) + (1 - T_i)Y_i(0)$. The variable T_i can be thought of as a missing data indicator that tells us which potential outcome is seen and which one is hidden for the given unit.

In addition, let $\boldsymbol{X}_i$ denote a p-dimensional vector of covariates associated with unit i. We assume that the variables in $\boldsymbol{X}_i$ were realized before the treatment and are not causally affected by the treatment. These variables may be divided into four

main types: confounding variables (confounders) related to both the treatment and outcome, prognostic variables related only to the outcome, variables related only to treatment, and nuisance covariates related to neither treatment nor outcome. Confounders play a major role in causal inference and can have a large influence on the estimates.

Since only one outcome is observed, estimating causal effects can be difficult. Only the available observed outcomes are used in the estimation process. When estimating causal effects, the estimands relate to a group of units and unit-level estimates are unavailable. In this dissertation, the focus is on methods using the potential outcomes framework applied to observational data containing groups of units.

1.2.3 Causal effects

Rubin (1978) defines the causal effect of a treatment on unit i as $Y_i(1) - Y_i(0)$. In certain contexts, it may be useful to characterize the unit-level causal effect as a ratio, $Y_i(1)/Y_i(0)$, or some other measure of discrepancy between the potential outcomes. Unfortunately, only one of the potential outcomes, $Y_i(1)$ or $Y_i(0)$, is seen for unit i, so the individual-level causal effect is inherently unobservable. Holland (1986) calls this "the fundamental problem of causal inference." Although unit-level causal effects are sometimes of interest, the target of statistical inference is usually taken to be the average causal effect (ACE) for a given population,

$$ACE = \mu(1) - \mu(0) = E[Y(1)] - E[Y(0)]. \tag{1.1}$$

Note that the expectations in (1.1) are taken with respect to the distributions of $Y_i(1)$ and $Y_i(0)$ over the same population of units. A key feature of causal inference is that it compares outcomes for different treatments that are hypothetically applied to the same units. The well known saying "correlation is not causation" is a necessary qualification for analyses in which the groups receiving different treatments are not directly comparable. In a randomized experiment, comparability is guaranteed by the random assignment of treatments to units. In an observational study, the groups may differ with respect to variables in $\boldsymbol{X}_i$ or other characteristics that have not been measured, and some method of statistical adjustment based on $\boldsymbol{X}_i$ (along with assumptions that cannot be verified) will be needed to consistently estimate the ACE.

1.2.4 No causation without manipulation

In an observational study, the treatments are not assigned to units in a controlled manner. Nevertheless, many authors have cautioned that causal effects should not be estimated without first envisioning a hypothetical experiment in which the treatments applied to units could be altered by an intervention. For example, the effect of a new drug pill on blood pressure compared to the standard pill is well-defined because each study participant could conceivably be induced to take either one. In contrast, race or gender are not usually regarded as having causal effects because it is difficult to imagine precise real-life interventions to change someone's race or gender (Imbens and Rubin, 2015).

Another example discussed by Gelman and Hill (2006) (pp. 186-187) concerns the effects of single motherhood on children's well being. The treatment variable can be defined as $T_i = 1$ if a mother is single and $T_i = 0$ if she is married. However, one can imagine various interventions that might alter values of T_i, such as changes in tax laws, marriage encouragement programs for unwed parents, and so on. These potential treatments would impact families in varying ways, and there is no compelling reason to believe that their effects on well being of children would be identical. Results from an observational study of the relationship between single motherhood and children's outcomes would need to be interpreted in light of the possible ways that the treatment variable might be manipulated in real life.

1.2.5 The missing-data perspective

In any actual study, simultaneously observing the behavior of a particular unit under different treatments is not possible. It is sometimes helpful to cast causal inference as a problem of incomplete data, where one potential outcome is seen and the other is missing for each unit.

A visual representation of the missing data perspective for the binary treatment setting is shown in Figure 1.1. In this picture, the study units have been ordered so that the top half of the data set represents units that received $T_i = 1$ and the bottom half represents units that received $T_i = 0$. If there were no missing

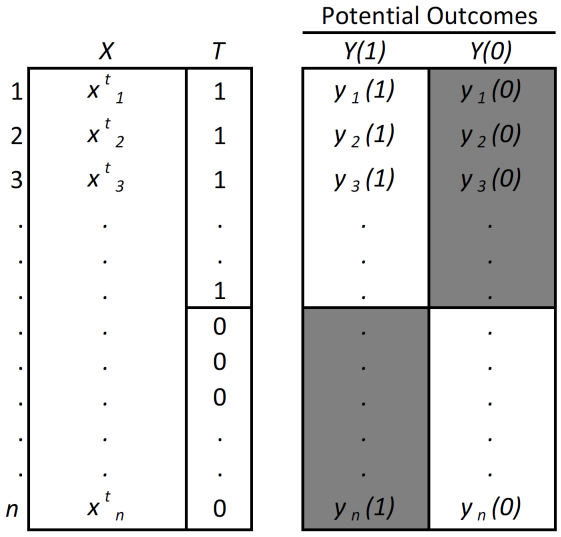

Figure 1.1: Potential outcomes representation of the binary treatment setting. Shaded regions are not observed.

values, the ACE could be consistently estimated under very general conditions by

$$\widehat{ACE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)). \tag{1.2}$$

Of course, because not all potential outcomes are observable, this quantity cannot be computed in any actual study. However, it is sometimes illuminating to compare the theoretical properties of an actual estimator to that of (1.2), because there may be limiting conditions under which the method approaches the behavior of (1.2) (Schafer and Kang, 2008).

The notation developed thus far can be easily extended to handle more than two treatments. The situation with three treatment levels is shown in Figure 1.2. In this setting, there are three potential outcomes for each unit, $Y_i(0)$, $Y_i(1)$, and $Y_i(2)$, and three causal comparisons: $Y_i(1) - Y_i(0)$, $Y_i(2) - Y_i(0)$, and $Y_i(2) - Y_i(1)$. The observed outcome can be written as $Y_i = Y_i(T_i) = \sum_{t=0}^{2} \mathbb{1}(T_i = t) \cdot Y_i(t)$.

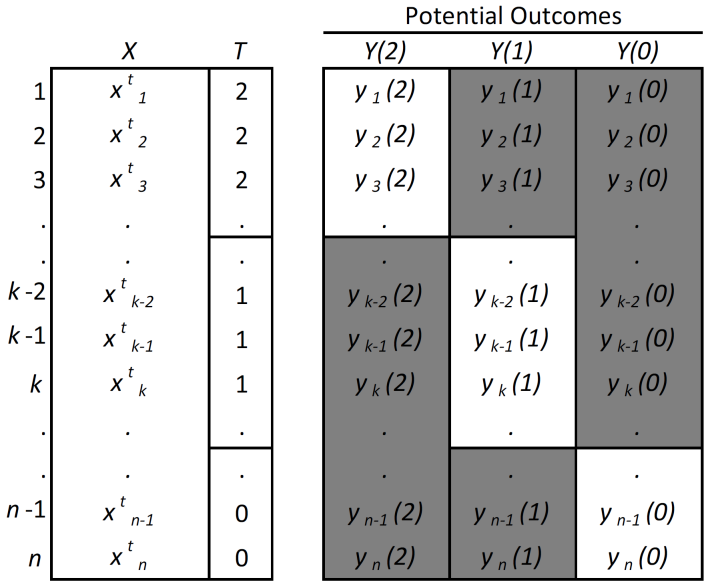

Figure 1.2: Potential outcomes representation of the multiple treatment setting of three treatment options. Shaded regions are not observed. One potential outcome is realized, but the others can be regarded as missing.

1.2.6 Notation for the continuous treatment

From a notational standpoint, the framework of potential outcomes is easily adapted to situations where the treatment variable is continuously distributed. Suppose now that T_i takes values within a real interval $\mathcal{T} = [t_{min}, t_{max}]$. The potential outcomes are now an uncountably infinite set $\mathcal{Y}_i = \{Y_i(t) : t \in \mathcal{T}\}$. The observed outcome can still be written as $Y_i = Y_i(t)$, and the causal effect for unit i of moving from treatment dose t to t^* is $Y_i(t^*) - Y_i(t)$. This unit-level effect is unobservable, but under certain conditions, we may be able to construct a reasonable estimate of the population average effect of moving from t to t^*, $E(Y_i(t^*)) - E(Y_i(t))$. All such comparisons for $t, t^* \in \mathcal{T}$ are contained in the average dose-response function (ADRF), which we write as

$$\mu(t) = E(Y_i(t)). \tag{1.3}$$

In this thesis, we focus on methods for estimating $\mu(t)$ over the domain $t \in \mathcal{T}$. Other estimands that may be of interest are mentioned in the next section. Data for the continuous treatment setting are shown in Figure 1.3.

Although the potential-outcomes notation extends easily to a continuous treatment, methods for estimating the ACE in the binary case do not easily adapt to estimation of $\mu(t)$ in the continuous case, for reasons that we describe later.

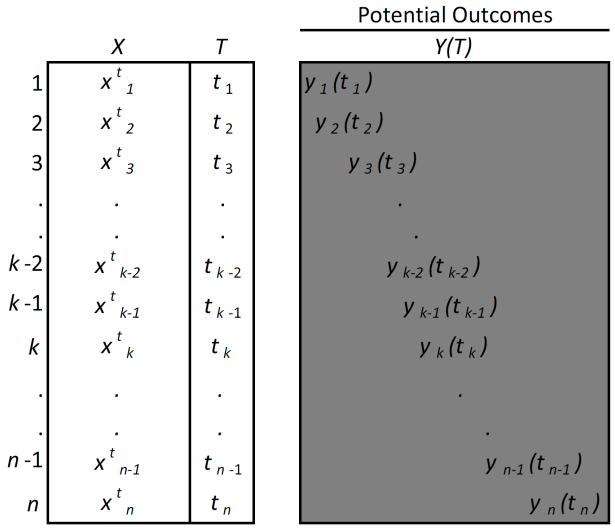

Figure 1.3: In the continuous treatment setting, only one of an uncountably infinite number of potential outcomes is observed for each unit.

1.2.7 Causal inference versus regression for a continuous treatment

The potential-outcomes notation helps us to clarify the difference between causal inference and regression analysis. Consider the dose-response behavior for a sample of individual units as illustrated in Figure 1.4(a). This hypothetical example shows each individual unit having its own potential-outcomes path describing the response at every possible treatment level. If we look at the vertical strip at dose level t_1, we can approximate the ADRF at t_1 with $\mu(t_1) = E(Y_i(t_1))$ by averag-

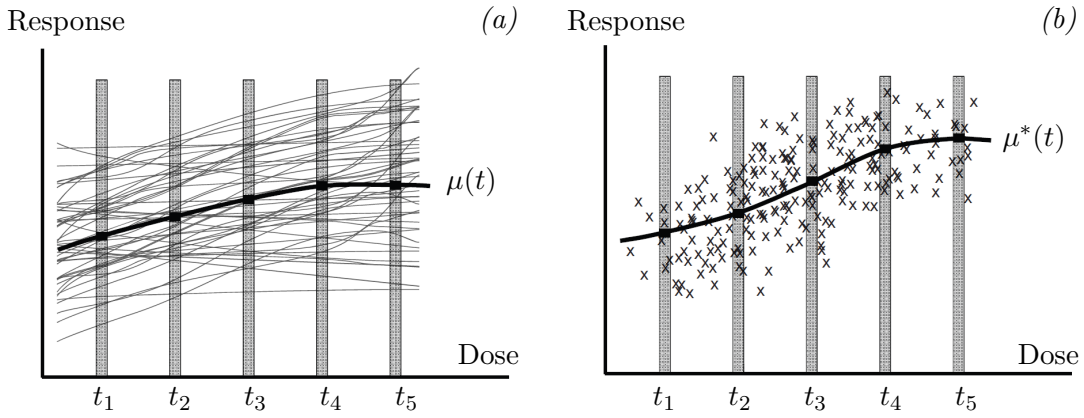

Figure 1.4: (a) Average dose-response function, $\mu(t) = E(\, Y_i(t)\,)$, and (b) regression relationship between treatment and observed outcomes, $\mu^*(t) = E(\, Y_i(t) \mid T_i = t)$.

ing the curves in the vertical strip. Computing these average values at different dose levels $t_1, t_2, t_3, \ldots$ traces out the ADRF. In contrast, fitting a curve to the observed data points (T_i, Y_i) will not necessarily approximate the ADRF. Figure 1.4(b) shows the observed data with a curve traced out by averaging at five example points $t_1, t_2, \ldots, t_5$. Computing the observed averages at $t_1, t_2, \ldots, t_5$ traces out the regression curve $\mu^*(t) = E(Y_i(t)|T_i = t) = E(Y_i|T_i = t)$. If T_i were independent of $Y_i(t)$ at each t, such as in a randomized experiment, then the units at each vertical strip t_k would be a representative sample from the full population. In that case, $\mu^*(t)$ and $\mu(t)$ would coincide. In nonrandomized or observational studies, however, $\mu^*(t)$ may not equal $\mu(t)$. In general, a regression curve does not have a causal interpretation. Causal effects are comparisons among potential outcomes in the same population of units, but the regression curve may represent different populations at different values of t.

1.2.8 Contrasting causal inference and conventional dose-response modeling

Dose-response modeling, also known as exposure-response modeling, is a general term used in pharmacology, environmental health, and other areas of life science (Wang, 2015; Dominici et al., 2002; Altshuler, 1981). It commonly refers to situations where the predictor is a continuously valued level of exposure (e.g., amount of drug taken, concentration of potential harmful substance), and the response is an outcome of interest that is thought to be causally related to the exposure (e.g. probability of a certain reaction, time to some event). Data for dose-response modeling often come from experiments where the treatment or dose is under direct control of the investigator. In some cases they can assign multiple doses to the same subject. In those contexts, dose-response modeling may be carried out using standard linear or nonlinear regression, generalized linear modeling, survival analysis, and longitudinal modeling.

This dissertation addresses the same general question as conventional dose-response modeling: How does changing the level of a treatment variable causally affect the mean of an outcome variable? However, causal inference with a continuous treatment is different from conventional dose-response modeling in these ways:

1. It deals with non-experimental situations where the dose was not assigned by the investigator, and thus confounders may be present.

2. It explicitly uses the framework of potential outcomes.

3. It uses data that contains only one dose level for each subject.

4. It limits attention to situations where the response is numeric and continuous.

1.2.9 Causal estimands

Depending on the problem, different causal quantities may be of interest. As mentioned, we will focus on the population ADRF, $\mu(t) = E[Y_i(t)]$, but some problems may suggest other estimands. Quantities such as $E(Y_i(t)) - E(Y_i(s))$, $(d/dt)(\mu(t))$, $E[Y_i(t)]/E[Y_i(s)]$, and $\log(E[Y_i(t)]) - \log(E[Y_i(s)])$ are all determined by the ADRF, so if we estimate the ADRF we have estimated these as well.

In some studies, it is difficult to imagine a certain dosage being applied to a whole population. In the binary-treatment setting, we are sometimes interested in $E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1]$, the average causal effect among the treated or $E[Y_i(1)|T_i = 0] - E[Y_i(0)|T_i = 0]$, the average causal effect among the controls. Likewise, in the continuous-treatment setting, we may wish to estimate $\{E[Y_i(t + \delta)|T_i = t] - E[Y_i(t)|T_i = t]\}$ for various values of δ.

1.3 Common assumptions

1.3.1 The Stable Unit Treatment Value Assumption

To produce consistent estimates of causal quantities, we will make a few assumptions throughout this thesis. First, we assume that treatments applied to one unit will not affect the outcome for any other unit. For example, if you take a drug

for lowering blood pressure, we assume that this will not affect the blood pressure of anyone else. This assumption is called the Stable Unit Treatment Value Assumption (SUTVA) (Rosenbaum and Rubin, 1983, 1984), which means that units do not interact with each other, or that there is no interference between units. SUTVA also includes the idea that there is only one version of each treatment value; for example, taking a new drug for blood pressure means that the pills are identical in dosage when making comparisons.

1.3.2 Positivity

Positivity is assumed for all units. This means each unit must have a chance of receiving each treatment level. When the treatment variable is discrete, $P(T_i = t|\boldsymbol{X}_i) \in (0,1)$ for $t \in \{t_1, t_2, \ldots, t_n\}$ when T_i is binary or discrete; the probability of receiving a particular treatment level cannot be 0 or 1. When the treatment is continuous, we make the analogous assumption that each unit in the population has some possibility of receiving any dose of the treatment. That is, $P(T_i \in \mathcal{T}_i|\boldsymbol{X}_i) > 0$ for every $\boldsymbol{X}_i$ in the population and every set $\mathcal{T}_i \subset \mathcal{T}$ with positive measure. In other words, in any part of the covariate space, it must be possible to receive every level of the treatment. Without this assumption, some portions of the ADRF may become nonestimable.

1.3.3 Unconfoundedness

Of utmost importance in causal inference is the treatment mechanism underlying the data. In an observational study, the distribution of T_i is beyond the control of the investigator and largely unknown. To make headway in estimating causal effects, investigators will typically assume that any relationship between potential outcomes and the treatment can be fully explained by measured covariates. For binary treatments, it is often assumed that

$$T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} | \boldsymbol{X}_i \tag{1.4}$$

for all $i = 1, \ldots, N$, which is called strong ignorability (Rosenbaum and Rubin, 1983, 1984). For the continuous treatment setting, the analogous assumption is

$$T_i \perp\!\!\!\perp \mathcal{Y}_i | \boldsymbol{X}_i \tag{1.5}$$

for all $i = 1, \ldots, N$, T_i, where $\mathcal{Y}_i = \{Y_i(t) : t \in \mathcal{T}\}$. Some authors have made the weaker assumption that $T_i \perp\!\!\!\perp Y_i(t) | \boldsymbol{X}_i$ for all $t \in \mathcal{T}$ (Hirano and Imbens, 2004). Both of these assumptions are statements about the joint distribution of $\mathcal{Y}_i, T_i,$ and $\boldsymbol{X}_i$ at the unit level. An alternative definition of confounding at the population level is given by Greenland and Robins (1986).

Strong ignorability implies that all confounders have been measured and are present in $\boldsymbol{X}_i$. In real-world applied problems, ignorability is often violated due to

unmeasured confounders, but we will not consider that possibility here. As a way to resolve this problem, sensitivity analyses can be conducted to understand the effect of possible omitted covariates in the estimation process (Rosenbaum, 2002).

1.3.4 Other assumptions

For some of our techniques, we will suppose that $Y_i(t)$ is continuous in t. We will also suppose that the triplets $(\boldsymbol{X}_i, T_i, \mathcal{Y}_i)$ are independent and identically distributed for all units $i = 1, \ldots, N$, in the population and in the sample. Many of the methods in this thesis can be extended to complex sampling designs, but we defer those extensions to future research. In some real world studies, treatments are not applied to individual units but to clusters of units (e.g., students in a classroom). Methods for cluster-level treatments are beyond the scope of this thesis. However, in some studies, we can circumvent the problem of cluster-level treatments by aggregating the responses and covariates to the cluster level, regarding the clusters as the units of analysis.

1.3.5 The observed data

The observed data for each unit in the sample consist of the covariate vector $\mathbf{X}_i$, the treatment T_i, and the observed outcome $Y_i = Y_i(T_i)$. We will investigate methods for estimating the ADRF $\mu(t) = E(Y_i(t))$ from the observed data $(\boldsymbol{X}_i, T_i, Y_i)$, $i = 1, \ldots, N$. In many real world applications, some portions of $(\boldsymbol{X}_i, T_i, Y_i)$ may be missing for some units. We will not consider those types of missing data problems

here.

1.3.6 Modeling assumptions

In addition to the assumptions already mentioned, causal inference will leverage the information in the covariates $\boldsymbol{X}_i$ by applying various types of models. Before the 1980s, it was common for analysts to regress Y_i on $(\mathbf{X}_i, T_i)$ and then use that model to estimate treatment effects. For example, one might fit a linear regression with main effects for $\boldsymbol{X}_i$ and T_i, and interpret the coefficient for T_i as an estimated causal effect. One drawback of this outcome model approach is that without the rigor of potential outcomes, the estimated values are not necessarily causal quantities. The relationship between the estimated regression function $E(Y_i|T_i, \boldsymbol{X}_i)$ and the ADRF is far from clear.

In the early 1980s, Rosenbaum and Rubin (1983, 1984) pioneered another approach to estimate causal effects by modeling the treatment mechanism. They model the treatment T_i given the covariates $\mathbf{X}_i$, which is now called a propensity model. In the binary-treatment setting, the treatment assignment is typically described by a binary (e.g. logistic) regression for predicting T_i given covariates obtained from $\boldsymbol{X}_i$. In the continuous-treatment setting, various other models (e.g. linear regression) could used. Models for T_i are an important building block for many types of causal estimators.

Another strategy combines a propensity model with additional assumptions about how $Y_i(t)$ may vary with $\boldsymbol{X}_i$. By combining models for the treatment and the

potential outcomes, the hope is to take advantage of some of the positive aspects of each one while mitigating the possible negative effects of model failure.

1.3.7 Why the continuous treatment setting is more complicated than the binary case

From a notational standpoint, the extension of a pair of potential outcomes $\mathcal{Y}_i = \{Y_i(0), Y_i(1)\}$ to a set indexed by a continuous variable $\mathcal{Y}_i = \{Y_i(t) : t \in \mathcal{T}\}$ seems very straightforward. However, methods for estimating the average causal effect $E(Y_i(1)) - E(Y_i(0))$ in the binary case (see, for example Schafer and Kang (2008)) are more numerous and have received far more attention than methods for estimating an ADRF $\mu(t) = E(Y_i(t))$ for a continuous treatment. The continuous-treatment problem is relatively poorly studied and is significantly more complicated than the binary one. The reasons for this will be made clear in the chapters ahead, but here we provide a thumbnail sketch.

One common approach for the binary case is to weight the observed outcomes to resemble a sample of potential outcomes from the full population. That is, we can apply weights to the observed values of $Y_i(T_i)$ from the sample units having $T_i = 1$ to estimate $E(Y_i(t))$, and we can apply weights to the observed values of $Y_i(0)$ from the sample units having $T_i = 0$ to estimate $E(Y_i(0))$. These weights are derived from a propensity model that predicts $P(T_i = 1|\boldsymbol{X}_i)$ and $P(T_i = 0|\boldsymbol{X}_i)$.

Another common approach is to predict the missing potential outcomes themselves with a pair of regression models. For example, we can regress the observed

values of $Y_i(1)$ on $\boldsymbol{X}_i$ for units with $T_i = 1$, and use the fitted model to predict the unseen values of $Y_i(1)$ when $T_i = 0$. Similarly, we can regress the observed values of $Y_i(0)$ on $\boldsymbol{X}_i$ for units with $T_i = 0$, and use the fitted model to predict the unseen values of $Y_i(0)$ when $T_i = 1$.

Neither of these simple strategies immediately generalizes to a continuous treatment. Suppose we want to estimate $\mu(t) = E(Y_i(t))$ for a specific value of t. Because T_i is a continuous random variable, there may be no units in the sample with $T_i = t$, and thus no observed values of $Y_i(t)$ to reweight. Similarly, there may be few or no observed values of $Y_i(t)$ to use in a regression model for predicting the missing values of $Y_i(t)$ when $T_i \neq t$. To make headway using either of these approaches (weighting or prediction), we will have to somehow borrow strength from units with values of T_i in the neighborhood of t, which may necessitate additional assumptions (continuity, smoothness) about the ADRF or the unit-level stochastic process $Y_i(t)$.

1.4 Motivating examples

1.4.1 A simulated example

1.4.1.1 Treatment and potential outcomes

This example, which will be used throughout Chapter 3, uses simulated data. One advantage of simulated data is that the true causal quantity is observed; this allows for a fair evaluation of the methods. The essential features of this example

are that the unit-level potential outcome paths are linear, where the treatment is correlated with the slope and intercept, that the covariates are measured, but that the relationships among the covariates, treatment, and potential outcomes could be misspecified.

We now describe the example in more detail. The potential outcome paths are linear, $Y_i(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t)$, with $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^\top$ and $\boldsymbol{b}(t) = (1, t)^\top$. The population for $(\theta_{i1}, \theta_{i2}, T_i)^\top$ is multivariate normal with mean vector $(\xi_1, \xi_2, \kappa)^\top = (50, 0, 12)^\top$ and covariance matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{12} & \omega_{22} & \omega_{23} \\ \omega_{13} & \omega_{23} & \omega_{33} \end{bmatrix} = \begin{bmatrix} 51.0 & 3.80 & 5.92 \\ 3.80 & 0.55 & 0.51 \\ 5.92 & 0.51 & 2.02 \end{bmatrix}.$$

A scatterplot of $Y_i = \theta_{i1} + \theta_{i2} T_i$ versus T_i for a sample of $N = 200$ units is shown in Figure 1.5, along with the $Y_i(t)$'s and $\mu(t) = E(Y_i(t))$. The ADRF is constant, but the nonzero values for ω_{13} and ω_{23} induce a strong positive correlation between T_i and Y_i.

Figure 1.5 also shows the population regression curve $\mu^*(t) = E(Y_i \mid T_i = t)$. By well known properties of the multivariate normal distribution, $\boldsymbol{\theta}_i$ given $T_i = t$ is bivariate normal with $E(\theta_{i1} \mid T_i = t) = \xi_1 + (\omega_{13}/\omega_{33})(t - \kappa)$, $E(\theta_{i2} \mid T_i = t) = \xi_2 + (\omega_{23}/\omega_{33})(t - \kappa)$, $V(\theta_{i1} \mid T_i = t) = \omega_{11} - \omega_{13}^2/\omega_{33}$, $V(\theta_{i2} \mid T_i = t) = \omega_{22} - \omega_{23}^2/\omega_{33}$, and $\mathrm{Cov}(\theta_{i1}, \theta_{i2} \mid T_i = t) = \omega_{12} - \omega_{13}\omega_{23}/\omega_{33}$. It follows that $Y_i = \theta_{i1} + \theta_{i2} T_i$ given

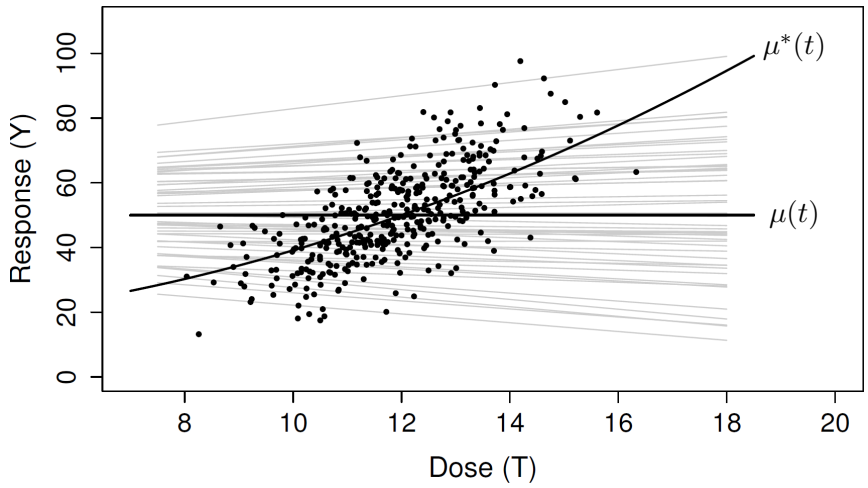

Figure 1.5: Simulated sample of $N = 200$ observed points (T_i, Y_i), with representative potential-outcome paths (gray lines), average causal dose response function $\mu(t)$, and regression curve $\mu^*(t)$.

$T_i = t$ is normal with

$$
\begin{aligned}
E(Y_i \mid T_i = t) &= \left(\xi_1 + \frac{\omega_{13}}{\omega_{33}}\kappa\right) + \left(\xi_2 + \frac{\omega_{13}}{\omega_{33}} - \frac{\omega_{23}}{\omega_{33}}\kappa\right) t + \left(\frac{\omega_{23}}{\omega_{33}}\right) t^2, \\
V(Y_i \mid T_i = t) &= \left(\omega_{11} - \frac{\omega_{13}^2}{\omega_{33}}\right) + 2\left(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}}\right) t + \left(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}\right) t^2.
\end{aligned}
$$

The curvature in $\mu^*(t)$ would vanish if the treatment were uncorrelated with the random slopes ($\omega_{23} = 0$). Even in that case, however, the slopes of $\mu^*(t)$ and $\mu(t)$ would differ unless the treatment were also uncorrelated with the random intercepts ($\omega_{13} = 0$).

1.4.1.2 Covariates

To create these data, we generated eight variables $\boldsymbol{A}_i^* = (A_{i1}^*, \ldots, A_{i8}^*)^\top$ for each unit from $\boldsymbol{A}_i^* \sim N(\boldsymbol{0}, \boldsymbol{I})$, and then we generated $\boldsymbol{\theta}_i$ and T_i independently given

$\boldsymbol{A}_i^*$ from $\boldsymbol{\theta}_i \mid \boldsymbol{A}_i^* \sim N(\boldsymbol{\nu}+\boldsymbol{\Gamma}^\top \boldsymbol{A}_i^*, \boldsymbol{\Sigma})$ and $T_i \mid \boldsymbol{A}_i^* \sim N(\tau+\boldsymbol{\delta}^\top \boldsymbol{A}_i^*, \lambda^2)$ with $\boldsymbol{\nu} = (50, 0)^\top$,

$$\boldsymbol{\Gamma} = \begin{bmatrix} 5 & 2 & -1 & 1 & 2 & 3 & -1 & 1 \\ .4 & .3 & -.2 & .1 & .3 & .2 & .1 & -.1 \end{bmatrix}^\top, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 5 & -.1 \\ -.1 & .1 \end{bmatrix},$$

$\tau = 12$, $\boldsymbol{\delta} = (1, 1/2, 1/3, 1/4, 0, 0, 0, 0)^\top$ and $\lambda^2 = 0.6$. This represents a situation where the dose-response relationship is distorted by covariates whose associations with the treatment and outcomes are strong. Variables $A_{i1}^*, \ldots, A_{i4}^*$ are confounders, associated with both $\boldsymbol{\theta}_i$ and T_i, whereas $A_{i5}^*, \ldots, A_{i8}^*$ are prognostic variables, associated with $\boldsymbol{\theta}_i$ but not T_i.

In a real application, it is unlikely that an analyst would have set of variables that fully account for the associations between the potential outcomes and the treatment and whose relationships to $\boldsymbol{\theta}_i$ and T_i are perfectly linear. To make this example more realistic, we imagine that the A_{ij}^*'s are hidden from view, and instead

the analyst sees

$$
\begin{aligned}
A_{i1} &= -4.975\,\exp((A^*_{i1}+1)/6) + 5.960, \\
A_{i2} &= 3\log(A^*_{i1}/5 + \exp(A^*_{i2})/10 + 2), \\
A_{i3} &= 4\log(A^*_{i1}+8) + (A^*_{i3}+6)^{1.8}, \\
A_{i4} &= (A^{*2}_{i1}/(A^*_{i4}+10))^{0.25}, \\
A_{i5} &= \Phi(\,(A^*_{i3}+A^*_{i4}+2A^*_{i5})/\sqrt{6}\,), \\
A_{i6} &= A^{*2}_{i1}+A^{*2}_{i2}+A^{*2}_{i3}+A^{*2}_{i4}+A^{*2}_{i5}+A^{*2}_{i6}, \\
A_{i7} &= 1 \text{ if } (A^*_{i2}+A^*_{i7}) > 1 \text{ and } 0 \text{ otherwise}, \\
A_{i8} &= 1 \text{ if } (A^*_{i2}-A^*_{i7}) \leq -1 \text{ and } 0 \text{ otherwise},
\end{aligned}
$$

and where $\Phi(\cdot)$ denotes the $N(0,1)$ cumulative distribution function. These transformations were chosen rather arbitrarily, but we strove to create a realistic situation where the predictors available to the analyst were non-normal, intercorrelated, and related to T_i and Y_i in a nonlinear fashion. The transformations produce skewed, bounded and binary covariates whose relationships to θ_{i1}, θ_{i2} and T_i are moderately strong and mildly nonlinear. The best linear functions of the A^*_{ij}'s explain 90% of the variance in θ_{i1}, 82% of the variance in θ_{i2}, and 70% of the variance in T_i, whereas the corresponding figures for the A_{ij}'s are 68%, 66%, and 65%, respectively.

In Chapter 3, we present techniques to estimate the ADRF and simulate their performance over random samples of $N = 200$ and $N = 1,000$. The case $N = 200$ represents a condition where large-sample approximations may not be accurate,

and where bias may be less important than variability; with $N = 1,000$, asymptotic arguments should work well, and the major concern becomes bias due to model misspecification. In all of our analyses, we will suppose that the ADRF is linear, $\mu(t) = \xi_1 + \xi_2 t$, and we will see how well the estimate of ξ_2 reproduces the true value of zero.

1.4.2 Lottery data

This data set, which was previously analyzed by Imbens et al. (2001) and Hirano and Imbens (2004), contains information from a survey on a group of $N = 237$ Massachusetts lottery winners. Variables include information about their salary before and after winning the lottery, winning prize amount, education, savings, spending, gender, and age. In our analyses, we focus only on people who won the lottery, and want to know how lottery prize amount affects their earned income six years after winning. Hirano and Imbens (2004) analyzed the data and estimated the derivative of the ADRF. In Chapter 4, we revisit the lottery data set and apply a new set of techniques.

1.4.3 Simulation from Hirano and Imbens (2004) and Moodie and Stephens (2012)

In this example, the covariates are exponentially distributed, the treatment is an exponential RV that depends on the covariates, and the outcome is normally distributed with a mean that depends on the covariates and has a standard deviation

of one.

To be more precise, let $Y_1(t)|X_1, X_2 \sim \mathcal{N}\left(t + (X_1 + X_2)\exp[-t(X_1 + X_2)], 1\right)$, let X_1, X_2 be unit exponentials, and let $T_1 \sim \exp(X_1 + X_2)$. The ADRF can be calculated by integrating out the covariates analytically (Moodie and Stephens, 2012),

$$\mu(t) = E(Y_i(t)) = t + \frac{2}{(1+t)^3}. \tag{1.6}$$

We can compare the performance of different estimators to this true ADRF. This example is included in Chapter 5.

1.4.4 National Medical Expenditure Survey data

The National Medical Expenditure Survey (NMES) contains information about smoking amount and the cost of medical care. The 1987 medical costs are verified by multiple interviews and other data from clinicians and hospitals. From Johnson et al. (2003):

> The 1987 National Medical Expenditure Survey (NMES, US Department of Health and Human Services, Public Health Service, 1987) provides data on annual medical expenditures and disease status for a representative sample of the U.S. civilian, non-institutionalized populations.

In the early 2000s, the NMES data was important for the tobacco litigation cases on the effect of cigarettes on medical expenditures. The R package `causaldrf` includes these data. In Chapter 5, we briefly analyze the relationship of smoking amount on medical expenditures.

1.4.5 Infant Health and Development Program data

The last example in Chapter 5, the Infant Health and Development Program (IHDP), is described by Gross (1992):

> The Infant Health and Development Program was a collaborative, randomized, longitudinal, multisite clinical trial designed to evaluate the efficacy of comprehensive early intervention in reducing the developmental and health problems of low birth weight, premature infants. An intensive intervention extending from hospital discharge to 36 months corrected age was administered between 1985 and 1988 at eight different sites. The study sample of infants was stratified by birth weight (2,000 grams or less, 2,001-2,500 grams) and randomized to the Intervention Group or the Follow-Up Group.

In this study, even though families are randomly selected for intervention, we restrict our analysis to those selected for the treatment group. These families choose the amount of days they attend the child development centers and this makes the data set, for practical purposes, an observational data set. We apply our methods to this subset of the data to estimate the relationship of days spent at the child development centers on the cognitive benefit to children.

1.4.6 National Growth and Health Study data

In Chapter 6, we analyze the National Growth and Health Study (NGHS) which is a multicenter, 10-year longitudinal study that covers 2379 girls from the ages of 9-10 through 18-19. In addition to the activity measures, the NGHS data set contains many other covariates: demographic, history, physical measurements, biochemical determinations, diet, physical activity, and psychosocial. One goal is to understand the role of changing physical activity from year 3 to year 7 in the study and how this change affects body weight and activity at year 10.

1.5 Looking ahead

This dissertation explores the understudied problem in causal inference of the continuous treatment. Chapter 2 reviews the current available methods. These include methods from Imai and van Dyk (2004), Hirano and Imbens (2004), Hill (2011), Flores et al. (2012).

Chapter 3 introduces new ideas and methods addressing parametric dose response functions when the true ADRF has a parametric form. The continuous treatment problem can be approached by parameterizing the curve as a linear combination of a finite number of basis functions whose coefficients vary across the units. The ADRF is estimated by averaging over all the units.

Chapter 4 revisits the lottery data set, reanalyzes the data, performs diagnostics, and applies a new set of techniques to estimate the ADRF. The results are compared and contrasted with Hirano and Imbens (2004) and Bia et al. (2014).

Chapter 5 contains the `causaldrf` R package vignette (Galagate and Schafer, 2015b). Examples include simulated data from Hirano and Imbens (2004) and Moodie and Stephens (2012), NMES data from Imai and van Dyk (2004), and the IHDP data from Hill (2011). We illustrate different methods and demonstrate the flexibility of the `causaldrf` R package.

Chapter 6 provides an in-depth example with the NGHS data set. The NGHS has not yet been analyzed by using causal inference methods. This chapter performs the analysis and provides step-by-step commentary and suggestions for how to apply causal inference techniques to a real data set that includes missing data.

Chapter 7 wraps up with the conclusion and possible future work.

Chapter 2: A review of existing causal inference methods with a continuous treatment

2.1 Defining the problem

To review our notation, let N be the number of sample units in the data set. The treatment received by unit i is T_i, which takes values in a real interval $\mathcal{T} = [t_{min}, t_{max}]$. We imagine a set of potential outcomes for each unit, $\mathcal{Y}_i = \{Y_i(t) : t \in \mathcal{T}\}$, and the observed outcome is $Y_i = Y_i(T_i)$. We call $Y_i(t)$ the individual-level dose response function. The variable T_i has a positive density with support $\mathcal{T} = [t_{min}, t_{max}]$

The target estimand is, $\mu(t) = E[Y_i(t)]$, the average dose response function (ADRF) for the population of interest. In some cases, the ADRF might refer to a subset of the whole population. Another estimand of interest is mentioned by Hill (2011), namely $E[Y_i(t) - Y_i(t_0)|T_i = t]$. This quantity compares the observed outcome in the population receiving dose t to what the outcome would have been at another dose.

We also have available $\mathbf{X}_i$, a vector of background covariates. The data available to us for estimating the ADRF are $(\mathbf{X}_i, T_i, Y_i = Y_i(T_i))$ for $i = 1, \ldots, N$. We assume that $(\mathbf{X}_i, T_i, Y_i)$ are defined in a common probability space, that T_i is continuously distributed with respect to Lebesgue measure in $\mathcal{T}$, and that $Y_i(T_i)$ is a well defined random variable. No parametric assumptions are made on $\mathbf{X}_i$. If modeling of $\mathbf{X}_i$ is necessary, we use an empirical distribution.

Statisticians have used three broad approaches for estimating causal effects: strategies based on modeling the outcome, strategies based on modeling the treatment, and dual modeling strategies (Schafer and Kang, 2008). These methods are well understood in the binary treatment setting, but not when the treatment is continuous. Some methods in the binary treatment setting can easily generalize to the continuous setting, while others do not have obvious extensions. Because the continuous treatment problem has not been well studied, it is helpful to review the methods that have been applied to estimating the average causal effect (ACE) in the binary treatment case.

When the treatment is binary ($T_i = 1$ or $T_i = 0$), the ACE is defined as $E(Y_i(1) - Y_i(0))$. If the treatment takes a finite number of values $0, 1, 2, \ldots, k$, we can define additional ACEs as pairwise comparisons, such as $E(Y_i(2) - Y_i(0))$ or $E(Y_i(2) - Y_i(1))$. This multiple treatment problem has not received much attention either, but it is not substantially more difficult than the binary situation.

The leap to continuous treatments is harder and has been discussed by only a few authors (Moodie and Stephens, 2012; Kluve et al., 2012; Hirano and Imbens, 2004; Imai and van Dyk, 2004). Before describing that work, in the next section we review methods for estimating causal effects when the treatment is binary. The later sections of this chapter review methods for a continuous treatment.

2.2 Estimating an average causal effect when the treatment is binary

2.2.1 The *prima facie* estimator

In the binary treatment setting, the main estimand of interest is the average causal effect (ACE),

$$ACE = \mu(1) - \mu(0) = E[Y_i(1)] - E[Y_i(0)], \tag{2.1}$$

which is the average of the unit-level causal effects $D_i = Y_i(1) - Y_i(0)$. If all potential outcomes could be seen, the ACE could be estimated by

$$\widehat{ACE}_{Gold} = \frac{1}{N}\sum_{i=1}^{N} D_i = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0). \tag{2.2}$$

Of course, this estimate of the ACE cannot be computed from the observed data, because only one potential outcome is observed for each individual. Nevertheless, the estimate in (2.2) can be regarded as a gold standard, because it provides a consistent estimate of the ACE under very general conditions, and to the extent that other estimates mimic the behavior of (2.2), they will also tend to perform well (Schafer and Kang, 2008). An estimator that can always be computed is the average response among units with $T_i = 1$ minus the average response when $T_i = 0$,

$$\widehat{ACE}_{Prima} = \frac{\sum_{i=1}^{N} T_i Y_i(1)}{\sum_{i=1}^{N} T_i} - \frac{\sum_{i=1}^{N}(1 - T_i) Y_i(0)}{\sum_{i=1}^{N}(1 - T_i)}, \tag{2.3}$$

which consistently estimates the *prima facie* effect, $PFE = E(Y_i(1)|T_i = 1)$ - $PFE = E(Y_i(0)|T_i = 0)$ (Holland, 1986). If the treatment is independent of $Y_i(1)$ and $Y_i(0)$, such as in a randomized study, then the *prima facie* effect coincides with the ACE. If T_i is not independent of the potential outcomes, $\widehat{ACE}_{Prima}$ can be badly biased.

2.2.2 Outcome-prediction methods

2.2.2.1 Regression and ANCOVA

The *prima facie* estimator $\widehat{ACE}_{Prima}$ is a naive method that does not take covariates $\boldsymbol{X}_i$ into account. If treatment is not randomly assigned, the treated and untreated subgroups may not be representative of the overall population, and this simple difference estimate may be biased. One simple way to take $\boldsymbol{X}_i$ into account

is commonly known as regression adjustment or analysis of covariance (ANCOVA), a model-based technique that predicts what might have happened if the $T_i = 1$ and $T_i = 0$ groups had no baseline differences in the covariates.

The simple version of ANCOVA supposes that $E[Y_i|T_i, \mathbf{X}_i] = \alpha + T_i\theta + \mathbf{X}_i^T\boldsymbol{\beta}$, but more general version are possible with interactions, nonlinear relationships, and heteroscedastic errors (Little et al., 2000; Schafer and Kang, 2008). The treatment effect is associated with the parameter θ. This parameter coincides with the ACE only under certain circumstances. ANCOVA was originally proposed by R.A. Fisher to help reduce variance of the estimated treatment effects in randomized studies. For nonrandomized studies, the rationale for adjusting for $\mathbf{X}_i$ is to reduce bias attributable to the measured confounders (Schafer and Kang, 2008). ANCOVA is easy to use, but requires that the model be correct in order for the parameter estimate θ to correspond to the ACE. The critical assumptions being made are that the potential-outcome regression surfaces $E(Y_i(1)|\boldsymbol{X}_i)$ and $E(Y_i(0)|\boldsymbol{X}_i)$ are both linear functions of $\boldsymbol{X}_i$ and that all of the slopes are equal. If the slopes are unequal, the model can be corrected by including interactions between $\boldsymbol{X}_i$ and T_i. Sensitivity to model failure grows as the distance between $E(X_i|T_i = 1)$ and $E(X_i|T_i = 0)$ increases (Schafer and Kang, 2008).

2.2.2.2 Regression estimation

Another way to use regression for estimating the ACE is to separately model the response for the treatment and control groups and to use these models to predict the unseen potential outcomes. Suppose the two models are $E[Y_i|T_i = 0, \mathbf{X}_i] = \mathbf{X}_i^T\boldsymbol{\beta}_0$ and $E[Y_i|T_i = 1, \mathbf{X}_i] = \mathbf{X}_i^T\boldsymbol{\beta}_1$. After obtaining estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ from their respective groups, we can predict the missing potential outcomes by taking $\hat{Y}_i(0) = \mathbf{X}_i^T\hat{\boldsymbol{\beta}}_0$ for every unit having $T_i = 1$ and $\hat{Y}_i(1) = \mathbf{X}_i^T\hat{\boldsymbol{\beta}}_1$ for every unit having $T_i = 0$. We can also replace the observed potential outcomes by their regression

predictions with no ill effect, because standard methods for estimating regression coefficients (e.g. ordinary least squares) yield residuals with mean zero over the fitting sample. A regression-prediction estimator for the ACE is then

$$\widehat{ACE}_{Reg} = \frac{1}{N}\sum_{i}^{N}(\hat{Y}_i(1) - \hat{Y}_i(0)). \tag{2.4}$$

When the regression models are fit by ordinary least squares, the regression estimation method for estimating the ACE gives a result that is numerically equal to the ANCOVA parameter estimate under an ANCOVA model that includes the $\boldsymbol{X}_i T_i$ interactions (Schafer and Kang, 2008).

In this binary setting, there will be multiple units with $T_i = 1$ and with $T_i = 0$. In the continuous treatment setting, there is a lack of replication at each particular treatment value t. If a unit has an observed value $T_i = t$ in the continuous treatment setting, then there is not likely to be another unit with that same t value. This makes regression estimation difficult for the continuous treatment and requires pooling units across t values, which is discussed in later chapters.

2.2.3 Introducing the propensity score

2.2.3.1 Definition

In the binary treatment setting, the propensity score (PS) is the probability of receiving a treatment given a set of covariates, $\pi_i(\mathbf{x}) = P(T_i = 1|\mathbf{X}_i = \boldsymbol{x})$. The realized propensity score for sample unit i will be written as $\pi_i = \pi_i(\boldsymbol{X}_i)$. The PS was defined by Rosenbaum and Rubin (1983, 1984) and has been used in many studies (Guo and Fraser, 2014) to make causal inferences in observational studies.

The PS can balance covariates across the treatment groups, and it is the coarsest function of covariates with the balancing property, which is: $P(\mathbf{X}_i = \mathbf{x}_i|\pi_i = c, T_i = 1) = P(\mathbf{X}_i = \mathbf{x}_i|\pi_i = c, T_i = 0)$. The PS distills the covariate information

into a scalar value that can be used to estimate the ACE in a variety of ways. Units with the same propensities have the same probability of treatment. This means that units randomly received treatment or control and have the same distribution of covariates (same distribution of $\mathbf{X}_i$), on average (Rosenbaum and Rubin, 1983, 1984). Adjusting the sample by their PS values means that units with equal propensities appear to be allocated to a treatment group randomly. If units receive treatment levels randomly, then estimating causal effects becomes straightforward. For example, if we could divide the sample units into groups in which the PSs are homogeneous, then the *prima facie* estimator of the ACE within any group will consistently estimate the ACE for the subpopulation represented by the group (i.e. the population of units with the same PS).

2.2.3.2 Estimating the propensity score

In a nonrandomized study, the PS is unknown and must be estimated. The most common way to model the PS is by logistic regression, although other methods such as probit regression, boosted regression (McCaffrey et al., 2004; Zhu et al., 2015), CART, and random forests have been used (Lee et al., 2010).

For the multiple treatment setting, Joffe and Rosenbaum (1999) described a single variable balancing score that uniquely determines the distribution of doses given $\mathbf{X}_i$. McCullagh (1980) describes an ordinal logistic regression model that can be used to describe the distribution of doses given $\mathbf{X}_i$. The $\mathbf{X}_i^T\boldsymbol{\beta}$ component from the ordinal logistic model would be the balancing score or propensity scalar quantity for a multiple treatment setting.

2.2.3.3 Checking the propensity score

To check the fit of the PS model, the criterion traditionally applied is covariate balance. Ideally, the PS balances the distribution of covariates across all treatment

levels. For example, in the binary treatment setting, if we restrict our attention to a set of units with similar propensity scores, we should find no major differences between $T_i = 1$ and $T_i = 0$ on any component of $\boldsymbol{X}_i$ or function of $\boldsymbol{X}_i$. Analysts will typically divide the sample into subgroups (e.g. defined by the quantiles of estimated propensities) and compute t-tests and standardized mean differences for each covariate. Graphical representations of the covariate distributions (histograms, boxplots, etc.) at different treatment levels within the propensity-defined classes is another popular way to check covariate balance.

It is also important to check PS overlap to determine if estimating causal effects is even feasible. Comparing histograms of the PS values of the treatment group and control group is often used to evaluate the degree of overlap. If there is an insufficient number of units with $T_i = 1$ in regions of low propensity, or an insufficient number of units with $T_i = 0$ in regions of high propensity, then causal inference in those regions is ill advised because it may require excessive amounts of extrapolation (e.g., Gelman and Hill, 2006, Chap. 10).

Many authors have claimed that overfitting is not a serious problem when estimating the PS, because the main goal is predicting the treatment probabilities. In fact, overfitting the PS is not detrimental to the estimation of the ACE since prediction and covariate balance are the main criteria for estimating the PS (Rubin, 2004; Brookhart et al., 2006).

2.2.4 Using propensity scores to estimate causal effects

2.2.4.1 Matching

Propensity-score matching, an intuitive solution to causal inference, constructs treatment and control samples that have similar covariate distributions (Rubin and Thomas, 2000; Stuart, 2010; Rubin and Thomas, 1996). With a binary treatment,

matching is usually performed in this manner. For each unit in the smaller group ($T_i = 0$ or $T_i = 1$), an algorithm is used to select a unit in the larger group with a similar propensity score, and perhaps similar values of $\boldsymbol{X}_i$ as well (as measured, for example, by Mahalanobis distance). After matches have been found for all units in the smaller group, the excess units in the larger group are discarded. This method is called 1-to-1 matching. Because the pairs have similar covariates but different treatments, this method can provide insight into the effect of a treatment on the outcome while removing biases due to confounding that can be attributed to $\boldsymbol{X}_i$. Matching is thoroughly reviewed in Stuart (2010) and in Rubin and Thomas (1996), Rubin (2006), Ho et al. (2006), and Pattanayak et al. (2011).

There are different options for how to decide which matches to make. After fitting the PS model, the estimated PS values are used to match units in the treatment group with units in the control group. The idea is that units with similar PS values should have similar distributions on their covariates, if the PS is modeled correctly. After creating the matched dataset, the ACE can be calculated by comparing the mean outcomes of each group. It is also possible to combine model-based regression adjustments with matching (Rubin and Thomas, 2000).

One drawback to matching is that the covariate distributions in the resulting matched samples may be atypical of the population from which the original full sample was drawn. The causal effect estimated from the matched sample may not generalize to the ACE for the whole population. This is not necessarily a drawback; matching forces the analyst to be realistic about the limitations of the data, which may not be amenable to estimating an overall population ACE. When matching cannot estimate the ACE for the whole population, it may provide a realistic estimate for a smaller subpopulation.

Matching does not easily extend to a continuous treatment, because of the lack of replication at each particular $t \in \mathcal{T}$. Lu et al. (2001) applied matching on

doses, but they did so by dichotomizing the dose, estimating a quantity comparing high and low doses rather than the ADRF.

2.2.4.2 Subclassification

Propensity-score subclassification was originally proposed by Rosenbaum and Rubin (1983, 1984) to adjust for the selection bias. In this method, the PS is estimated and then units are divided into groups with similar propensities, usually based on quantiles of the estimated PS values. The number of classes may depend on the size of the data set, but many analysts use five classes with quintiles of the PS distribution as endpoints. Units within the same PS subclasses should be similar in their covariate distributions. Within each subclass, the ACE is estimated in a straightforward manner, such as the *prima facie* method, although regression adjustments could also be used. To get the overall estimate of the ACE, the subgroup estimates are combined by a weighted average depending on subclass size,

$$\widehat{ACE} = \sum_{s} \frac{N_s}{N} \hat{\theta}_s, \tag{2.5}$$

where s is subclass index, N_s is the number of units in subclass s, N is the total sample size, and $\hat{\theta}_s$ is the ACE estimate within subclass s.

2.2.4.3 Weighting

Another method used to remove the bias attributable to covariate imbalance is to apply weights derived from propensity scores to units (Robins et al., 2000). The weighting procedure expands each treatment group ($T_i = 0$ or $T_i = 1$) in the sample to a pseudo-population that mimics the properties of the overall population. The weighting procedure starts by estimating the PS. Each unit in the $T_i = 1$ group is assigned a weight proportional to $1/\pi_i$, and each unit in the $T_i = 0$ group is assigned

a weight proportional to $1/(1-\pi_i)$. This technique bears a strong resemblance to the Horvitz-Thompson weighting method used in sample surveys (Horvitz and Thompson, 1952). Units with lower probabilities of getting selected into their respective treatment group will have more weight allocated to them, whereas units with higher probabilities of selection will be assigned less weight. In other words, the weights correct the distortions that arises from differential probabilities of selection. This method implicitly assumes that the probability of treatment is modeled correctly, and if so, then the weighting process will make the observational data appear to come from a randomized trial (Hernan and Robins, 2016).

Let $w_i(1) = 1/\pi_i = 1/P(T_i = 1|\mathbf{X}_i)$ and $w_i(0) = 1/(1-\pi_i) = 1/P(T_i = 0|\mathbf{X}_i)$, where $w_i(t)$ represents the weight of unit i at treatment level t. The inverse probability of treatment weighting estimator for the ACE is

$$\widehat{ACE} = \frac{\sum \mathbb{1}(T_i = 1)Y_i w_i(1)}{\sum \mathbb{1}(T_i = 1)w_i(1)} - \frac{\sum \mathbb{1}(T_i = 0)Y_i w_i(0)}{\sum \mathbb{1}(T_i = 0)w_i(0)}, \tag{2.6}$$

where summations are taken over the entire sample.

2.2.5 Dual-modeling techniques

2.2.5.1 Dual-modeling background

In the previous sections, we reviewed methods for estimating an ACE based on models for the outcome (regression adjustment and ANCOVA) and methods based on models for the treatment assignment mechanism (propensity-based matching, subclassification, and weighting). In recent years, it has become increasingly popular to combine the two types of models into a single estimator that protects against bias due to model misspecification. These dual-modeling approaches have a property known as double robustness, which means that they remain consistent if either of the two models has been correctly specified (Van der Laan and Robins, 2003).

Tsiatis (2007) provides theory for a wide range of semiparametric estimators, some of which are doubly-robust, that are applicable to incomplete-data and causal inference problems, based on the idea of influence functions.

2.2.5.2 Weighted residual bias correction

Weighted residual bias correction is a dual-modeling strategy that has two components: an outcome regression component and a bias correction component. The outcome regression component relates covariates and treatment to the outcome, and the bias correction component relates covariates to treatment selection. The resulting estimates are asymptotically unbiased if either of these two models are correctly specified. The general form of the weighted residual bias correction estimator is

$$\widehat{ACE} = \frac{1}{N}\sum_i (\hat{Y}_i(1) - \hat{Y}_i(0)) + \frac{\sum_i T_i \hat{\pi}_i^{-1} \hat{\epsilon}_i(1)}{\sum_i T_i \hat{\pi}_i^{-1}} - \frac{\sum_i (1-T_i)(1-\hat{\pi}_i)^{-1} \hat{\epsilon}_i(0)}{\sum_i (1-T_i)(1-\hat{\pi}_i)^{-1}}, \quad (2.7)$$

where $\hat{Y}_i(t)$ is a regression prediction for $Y_i(t)$, $t = \{0, 1\}$ based on an outcome regression model, and $\hat{\epsilon}_i(t) = \hat{Y}_i(t) - Y_i(t)$ is the residual for unit i under treatment t. The first term in (2.7) is the standard regression estimate, the second term is the bias correction for $E(Y_i(1))$ and the third term is the bias correction for $-E(Y_i(0))$. If either the outcome model or the treatment allocation model are correctly specified, then the estimates will be unbiased.

Weighted residual bias correction has strong connections to the generalized regression (GREG) estimator in survey statistics literature (Deville and Särndal, 1992). In the survey context, weights come from the sample design. Tan (2010) shows that doubly robust estimators have additional desirable properties, other than consistency, if either the propensity score or outcome regression models are correct.

2.2.5.3 Weighted regression estimation

Weighted regression estimation is a doubly robust estimator that uses inverse probability weights in a model for outcome prediction (Schafer and Kang, 2008). This method is similar to regression estimation, but includes weights in the estimation step to give consistent estimates of the regression coefficients that one would get by fitting the outcome regression models to the full population. The weighted estimate of the coefficients for the model predicting $Y_i(1)$ is

$$\hat{\boldsymbol{\beta}}_{1\_wt} = \left(\sum_i T_i \hat{\pi}_i^{-1} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \left(\sum_i T_i \hat{\pi}_i^{-1} \boldsymbol{X}_i Y_i \right), \tag{2.8}$$

and the weighted estimate of the coefficients for the model predicting $Y_i(0)$ is

$$\hat{\boldsymbol{\beta}}_{0\_wt} = \left(\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \left(\sum_i (1 - T_i)(1 - \hat{\pi}_i)^{-1} \boldsymbol{X}_i Y_i \right) \tag{2.9}$$

Once the parameters are estimated for each of these potential outcome models, the regression predictions for $Y_i(1)$ and $Y_i(0)$ are computed for each unit. The average difference in these predictions estimates the ACE,

$$\widehat{ACE} = \frac{1}{N} \sum_i^N \left(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_{1\_wt} - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_{0\_wt} \right). \tag{2.10}$$

2.2.5.4 Regression estimation with propensity-related covariates

The regression models used for the estimator in (2.4) may have incorrectly specified the relationship between the potential outcomes and the covariates, causing the estimate of the ACE to be biased. Another way to remedy this problem is by including propensity-related covariates (Little and An, 2004; Schafer and Kang, 2008). The outcome Y_i is regressed on T_i, $\boldsymbol{X}_i$, and covariates related to the estimated

PS. Little and An (2004) allow the mean response to vary with propensities in a flexible way using cubic splines. Kang and Schafer (2007) recommend adding dummy variables that identify propensity classes where units have homogeneous propensities, which is equivalent to fitting a piecewise constant function (i.e. a zero-order spline).

2.3 Estimating causal quantities when the treatment is continuous

2.3.1 Difficulties with the continuous treatment

As we have mentioned, when the setting changes from a binary treatment to a continuous one, the number of potential outcomes for each unit becomes uncountably infinite, and the data available at any particular treatment value t become sparse. The strategy of matching units with similar covariates, but with different treatment levels is more difficult in the continuous-treatment setting, because the treatment values (doses) reflected in the matches would need to span the domain $\mathcal{T}$. Hypothetically, if the amount of data collected were very large, then for each fixed level of covariates, it would be possible to find units with different treatment levels and be able to trace out the unit-level dose-response function. With realistic sample sizes, however, this is usually impossible. The sparsity of data at or near any particular dose t also makes it difficult to extend methods based on regression prediction, subclassification, or weighting, as we have previously mentioned. Moreover, there is no single way to extend the propensity score to a continuous treatment; at least two generalizations have been proposed (Hirano and Imbens, 2004; Imai and van Dyk, 2004).

Thus far, most of the methods proposed for continuous treatments have not assumed that the ADRF follows any particular parametric form. Parametric assumptions can provide more structure and make the problem more tractable. If no

parametric form is assumed, the space of possible answers becomes much larger and the solution is more difficult to pinpoint.

2.3.2 Methods based on outcome prediction models

Hill (2011) proposed using Bayesian additive regression trees (BART) for causal inference. BART is focused on predicting potential outcomes and does not require fitting a model for the treatment mechanism. BART accommodates large numbers of covariates, and can handle binary, categorical, and continuous treatments. However, current implementations of BART require the outcome variable to be continuous.

BART can be regarded as a Bayesian adaptation of a random forest (Chipman et al., 2010). It averages the predictions over a space of regression trees, and the influence of each tree is modified by a regularization prior so that each tree only contributes a small amount to the overall fit. The priors are also designed to avoid overfitting the data (Chipman et al., 2010).

In effect, BART fits a response surface to predict $E(Y_i(t)|\boldsymbol{X}_i = \boldsymbol{x})$ over the space of t and $\boldsymbol{x}$. The trees used to construct these predictions are simulated from a posterior distribution of trees using an elaborate Markov chain Monte Carlo (MCMC) procedure.

Although most applications of BART have been for the binary treatment setting, Hill (2011) also mentions an extension to a continuous treatment. Instead of estimating an ADRF, Hill (2011) compares the outcomes of units with treatment level $T_i = t$ to their outcomes had they received $T_i = 0$ (or some other meaningful baseline). The actual comparison is between $Y_i(0)|(T_i = t)$ and $Y_i(t)|(T_i = t)$. In other words, the causal comparison is the outcome for units that received a given treatment dose with what their outcome would have been had they received no dose.

Hill (2011) reports that this method performs well in simulation studies. A

major drawback of BART is the large amount of computing resources needed, especially as the number of covariates grows.

2.3.3 Methods based on treatment-focused models

2.3.3.1 Generalizing the propensity score to the continuous-treatment setting

For the continuous-treatment setting, Hirano and Imbens (2004) defined the generalized propensity score (GPS) as the treatment assignment density evaluated at a particular treatment value and set of covariates. This requires a model for the conditional distribution of T_i given $\boldsymbol{X}_i$. When evaluated at the realized $\boldsymbol{X}_i$ and any specific t, the GPS becomes a random variable that can be used to balance covariates. The GPS is defined as $r(t, \boldsymbol{x}) = f(T_i = t|\boldsymbol{X}_i = \boldsymbol{x})$. In applications, the GPS is applied in two stages. The first stage involves fitting a model to estimate $f(T_i|\boldsymbol{X}_i)$. The second stage uses that model to obtain an estimate of $E(Y_i(t))$ as we shall describe later.

Another generalization of the propensity score, called the propensity function (PF), was proposed by Imai and van Dyk (2004). They suppose that the conditional density of T_i given $\boldsymbol{X}_i$ is indexed by a parameter $\boldsymbol{\psi}$. If this density depends on $\boldsymbol{X}_i$ only through a finite-dimensional quantity $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$, then $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$ is a PF. Thus, Imai and van Dyk (2004) make the extra assumption that the density of T_i given $\boldsymbol{X}_i$ is uniquely parametrized by $\pi(\boldsymbol{X}_i, \boldsymbol{\psi})$, which uniquely identifies the PF, $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$. In other words, identifying $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$ will specify the propensity function. In many cases, the parameter $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$ will be a scalar value which distills all the information in $\mathbf{X}_i$ into a single number. For example, if $f(T_i|\boldsymbol{X}_i)$ is modeled as a normal linear regression with a constant variance, $\mathcal{N}(\boldsymbol{X}_i^T\boldsymbol{\beta}, \sigma^2)$, then $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2)$, and the linear predictor becomes a scalar PF, $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi}) = \boldsymbol{X}_i^T\boldsymbol{\beta}$. This dimension reduction greatly

simplifies the problem of estimating causal effects.

Imai and van Dyk (2004) show that, under strong ignorability, the PF has two key properties. First, it is a balancing score, meaning that conditionally given the PF, the treatment is independent of the covariates. Second, given the PF, the treatment assignment is strongly ignorable, $T_i \perp\!\!\!\perp \mathcal{Y}_i|\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$. In other words, the low-dimensional PF is capable of removing all the bias in a causal comparison attributable to the higher dimensional $\boldsymbol{X}_i$. In actual applications, the PF is not known and must be estimated by fitting a model $P(T_i = t|\boldsymbol{X}_i; \boldsymbol{\psi})$ to sample data $(\boldsymbol{X}_i, T_i), i = 1, \ldots, N$ and using a plug-in estimate of the parameter $\hat{\boldsymbol{\psi}}$. We will denote the estimated PF by $\boldsymbol{\pi}(\boldsymbol{X}_i, \hat{\boldsymbol{\psi}}) = \hat{\boldsymbol{\pi}}_i$, or $\hat{\pi}_i$ for a scalar.

When fitting the treatment model, there are many possible methods. To evaluate the treatment model fit, covariate balance is a key criterion. There is no clear consensus for checking covariate balance in the continuous treatment setting. It is a topic of current research, but some different ways are described in Hirano and Imbens (2004); Imai and van Dyk (2004); Flores et al. (2012).

2.3.3.2 Inverse probability of treatment weighting

For the binary treatment setting in section (2.2.4.3), we described a weighting procedure that starts by estimating the propensity $\pi(\mathbf{X}_i) = P(T_i = 1|\boldsymbol{X}_i)$ of getting the treatment, and then weighting the units that received the treatment by $1/\pi_i$ and weighting the units that did not receive the treatment by $1/(1-\pi_i)$. This weighting is similar to the Horvitz-Thompson weighting (Horvitz and Thompson, 1952) in survey methodology. Units with lower probabilities of selection in the sample will receive more weight because they are rare, whereas units that have high probability of selection will be well represented in the sample and so will be assigned less weight. This method assumes that the probability of treatment is modeled correctly.

Robins et al. (2000) take this idea of weighting and apply it to the continuous-treatment setting under an assumed parametric form for the ADRF. For example, if we suppose that the function is linear, $\mu(t) = \theta_0 + \theta_1 t$, the method estimates θ_0 and θ_1 by a least-squares calculation that minimizes

$$\sum_{i}^{N} w_i(T_i)(Y_i - \theta_0 - \theta_1 T_i)^2, \tag{2.11}$$

where $w_i(T_i)$ is a weight. This will lead to $\hat{\mu}(t) = \hat{\theta}_0 + \hat{\theta}_1 t$. Alternatively, if we suppose that $\mu(t) = \theta_0 + \theta_1 t + \theta_2 t^2$, we minimize

$$\sum_{i}^{N} w_i(T_i)(Y_i - \theta_0 - \theta_1 T_i - \theta_2 T_i^2)^2, \tag{2.12}$$

which leads to $\hat{\mu}(t) = \hat{\theta}_0 + \hat{\theta}_1 t + \hat{\theta}_2 t^2$. Models of this type are called marginal structural models. If the form of $\mu(t)$ has been correctly specified, the parameters can be consistently estimated by using weights $w_i(T_i) = 1/P(T_i|\boldsymbol{X}_i)$ from a correctly specified treatment model.

One problem with using those weights is the possibility of highly variable values. Weighting by the reciprocals of treatment densities tends to be highly unstable, and the method is especially sensitive to misspecification of $P(T_i|\boldsymbol{X}_i)$ in the extreme tails. For this reason, Robins et al. (2000) recommend using stabilized weights of the form $w_i(t) = g(T_i)/P(T_i = t|\boldsymbol{X}_i)$, where $g(T_i)$ is the marginal density for T_i. As we will show in the next chapter, this weighting scheme bears a strong similarity to the well known simulation method called importance sampling. The stabilized weights are intended to adjust the fit to what it might have been had the treatment been assigned independently of the covariates, that is, if it had been distributed according to $g(T_i)$ rather than $P(T_i|\boldsymbol{X}_i)$.

2.3.3.3 Method of Imai and van Dyk

After introducing the PF, Imai and van Dyk (2004) describe its use in causal inference. They recommend subclassifying units based on the PF and estimating causal effects within these subclasses. If the PF model is correctly specified, then within each subclass the covariates are distributed evenly across target groups. Within each subclass, the outcome can be modeled as a function of treatment such as $Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \epsilon$. The final step is to combine the estimated function across subclasses to get an overall estimate, using a weighting scheme based on the number of observations in each subclass.

A modified version of this method, suggested by Zhao et al. (2014) fits a smooth coefficient model of the form $E[Y_i|T_i, \hat{\theta}] = j(\hat{\theta}) + k(\hat{\theta})T_i$ where j and k are assumed to be smooth but are not otherwise specified, and $\hat{\theta}_i$ is an estimated PF. This smooth-coefficient model is less rigid than subclassification. It allows the estimates to change gradually as θ changes. In order to estimate the ADRF, the predictions induced by the smooth-coefficient model are averaged over the empirical distribution of $\boldsymbol{X}_i$. For example, the set of steps for estimation in this method are:

1. Fit a model to describe the treatment given the covariates, and extract the propensity function $\hat{\theta}_\psi(\mathbf{X}_i)$.

2. Fit an observable model $Y_i|(\hat{\theta}_\psi(\mathbf{X}_i), T_i) = f(\hat{\theta}_\psi(\mathbf{X}_i), T_i)$, and estimate the model $\hat{f}(\cdot)$

3. Calculate the estimated ADRF as $\hat{E}[Y_i(t)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\theta}_\psi(\mathbf{X}_i), t)$ at each particular t value of interest.

This method can be regarded as a hybrid of methods from Hirano and Imbens (2004) (which we describe next) and Imai and van Dyk (2004). The outcome model used in Step 2 can be a flexible model (e.g., a generalized additive model). It is also

possible to include additional components or functions of $\boldsymbol{X}_i$ in the outcome model when modeling the relationship between Y_i and T_i, which may increase efficiency.

2.3.3.4 Method of Hirano and Imbens, including modifications

Hirano and Imbens (2004) (HI) introduced an imputation-type method that includes a GPS component. The idea is to fit a parametric model to the observed outcomes, including the estimated GPS (a function of t) as a predictor, and use the model to predict missing potential outcomes at specific values of t.

The method requires several steps. First, a model is used to relate treatment to the recorded covariates. For example, $T_i|\mathbf{X}_i \sim \mathcal{N}(\mathbf{X}_i^T\boldsymbol{\beta}, \sigma^2)$. We estimate the parameters of this model. Next, the GPS for each unit is computed under the model, for example,

$$\hat{R}_i(t) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{(t-\mathbf{X}_i^T\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}}, \tag{2.13}$$

using the estimated parameters. These GPS estimates become predictors in the outcome model. The outcome is modeled as a function of T_i and $\hat{R}_i$ parametrically. For example,

$$E[Y_i|T_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 \hat{R}_i + \alpha_4 \hat{R}_i^2 + \alpha_5 \hat{R}_i T_i. \tag{2.14}$$

After collecting the estimated parameters in the outcome and treatment models, we plug in the treatment values into the model to predict the unknown potential outcomes of each unit at a given treatment level. For example, if we plug $T_i = t$ into the estimated models, each unit will have a potential outcome estimated at treatment level $T_i = t$,

$$\hat{Y}_i(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{R}_i(t) + \hat{\alpha}_4 \hat{R}_i^2(t) + \hat{\alpha}_5 \hat{R}_i(t) t. \tag{2.15}$$

The final step is to aggregate these estimated potential outcomes to get an average treatment effect at dose level $T_i = t$. The mean outcome at dose level $T_i = t$ is given by

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i}^{N} \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{R}_i(t) + \hat{\alpha}_4 \hat{R^2}_i(t) + \hat{\alpha}_5 \hat{R}_i(t) t. \tag{2.16}$$

Different treatment levels t are plugged into (2.15) and (2.16) to trace out an estimated ADRF. These models can be made more flexible by including more higher order terms and interactions, or by fitting generalized additive models. The treatment models may also be made more flexible.

Moodie and Stephens (2012) extended this method to the longitudinal setting to estimate the direct effect of a continuous treatment on a longitudinal response. They formulate a GPS approach that is suitable for the analysis of repeated measures response data with interval dependent treatments. Their example, which has two time intervals, estimates the potential outcomes at the first time interval and includes them as predictors for the second time interval.

2.3.3.5 Method of Flores et al.

Flores et al. (2012) observe that a Horvitz-Thompson style weighting scheme can be applied to a continuous treatment by weighting units inversely proportional to the GPS at any given dose level, and then smoothing over the dose levels with a kernel-type method used in scatterplot smoothing.

Imagine a setting with three treatment values. The estimate for $\mu(1)$ is

$$\frac{\sum \mathbb{1}(T_i = 1) Y_i w_i(1)}{\sum \mathbb{1}(T_i = 1) w_i(1)}, \tag{2.17}$$

when $\mu(2)$, use

$$\frac{\sum \mathbb{1}(T_i = 2) Y_i w_i(2)}{\sum \mathbb{1}(T_i = 2) w_i(2)}, \tag{2.18}$$

and when $\mu(3)$, use

$$\frac{\sum \mathbb{1}(T_i = 3) Y_i w_i(3)}{\sum \mathbb{1}(T_i = 3) w_i(3)} \tag{2.19}$$

with $w_i(t) = 1/P(T_i = t|\mathbf{X}_i)$. If the treatment is truly continuous, then the probability of getting a particular treatment level is zero. The expressions with indicator functions above will be zero almost surely. To get around this problem, a possible option is to group units into bins based on treatment but then make the bins more and more narrow to resemble the continuous treatment setting. Conceptually, when there is a finite number of treatments such as $t \in \{1, 2, \ldots, N\}$, we can imagine a histogram with N bars with the area of each bar representing the probability of getting a particular treatment level. We have $\sum_j Pr(T_i = j) = 1$ but for the continuous treatment setting, $\int_{\mathcal{T}} f(t)dt = 1$. An example of how to approximate the histogram with the smooth curve is $P(T_i \in \Delta|X) \approx 2hR_i^t$, where h is a sequence of numbers tending to 0 as $N \to \infty$ and $\Delta = [t - h, t + h]$ describes the region of interest around t. An estimator of the DRF is

$$\mu_N(t) = \frac{1}{N} \frac{\sum_{i=1}^N \mathbb{1}(T_i \in \Delta) Y_i}{2h\hat{R}_i^t}. \tag{2.20}$$

As the number of subclasses increases, $\Delta \to 0$, and an estimator of $\mu(t)$ that can smooth out the relationship between t and Y_i is

$$\mu(t) = \frac{\sum_{i=1}^N K_h(T_i - t) Y_i w_i(t)}{\sum_{i=1}^N K_h(T_i - t) w_i(t)}, \tag{2.21}$$

with $K_h(T_i - t)$ being a kernel function such as a Gaussian kernel, triangular kernel density, or other shape. Instead of only looking at each set of units that lie within Δ for each t, a kernel function can be used to give more influence to units closer to

t than farther away (the kernel function would give more weight to units closer to the particular treatment level of interest).

In a slight variation, let $\tilde{K}_h(T_i - t) = K_h(T_i - t)/\hat{R}_i^T = K_h(T_i - t)\hat{w}_i(T_i = t)$, then an estimate of $\mu(t)$ is

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{N} \tilde{K}_h(T_i - t)Y_i}{\sum_{i=1}^{N} \tilde{K}_h(T_i - t)}. \tag{2.22}$$

This method is an adaptation of the Nadaraya-Watson estimator (Nadaraya, 1964) which is a local constant regression but weighted by the inverse of the GPS. Another method described by Flores et al. (2012) is to use a local linear regression that takes the form

$$\hat{\mu}(t)_{IW} = \frac{D_0(t)S_2(t) - D_1(t)S_1(t)}{S_0(t)S_2(t) - S_1^2(t)}, \tag{2.23}$$

where $S_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,X}(T_i - t)(T_i - t)^j$ and $D_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,X}(T_i - t)(T_i - t)^j Y_i$.

Flores et al. (2012) also suggests semiparametric or nonparametric outcome models based on splines. These are also described in the Stata program in Bia et al. (2014). These methods are an extension of the version in Hirano and Imbens (2004) and allow for spline based terms of the GPS and t.

2.4 Discussion

The different estimators summarized in this chapter show the variety of options currently available. In the next chapter, we present new methods for estimating an ADRF that assumes the ADRF follows a parametric form.

Chapter 3: Causal inference with a continuous treatment and outcome: alternative estimators for parametric dose-response functions

3.1 Introduction

3.1.1 Parameterizing the dose response function

Previous strategies for estimating an ADRF make minimal assumptions about the shape of $Y_i(t)$. In contrast, we will suppose that $Y_i(t) = \sum_{j=1}^{k} \theta_{ij}\, b_j(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t)$, where $\boldsymbol{b}(t) = (\, b_1(t), \ldots, b_k(t)\,)^\top$ is a vector of known basis functions, and $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ik})^\top$ is a vector of real-valued coefficients specific to unit i. An important special case is $\boldsymbol{b}(t) = (1, t)^\top$, which specifies linear paths whose intercepts and slopes may vary. Higher-order polynomial and spline (piecewise polynomial) models can also be expressed in this form. The observed outcome is $Y_i = Y_i(T_i) = \boldsymbol{\theta}_i^\top \boldsymbol{B}_i$, where $\boldsymbol{B}_i = \boldsymbol{b}(T_i) = (B_{i1}, \ldots, B_{ik})^\top$. Letting $\boldsymbol{\xi} = \mathrm{E}(\boldsymbol{\theta}_i)$, inference for $\mu(t) = \boldsymbol{\xi}^\top \boldsymbol{b}(t)$ becomes a matter of estimating $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$. When $\boldsymbol{b}(t) = (1, t)^\top$, attention is focused on ξ_2, the increase in average response produced by increasing the dose from t to $t + 1$ for any t.

For a useful parallel to binary treatments, imagine $\boldsymbol{B}_i = (T_i, 1 - T_i)^\top$, $\boldsymbol{\theta}_i = (Y_i(1), Y_i(0))^\top$, and $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. In that setting, T_i plays the role of a missing-data indicator, exposing one of the potential outcomes and hiding the other. With a continuous treatment, T_i filters $Y_i(t)$ to reveal a randomly selected

linear combination of the components of $\boldsymbol{\theta}_i$. Inference about $\boldsymbol{\xi}$ is no longer a problem of missing data *per se*, but an example of the more general concept of coarsened data (Heitjan and Rubin, 1991).

Some may question the value of modeling $Y_i(t)$ when we observe this function only at a single t. This is similar to positing a joint distribution for $Y_i(1)$ and $Y_i(0)$ in the binary case. By restricting $Y_i(t)$ to lie within $\mathcal{M} = \{\boldsymbol{\theta}^\top \boldsymbol{b}(t) : \boldsymbol{\theta} \in \mathbb{R}^k\}$, we reduce the infinite-dimensional $Y_i(t)$ to a k-dimensional vector, simplifying the task of modeling it. Although $Y_i(t) \in \mathcal{M}$ cannot be refuted, we may test $\mu(t) \in \mathcal{M}$ against a more general alternative by adding basis functions and testing the new components of $\boldsymbol{\xi}$ to see if they are nonzero. For example, if $\boldsymbol{b}(t) = (1, t)^\top$ does not fit, we might switch to $\boldsymbol{b}(t) = (1, t, t^2)^\top$, or we might apply a monotonic transformation to the dose and/or response to straighten the relationship. Our assumption $Y_i(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t)$ implies $\mathrm{Cov}(Y_i(t_i), Y_i(t_2)) = \boldsymbol{b}(t_1)^\top V(\boldsymbol{\theta}_i)\boldsymbol{b}(t_2)$, but the estimators we propose will not require this covariance function to be correct. We could have expanded the model to $Y_i(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t) + \epsilon_i(t)$, where $\epsilon_i(t)$ is a mean zero process unrelated to $\boldsymbol{\theta}_i$, T_i, or $\boldsymbol{X}_i$, and our key results would still hold; this expansion would only add covariance parameters about which the data provide little or no information. We assume that all of the unit-level response curves have the same functional form. Relaxing this assumption is a topic of future research.

Although we have described the dose as continuous, the assumption $Y_i(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t)$ also applies when the treatment is multivalued and discrete. If the possible values of T_i are $t_1, \ldots, t_k$, then we may use basis functions $b_j(t) = I(t = t_j)$ for $j = 1, \ldots, k$, and the parameter of interest becomes $\boldsymbol{\xi} = (\mu(t_1), \ldots, \mu(t_k))^\top$.

3.2 Simulated example

The simulation described in Chapter 1 will be used for the remainder of this chapter to evaluate different estimation strategies. The data set contains eight true

$\boldsymbol{A}_i$ covariates, eight transformed versions of $\boldsymbol{A}_i$ denoted $\boldsymbol{A}_i^*$, T_i, Y_i, and the true individual parameters, θ_{i1} and θ_{i2}. The main goal is to estimate the ADRF, which is θ_{i2}, in this parametric example. For details of the simulation, see Section 4.1 in Chapter 1.

3.2.1 The propensity function

Let $P(T_i = t \mid \boldsymbol{X}_i; \boldsymbol{\psi})$ denote the probability density for the treatment given the covariates, indexed by a parameter $\boldsymbol{\psi}$. If this density depends on $\boldsymbol{X}_i$ only through a finite-dimensional quantity $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$, then $\boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$ is called a propensity function (PF) (Imai and van Dyk, 2004). The PF usually has fewer dimensions than $\boldsymbol{X}_i$ or $\boldsymbol{\psi}$, and in many cases it will be a scalar. (When the PF is a scalar, we will omit the boldface and write it as $\pi(\boldsymbol{X}_i, \boldsymbol{\psi})$.) For example, if the treatment is modeled by normal linear regression with homoscedastic errors, $T_i \mid \boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \boldsymbol{\beta}, \sigma^2)$, then $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2)$, and the linear predictor $\pi(\boldsymbol{X}_i, \boldsymbol{\psi}) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ or any one-to-one function of it is a PF.

Under strong ignorability, the PF has two key properties. First, it is a balancing score; conditionally given the PF, the treatment is independent of the covariates. Second, given the PF, the treatment assignment is strongly ignorable, $T_i \perp\!\!\!\perp \mathcal{Y}_i \mid \boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$. In typical applications, the PF is unknown and must be estimated by fitting a model $P(T_i = t \mid \boldsymbol{X}_i; \boldsymbol{\psi})$ to sample data $(\boldsymbol{X}_i, T_i)$, $i = 1, \ldots, N$ and plugging in the estimated parameter $\hat{\boldsymbol{\psi}}$. We denote the estimated PF by $\boldsymbol{\pi}(\boldsymbol{X}_i, \hat{\boldsymbol{\psi}}) = \hat{\boldsymbol{\pi}}_i$, or $\hat{\pi}_i$ if it is a scalar.

Imai and van Dyk (2004) estimate causal effects by subclassification of the PF. After fitting a treatment model, they divide the sample into groups where the $\hat{\boldsymbol{\pi}}_i$'s are roughly constant; they estimate the effects independently within each group and average the effects across groups. They also describe a flexible smooth coefficient model that parameterizes the causal effects and allows them to vary

Table 3.1: Classification of $N = 200$ sample observations by quantiles of the estimated propensity function $\hat{\pi}_i$ and realized dose T_i.

	Percentiles of T_i				
Percentiles of $\hat{\pi}_i$	0–20	20–40	40–60	60–80	80–100
0–20	25	11	4	0	0
20–40	11	15	11	3	0
40–60	4	6	16	10	4
60–80	0	7	8	13	12
80–100	0	1	1	14	24

smoothly with the $\hat{\boldsymbol{\pi}}_i$'s. That method is similar in spirit to a technique that we describe in Section 3.3.8.

The $\hat{\boldsymbol{\pi}}_i$'s are also useful for exploratory work. They help us to diagnose situations where causal inference requires extrapolation to regions of the covariate space where data are sparse. The balancing-score property implies that, in any group where the PF is constant, the treatment is unrelated to any covariate. Testing for associations between T_i and functions of $\boldsymbol{X}_i$ within subclasses of $\hat{\boldsymbol{\pi}}_i$ is useful for checking the adequacy of a treatment model.

In our simulated example, $\mathrm{E}(T_i \mid \boldsymbol{A}_i^*) = 12 + A_{i1}^* + A_{i2}^*/2 + A_{i3}^*/3 + A_{i4}^*/4$ is a true PF. Because the A_{ij}^*'s are hidden from view, we imagine that an analyst would estimate the PF by regressing T_i on $A_{i1} \ldots, A_{i8}$ and use the fitted values as $\hat{\pi}_i$'s. We did this for the sample of $N = 200$ shown in Figure 5 of Chapter 1, and then we assigned the units to categories defined by the sample quintiles of the $\hat{\pi}_i$'s and the T_i's. Frequencies for this cross-classification are shown in Table 3.1. The proportion of variance in T_i's explained by this model is moderately high ($R^2 = 0.66$), leading to a high concentration in cells near the main diagonal. If we were to estimate the ADRF within subclasses of the PF, this table reveals that no data are available for directly estimating $\mathrm{E}(Y_i(t) \mid \hat{\pi}_i)$ for low values of $\hat{\pi}_i$ and high values of t, and vice-versa; casual inference in those regions requires extrapolation. Within each $\hat{\pi}_i$-

class, however, there is adequate information to estimate a linear trend over regions of dose where the T_i's are seen. If those trends look similar across $\hat{\pi}_i$ classes, then it seems reasonable to pool across classes to estimate a common dose effect, and extrapolation becomes less worrisome.

To investigate this, we regressed Y_i on T_i, dummy indicators for the $\hat{\pi}_i$ classes, and the products of T_i with the dummy indicators. A test for equality of slopes gave an F-statistic of 1.45 with $(4, 190)$ degrees of freedom ($p = 0.22$). Evidence against equality is weak, and we will proceed to fit one ADRF for the population. If the variation across classes of $\hat{\pi}_i$ were significant, an analysis that estimated a single ADRF would not be invalidated, because our model does not suppose that treatment effects are homogeneous. However, there may be situations where variation in treatment effects across subclasses is so large that an aggregate ADRF could be misleading. The methods we will describe for estimating an ADRF may be extended to allow the dose-response relationship to vary in relation to the PF or other moderator variables, but those extensions are beyond the scope of this chapter.

3.3 Estimation

3.3.1 The *prima facie* estimator

A naive approach to estimating $\boldsymbol{\xi}$ is to ignore the covariates and regress Y_i on $\boldsymbol{B}_i = \boldsymbol{b}(T_i)$. Using ordinary least squares (OLS), this estimator is

$$\hat{\boldsymbol{\xi}} = \left(\sum_{i=1}^{N} \boldsymbol{B}_i \boldsymbol{B}_i^{\top} \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{B}_i Y_i \right). \tag{3.1}$$

Adapting terminology from Holland (1986), we call (3.1) the *prima facie* estimator; it would be consistent if T_i and $\boldsymbol{\theta}_i$ were independent.

The performance of (3.1) over 1,000 samples from our artificial population is

Table 3.2: Performance of the *prima facie* estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	7.05	0.350	1,190	7.08	7.06
$N = 1,000$	7.06	0.063	2,820	7.07	7.05

summarized in Table 3.2. "Bias" is the average difference between the estimated slope parameter $\hat{\xi}_2$ and the true value $\xi_2 = 0$; "Var." is the variance of $\hat{\xi}_2$; "% Bias" is the bias expressed as a percentage of $\hat{\xi}_2$'s standard deviation; "RMSE" is the square root of the average value of the squared error $(\hat{\xi}_2 - \xi_2)^2$; and MAE is the median value of the absolute error $|\hat{\xi}_2 - \xi_2|$. Percent bias gives some indication of how badly the bias affects inferences about ξ; a useful rule-of-thumb is that confidence intervals and hypothesis tests are seriously impaired when % Bias exceeds 50. MAE is a robust measure of precision unaffected by gross errors that occur in estimation procedures that occasionally go haywire. The *prima facie* estimator is badly biased whether $N = 200$ or $N = 1,000$ and its use is not recommended. We present these results mainly as a benchmark to assess the improvement of alternative methods.

Despite its poor performance, the *prima facie* method has one important virtue: it does not obscure the dose-response relationship by fixing covariates. Many data analysts adjust for confounders by including them on the right-hand side of a regression formula for predicting Y_i. That method creates a conceptual problem. The population ADRF describes the marginal distribution of $Y_i(t)$, which requires averaging over covariates, not conditioning on them. In special cases, averaging and conditioning lead to the same answer, but conceptually they are very different, and this difference often goes unappreciated. To avoid this confusion, we will not modify the *prima facie* estimator by merely adding covariates to the regression formula.

3.3.2 Estimating functions

The *prima facie* estimator (3.1) is the solution to $\boldsymbol{U}(\boldsymbol{\xi}) = \sum_{i=1}^{N} \boldsymbol{U}_i = \boldsymbol{0}$, where $\boldsymbol{U}_i = \boldsymbol{U}_i(\boldsymbol{\xi}) = \boldsymbol{B}_i(Y_i - \boldsymbol{B}_i^\top \boldsymbol{\xi}) = \boldsymbol{B}_i\boldsymbol{B}_i^\top(\boldsymbol{\theta}_i - \boldsymbol{\xi})$ is a vector of estimating functions. Estimators of this type, which are called generalized method-of-moments or Z-estimators, are $\sqrt{N}$-consistent and asymptotically normal if the components of $\boldsymbol{U}_i$ have expected values of zero when evaluated at the true $\boldsymbol{\xi}$ (Newey and McFadden, 1994; Van der Vaart, 2000) . In some cases, the mean-zero property requires knowing another finite-dimensional parameter $\boldsymbol{\phi}$, and the estimating function becomes $\boldsymbol{U}_i = \boldsymbol{U}_i(\boldsymbol{\xi}, \hat{\boldsymbol{\phi}})$, where $\hat{\boldsymbol{\phi}}$ is a $\sqrt{N}$-consistent estimate of $\boldsymbol{\phi}$. For our purposes, we will assume that the regularity conditions for consistency (e.g., Jesus and Chandler (2011)) are satisfied, and we will not compute variance estimates. Variances are not difficult to derive, but in this chapter, performance of point estimates is our main concern.

Because $\boldsymbol{B}_i$ is a function of T_i, it is easy to see that the *prima facie* estimating function has mean zero if T_i and $\boldsymbol{\theta}_i$ are independent. Under the weaker condition $T_i \perp\!\!\!\perp \boldsymbol{\theta}_i \mid \boldsymbol{X}_i$, there are multiple ways to modify the estimating function to give it a zero mean.

3.3.3 Importance weighting

For a related class of problems called marginal structural models, Robins et al. (2000) applied a modified estimating function of the form

$$\boldsymbol{U}_i = \frac{P(T_i)}{P(T_i \mid \boldsymbol{X}_i)} \boldsymbol{B}_i(Y_i - \boldsymbol{B}_i^\top \boldsymbol{\xi}), \tag{3.2}$$

where $P(T_i \mid \boldsymbol{X}_i)$ is the conditional density for T_i given $\boldsymbol{X}_i$, and $P(T_i)$ is the marginal density for T_i. The solution to $\sum_{i=1}^{N} \boldsymbol{U}_i = \boldsymbol{0}$ is the vector of coefficients from the weighted least-squares (WLS) regression of Y_i on $\boldsymbol{B}_i$, with weights

$P(T_i)/P(T_i \mid \boldsymbol{X}_i)$. RHB call these the stabilized weights, claiming that they improve upon the unstabilized versions $1/P(T_i \mid \boldsymbol{X}_i)$. Zhu et al. (2015) used (3.2) to estimate an ADRF by combining a spline basis in $\boldsymbol{B}_i$ with a generalized boosting algorithm for estimating $P(T_i \mid \boldsymbol{X}_i)$.

This technique is related to importance sampling (Hammersley, 2013). The goal of importance sampling is to approximate $\mathrm{E}_p(f(\boldsymbol{Z}))$, where f is a function and $\boldsymbol{Z}$ is a random vector with density p. One simulates a sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ from another density q, and by the law of large numbers, the weighted average $N^{-1}\sum_{i=1}^{N} w_i f(\boldsymbol{Z}_i)$ approaches $\mathrm{E}_p(f(\boldsymbol{Z}))$ as $N \to \infty$, where $w_i = p(\boldsymbol{Z}_i)/q(\boldsymbol{Z}_i)$ is the ratio of the target density to the actual density. For importance sampling to work, the support of q needs to cover that of p, and the method is efficient when q resembles p over the high-density regions. Applying this idea to (3.2), we see that the weights adjust the expectation of the *prima facie* estimating function to what it would be if the triplet $(\boldsymbol{X}_i, T_i, Y_i)$ had been sampled from a target population where T_i is independent of $\boldsymbol{X}_i$ — which, by strong ignorability, makes it independent of $\boldsymbol{\theta}_i$ as well. The non-stabilized weights use a uniform density for T_i as the target, whereas the stabilized weights use $P(T_i)$, which should be closer to $P(T_i \mid \boldsymbol{X}_i)$ and therefore more efficient.

In our example, a reasonable way to apply (3.2) is to estimate $P(T_i)$ by the density of a normal variate with mean $\bar{T} = N^{-1}\sum_{i=1}^{N} T_i$ and variance $(N-1)^{-1}\sum_{i=1}^{N}(T_i - \bar{(T)})^2$, and to estimate $P(T_i \mid \boldsymbol{X}_i)$ by the normal density with mean $\hat{T}_i$ and variance $(N-9)^{-1}\sum_{i=1}^{N}(T_i - \hat{T})^2$, where $\hat{T}_i$ is the fitted value from the OLS regression of T_i on covariates. We tried a condition where the treatment or T-model is correct, regressing T_i on $\boldsymbol{A}^*_{i1}, \ldots, \boldsymbol{A}^*_{i8}$, and a condition where the T-model is incorrect, regressing T_i on $\boldsymbol{A}_{i1}, \ldots, \boldsymbol{A}_{i8}$. Performance measures for these conditions are shown in Table 3.3. Comparing these results to those of the *prima facie* estimator, we see that the bias has been reduced by more than half. Unfortunately, these bias

Table 3.3: Performance of the importance-weighted estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	T-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	3.29	4.61	153	3.93	3.72
	Incorrect	3.41	4.07	169	3.96	3.69
$N = 1,000$	Correct	2.37	2.71	144	2.89	2.76
	Incorrect	2.68	3.88	136	3.32	2.97

reductions are accompanied by huge increases in variance. These variances do not drop as much as we would ordinarily expect as N increases from 200 to 1,000.

This is a classic case where importance sampling fails, because the target density $P(T_i)$ is more diffuse than the actual density $P(T_i \mid \boldsymbol{X}_i)$, causing the weights to be highly variable. The efficiency of importance sampling, relative to sampling N observations directly from the target, is $1/(1 + \mathrm{CV}(w)^2)$, where $\mathrm{CV}(w)$ is the coefficient of variation of the weights (Kong et al., 1994). Efficiency drops rapidly as the fit of the T-model improves. When R^2 from the T-model is 0.2, the relative efficiency is approximately 0.6, and when $R^2 = 0.4$, it is about 0.1, but in our example, R^2 exceeds 0.6, making the procedure very inefficient. Even when R^2 is low, the weights are sensitive to misspecification of the tails of $P(T_i \mid \boldsymbol{X}_i)$. For these reasons, we do not recommend importance weighting for a continuous treatment.

3.3.4 Inverse second-moment weighting

Fortunately, there is another weighting scheme that is more stable than importance weighting and does not rely on a full density $P(T_i \mid \boldsymbol{X}_i)$. Consider the observable random vector $\boldsymbol{B}_i Y_i = \boldsymbol{B}_i \boldsymbol{B}_i^\top \boldsymbol{\theta}_i$. Under strong ignorability, the mean of $\boldsymbol{B}_i Y_i$ is $\mathrm{E}[\, E(\boldsymbol{B}_i Y_i | \boldsymbol{X}_i)\,] = E[\, E(\boldsymbol{B}_i \boldsymbol{B}_i^\top | \boldsymbol{X}_i)\, E(\boldsymbol{\theta}_i | \boldsymbol{X}_i)\,]$. If we premultiply by the matrix $\boldsymbol{W}_i = [\, E(\boldsymbol{B}_i \boldsymbol{B}_i^\top | \boldsymbol{X}_i)\,]^{-1}$, which is a function of $\boldsymbol{X}_i$ but not T_i or $\boldsymbol{\theta}_i$, we get

a new random vector $\boldsymbol{W}_i\boldsymbol{B}_iY_i$ whose mean is

$$E[\,\boldsymbol{W}_i\,E(\boldsymbol{B}_i\boldsymbol{B}_i^\top|\boldsymbol{X}_i)\,E(\boldsymbol{\theta}_i|\boldsymbol{X}_i)\,]\;=\;E[\,E(\boldsymbol{\theta}_i|\boldsymbol{X}_i)\,]\;=\;\boldsymbol{\xi}. \tag{3.3}$$

Defining a new estimating function $\boldsymbol{U}_i = (\boldsymbol{W}_i\boldsymbol{B}_iY_i - \boldsymbol{\xi})$ and solving $\sum_{i=1}^N \boldsymbol{U}_i = \boldsymbol{0}$ gives

$$\hat{\boldsymbol{\xi}} \;=\; \frac{1}{N}\left(\sum_{i=1}^N \boldsymbol{W}_i\boldsymbol{B}_iY_i\right). \tag{3.4}$$

This method, which we call inverse second-moment weighting, is a natural extension of inverse probability of treatment weighting (IPTW) for a binary T_i (Hirano and Imbens, 2001). (Replacing $\boldsymbol{B}_i$ with $(T_i, 1-T_i)^\top$ and Y_i with $T_iY_i(1) + (1-T_i)Y_i(0)$ gives the usual IPTW estimator.) A modified version of (3.4) that normalizes the weights is

$$\hat{\boldsymbol{\xi}} \;=\; \left(\sum_{i=1}^N \boldsymbol{W}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\left(\sum_{i=1}^N \boldsymbol{W}_i\boldsymbol{B}_iY_i\right), \tag{3.5}$$

which can be regarded as the solution to $\sum_{i=1}^N \boldsymbol{U}_i = \boldsymbol{0}$ for $U_i = \boldsymbol{W}_i\boldsymbol{B}_i(Y_i - \boldsymbol{B}_i^\top\boldsymbol{\xi})$. Note that (3.5) is not a typical WLS regression; the weight applied to unit i is not a scalar but a matrix, and the "information" $\sum_i \boldsymbol{W}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top$ is asymmetric.

To use (3.4) or (3.5), we need to supply expected values for $\boldsymbol{B}_i\boldsymbol{B}_i^\top$, $i = 1, \ldots, N$ by fitting a model to predict T_i from $\boldsymbol{X}_i$. For example, if $\boldsymbol{B}_i = (1, T_i)^\top$, the T-model would have to consistently estimate $\mathrm{E}(T_i|\boldsymbol{X}_i)$ and $\mathrm{E}(T_i^2|\boldsymbol{X}_i) = \mathrm{Var}(T_i|\boldsymbol{X}_i) + [E(T_i|\boldsymbol{X}_i)]^2$. If the T-model is a linear regression with homoscedastic errors, we may set $\mathrm{E}(T_i|\boldsymbol{X}_i)$ to the ith fitted value and $\mathrm{Var}(T_i|\boldsymbol{X}_i)$ to the estimated residual variance.

We applied inverse second-moment weighting to our simulated example using a correct T-model (regressing T_i on $A^*_{i1}, \ldots, A^*_{i8}$) and an incorrect T-model (regressing T_i on $A_{i1}, \ldots, A_{i8}$). The performance of the normalized estimator (3.5) is summarized in Table 3.4; results for the non-normalized version (3.4) were very similar and

Table 3.4: Performance of the inverse second-moment weighted estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	T-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	-0.007	0.164	-1	0.405	0.274
	Incorrect	0.691	0.480	100	0.978	0.706
$N = 1,000$	Correct	-0.001	0.029	-0	0.171	0.112
	Incorrect	0.714	0.091	236	0.775	0.711

are not shown. When the T-model is correct, the estimator is unbiased and has smaller variance than the *prima facie* method. When the T-model is wrong, bias appears and variance increases. Even with an incorrect T-model, the bias relative to the *prima facie* method has dropped by about 90%, but for both sample sizes $N = 200$ and $N = 1,000$, the bias that remains is still large enough to impair confidence intervals and tests. Nevertheless, even with a misspecified T-model, this method is far superior to the *prima facie* estimator and to importance weighting.

Previous work has shown that IPTW for a binary treatment can be unstable; propensities close to zero or one produce weights that are large and error prone (Kang and Schafer, 2007). Inverse second-moment weighting for a continuous treatment fares better. To see why, note that in the linear case $\boldsymbol{B}_i = (1, T_i)^\top$, the determinant of $\boldsymbol{W}_i^{-1} = \mathrm{E}(\boldsymbol{B}_i \boldsymbol{B}_i^T | \boldsymbol{X}_i)$ is $\mathrm{Var}(T_i | \boldsymbol{X}_i)$. If the T-model is a standard linear regression, $\boldsymbol{W}_i$ will inflate only if $R^2 \to 1$, a scenario where causal inference should not even be attempted.

As basis functions are added to $\boldsymbol{b}(t)$, inverse second-moment weighting becomes less attractive, because it relies on higher moments of T_i. For the quadratic situation $\boldsymbol{b}(t) = (1, t, t^2)^\top$, we would need the T-model to correctly describe $\mathrm{E}(T_i^m \mid \boldsymbol{X}_i)$ for $m = 1, \ldots, 4$, which is almost as restrictive as requiring correct specification of the full density $P(T_i \mid \boldsymbol{X}_i)$.

3.3.5 Regression prediction

Estimators (3.4)–(3.5) rely on the random vector $\boldsymbol{W}_i\boldsymbol{B}_iY_i$ whose mean under a correct T-model is $\boldsymbol{\xi}$. Alternatively, if we had a model for $\boldsymbol{\theta}_i$ given $\boldsymbol{X}_i$, we could build an estimator from $\hat{\boldsymbol{\theta}}_i = \mathrm{E}(\boldsymbol{\theta}_i|\boldsymbol{X}_i)$, another function of observed data whose mean is $\boldsymbol{\xi}$. Taking $\boldsymbol{U}_i = (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\xi})$ and solving $\sum_{i=1}^N \boldsymbol{U}_i = \boldsymbol{0}$ gives $\hat{\boldsymbol{\xi}} = N^{-1}\sum_{i=1}^N \hat{\boldsymbol{\theta}}_i$. Because of the one-to-one correspondence between $\boldsymbol{\theta}_i$ and $Y_i(t)$, the model for predicting $\boldsymbol{\theta}_i$ from $\boldsymbol{X}_i$ will be called a Y-model. The estimator is consistent if the Y-model correctly describes $\mathrm{E}(\boldsymbol{\theta}_i \mid \boldsymbol{X}_i)$.

To use this regression-prediction method, the parameters of the Y-model must be estimated. A reasonable starting point for a Y-model is the standard multivariate regression $\boldsymbol{\theta}_i|\boldsymbol{X}_i \sim N(\boldsymbol{\nu} + \boldsymbol{\Gamma}^\top\boldsymbol{X}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\nu}$ $(k \times 1)$ and $\boldsymbol{\Gamma}$ $(p \times k)$ are the parameters of interest, and $\boldsymbol{\Sigma}$ $(k \times k)$ is a nuisance. (The assumptions of normality and constant covariance are not essential, and we include them only for illustration.) We can rewrite this model as $\boldsymbol{\theta}_i \mid \boldsymbol{X}_i \sim N(\boldsymbol{\Gamma}^{*\top}\boldsymbol{X}_i^*, \boldsymbol{\Sigma})$, where $\boldsymbol{\Gamma}^{*\top} = [\boldsymbol{\nu}, \boldsymbol{\Gamma}^\top]$ and $\boldsymbol{X}_i^{*\top} = (1, \boldsymbol{X}_i^\top)$. Under strong ignorability, conditioning on T_i does not change the model, $\boldsymbol{\theta}_i \mid T_i, \boldsymbol{X}_i \sim N(\boldsymbol{\Gamma}^{*\top}\boldsymbol{X}_i^*, \boldsymbol{\Sigma})$, which implies that $Y_i \mid T_i, \boldsymbol{X}_i \sim N(\boldsymbol{B}_i^\top\boldsymbol{\Gamma}^{*\top}\boldsymbol{X}_i^*, \boldsymbol{B}_i^\top\boldsymbol{\Sigma}\boldsymbol{B}_i)$. We can write $\boldsymbol{B}_i^\top\boldsymbol{\Gamma}^{*\top}\boldsymbol{X}_i^* = \boldsymbol{Z}_i^\top\mathrm{vec}(\boldsymbol{\Gamma}^*)$, where $\boldsymbol{Z}_i^\top = (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)^\top = (B_{i1}\boldsymbol{X}_i^{*\top}, \ldots, B_{ik}\boldsymbol{X}_i^{*\top})$, and where $\mathrm{vec}(\cdot)$ vectorizes a matrix by stacking its columns. Therefore, we can estimate $\boldsymbol{\Gamma}^*$ by ordinary least squares,

$$\mathrm{vec}\,(\hat{\boldsymbol{\Gamma}}^*) = \left(\sum_{i=1}^N \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)^{-1}\left(\sum_{i=1}^N \boldsymbol{Z}_iY_i\right), \tag{3.6}$$

which gives the regression predictions $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\nu}} + \hat{\boldsymbol{\Gamma}}^\top\boldsymbol{X}_i$, $i = 1, \ldots, N$. If the assumptions of normality and constant covariance hold, a more efficient estimator than (3.6) would come from maximizing the joint likelihood for $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\Sigma}$, which we will not pursue in this chapter. In this model, each covariate in $\boldsymbol{X}_i$ predicts every element of $\boldsymbol{\theta}_i$. Simpler models that remove some of the terms from $\boldsymbol{Z}_i$ are worth considering,

Table 3.5: Performance of the regression-prediction estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	Y-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	-0.005	0.157	-1	0.397	0.268
	Incorrect	0.616	0.500	87	0.938	0.654
$N = 1,000$	Correct	0.000	0.028	0	0.167	0.117
	Incorrect	0.660	0.089	221	0.725	0.661

along with strategies for variable selection when the pool of available covariates is large.

We applied this method to our example using a correct Y-model with $\boldsymbol{X}_i = (A^*_{i1}, \ldots, A^*_{i8})^\top$ and an incorrect Y-model with $\boldsymbol{X}_i = (A_{i1}, \ldots, A_{i8})^\top$. Results are shown in Table 3.5. These results are similar to those in Table 3.4. When the Y-model is correct, the estimator is unbiased. When the Y-model is wrong, the biases are similar to those of inverse second-moment weighting under the wrong T-model, and the variances are also comparable.

3.3.6 Prediction with a residual bias correction

It may be advantageous to build an estimator that combines features of a T-model and a Y-model. Adapting a strategy from Robins and Rotnitzky (1995), we may start with an estimating function based on weighting and augment it with a term that involves prediction,

$$\boldsymbol{U}_i \;=\; \boldsymbol{W}_i\boldsymbol{B}_i(Y_i - \boldsymbol{B}_i^\top\boldsymbol{\xi}) \;+\; (\boldsymbol{I} - \boldsymbol{W}_i\boldsymbol{B}_i\boldsymbol{B}_i^\top)\,(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\xi}).$$

Solving $\sum_{i=1}^N \boldsymbol{U}_i = \boldsymbol{0}$ gives

$$\hat{\boldsymbol{\xi}} \;=\; \frac{1}{N}\sum_{i=1}^N \hat{\boldsymbol{\theta}}_i \;+\; \frac{1}{N}\sum_{i=1}^N \boldsymbol{W}_i\boldsymbol{B}_i(Y_i - \hat{Y}_i), \tag{3.7}$$

Table 3.6: Performance of prediction with residual bias correction estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	T-model	Y-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	Correct	−0.005	0.157	−1	0.397	0.268
	Correct	Incorrect	0.043	0.247	9	0.499	0.343
	Incorrect	Correct	−0.003	0.166	−1	0.407	0.268
	Incorrect	Incorrect	0.616	0.500	87	0.938	0.654
$N = 1,000$	Correct	Correct	0.000	0.028	0	0.167	0.117
	Correct	Incorrect	0.009	0.041	4	0.203	0.145
	Incorrect	Correct	0.000	0.030	0	0.174	0.117
	Incorrect	Incorrect	0.660	0.089	221	0.725	0.661

where $\hat{Y}_i = \boldsymbol{B}_i^\top \hat{\boldsymbol{\theta}}_i$. This estimator bears a strong resemblance to general regression estimators in the survey literature, part of a more general class of calibration estimators (Deville and Särndal, 1992). It is doubly robust, which means that it is consistent if either of the models is true (Scharfstein et al., 1999). If the Y-model is correct, then the first term in (3.7) is unbiased for $\boldsymbol{\xi}$ and the second term has mean zero even if the T-model is wrong. If the Y-model is incorrect, the first term is biased, but the second term gives a consistent estimate of (minus one times) the bias from the Y-model if the T-model is correct.

The performance of this estimator is shown in Table 3.6 for various combinations of T and Y-models. As predicted by double robustness, the bias vanishes if either model is correct, but some bias remains if both models are wrong. Comparing Table 3.6 to Table 3.4, we see the effect of adding a Y-model to a method that relies on a T-model. If the T-model is correct, we see a small improvement in precision if the Y-model is correct and a small deterioration in precision if the Y-model is wrong. If the T-model is wrong, adding a correct Y-model wipes out the bias and drastically improves the precision. Adding a wrong Y-model to a wrong T-model gives a slight reduction in bias with little or no increase in variance. Similarly, if we compare these results to those in Table 3.5, we see the effect of adding a T-model to

a method that relies on a Y-model. The story is similar; there is a potential for gain and little risk. The combination of inverse second-moment weighting and regression prediction does no worse than either method alone, and sometimes it does much better. In the unfortunate but realistic situation where both models are wrong, a pernicious bias remains.

3.3.7 Prediction from a weighted regression

For each of our estimators (except for the *prima facie* method), the estimating function depends on a T-model and/or a Y-model whose parameters must themselves be estimated by another set of equations. For regression prediction, we estimated $\boldsymbol{\Gamma}^*$ by regressing Y_i on $\boldsymbol{Z}_i = (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)$, using the estimating function

$$\begin{aligned} \boldsymbol{S}_i &= \boldsymbol{Z}_i(\, Y_i - \boldsymbol{B}_i^\top \boldsymbol{\Gamma}^{*\top} \boldsymbol{X}_i^* \,) \\ &= (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)(\, Y_i - (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)^\top \mathrm{vec}(\boldsymbol{\Gamma}^*) \,). \end{aligned}$$

The solution to $\sum_{i=1}^N \boldsymbol{S}_i = \boldsymbol{0}$ given by (3.6) may also be written as

$$\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*) \;=\; \left(\sum_{i=1}^N \left[\, (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \otimes (\boldsymbol{X}_i^* \boldsymbol{X}_i^{*\top}) \,\right] \right)^{-1} \left(\sum_{i=1}^N (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)\, Y_i \right).$$

Suppose we perturb $\hat{\boldsymbol{\Gamma}}^*$ by applying a different fitting criterion so that the new $\hat{Y}_i$'s cause the second term on the right-hand side of (3.7) to vanish. This is another way to use a T-model to calibrate a faulty Y-model. The estimating function that accomplishes this is

$$\boldsymbol{S}_i \;=\; \boldsymbol{W}_i (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)(\, Y_i - (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)^\top \mathrm{vec}(\boldsymbol{\Gamma}^*) \,),$$

Table 3.7: Performance of prediction from weighted regression estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	T-model	Y-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	Correct	0.010	0.212	2	0.461	0.299
	Correct	Incorrect	−0.040	1.472	−3	1.213	0.449
	Incorrect	Correct	−0.025	0.449	−4	0.670	0.351
	Incorrect	Incorrect	0.767	3.533	41	2.029	0.803
$N = 1,000$	Correct	Correct	−0.000	0.028	−0	0.169	0.115
	Correct	Incorrect	−0.008	0.046	−3	0.215	0.149
	Incorrect	Correct	−0.001	0.042	−0	0.205	0.137
	Incorrect	Incorrect	0.743	0.093	243	0.803	0.742

and the new solution to $\sum_{i=1}^{N} \boldsymbol{S}_i = \mathbf{0}$ is

$$\text{vec}(\hat{\boldsymbol{\Gamma}}^*) = \left(\sum_{i=1}^{N} \left[(\boldsymbol{W}_i \boldsymbol{B}_i \boldsymbol{B}_i^\top) \otimes (\boldsymbol{X}_i^* \boldsymbol{X}_i^{*\top}) \right] \right)^{-1} \left(\sum_{i=1}^{N} \left[(\boldsymbol{W}_i \boldsymbol{B}_i) \otimes \boldsymbol{X}_i^* \right] Y_i \right). \quad (3.8)$$

To compute (3.8), we first estimate a T-model to obtain the inverse second-moment weights. After getting (3.8), we obtain regression predictions $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\nu}} + \hat{\boldsymbol{\Gamma}}^\top \boldsymbol{X}_i$ for $i = 1, \ldots, N$, and the resulting estimator $\hat{\boldsymbol{\xi}} = N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i$ is doubly robust. We applied this method to our example with T and Y-models that are correct and incorrect (Table 3.7). The pattern of bias is similar to that in Table 3.6, but the new method is less efficient, especially when $N = 200$.

3.3.8 Propensity-spline prediction

Little and An (2004) proposed an imputation method that allows the mean of an incomplete variable to vary as a flexible function of an estimated propensity score. The method, which they call propensity-spline prediction, produces an estimate of a population mean that is doubly robust. Schafer and Kang (2008) applied this to causal inference with a binary treatment, and a similar method can be used with a continuous treatment.

The regression prediction method of Section 3.3.5 capitalized on $P(\boldsymbol{\theta}_i \mid T_i, \boldsymbol{X}_i) = P(\boldsymbol{\theta}_i \mid \boldsymbol{X}_i)$, which follows from strong ignorability. If $\boldsymbol{\pi}_i = \boldsymbol{\pi}(\boldsymbol{X}_i, \boldsymbol{\psi})$ is a propensity function as defined in Section 3.2.1, then the same property holds conditionally on $\boldsymbol{\pi}_i$, $P(\boldsymbol{\theta}_i \mid T_i, \boldsymbol{\pi}_i) = P(\boldsymbol{\theta}_i \mid \boldsymbol{\pi}_i)$. It is possible to construct a prediction estimator based on a Y-model that allows $\mathrm{E}(\theta_i \mid \boldsymbol{\pi}_i)$ to vary with $\boldsymbol{\pi}_i$ in a flexible manner. For example, we may create basis functions that span a space of splines in $\hat{\boldsymbol{\pi}}_i$, and use this basis instead of $\boldsymbol{X}_i$ to predict $\boldsymbol{\theta}_i$. The resulting estimator will have low bias if the propensity model is correct, but it may be inefficient, because $\boldsymbol{X}_i$ may carry information about $\boldsymbol{\theta}_i$ beyond $\hat{\boldsymbol{\pi}}_i$ that should not be ignored. A better strategy is to add $\hat{\boldsymbol{\pi}}_i$-basis predictors to a Y-model that already contains $\boldsymbol{X}_i$ and proceed with the method of Section 3.3.5 (If the basis includes a term that is linear in $\boldsymbol{X}_i$, then that term or an element of $\boldsymbol{X}_i$ will need to be removed to prevent collinearity.) If the Y-model is correctly specified before the $\hat{\boldsymbol{\pi}}_i$-basis is added, no bias is incurred by the extra terms. If the Y-model is wrong, the extra terms will reduce the bias if the T-model model is correct.

We applied this method to our example by estimating $\pi_i = \mathrm{E}(T_i \mid \boldsymbol{X}_i)$ under a correct T-model (based on $A^*_{i1}, \ldots, A^*_{i8}$) and an incorrect T-model (based on $A_{i1}, \ldots, A_{i8}$). For each, we used a natural cubic spline basis with interior knots at the quintiles of $\hat{\pi}_i$ and boundary knots at the minimum and maximum. We added the basis functions as predictors to a Y-model that was correct (based on $A^*_{i1}, \ldots, A^*_{i8}$) and a Y-model that was wrong (based on $A_{i1}, \ldots, A_{i8}$), removing a predictor when necessary to prevent collinearity. Results are shown in Table 3.8. Performance is very similar to regression estimation with a residual bias correction (Table 3.6). The method is essentially unbiased when the T-model or Y-model is correct, but harmful bias remains when both models are wrong.

Without explicitly modeling $Y_i(t)$, Imai and van Dyk (2004) estimated the ADRF within subclasses of $\hat{\pi}_i$ and pooled the results across the subclasses. They

Table 3.8: Performance of the propensity-spline prediction estimator for ξ_2 over 1,000 samples from the artificial population.

Sample size	T-model	Y-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	Correct	−0.010	0.165	−3	0.406	0.282
	Correct	Incorrect	−0.011	0.269	−2	0.518	0.352
	Incorrect	Correct	−0.011	0.169	−3	0.411	0.273
	Incorrect	Incorrect	0.609	0.544	83	0.956	0.664
$N = 1,000$	Correct	Correct	0.000	0.028	0	0.167	0.117
	Correct	Incorrect	−0.005	0.036	−2	0.189	0.133
	Incorrect	Correct	−0.000	0.028	−0	0.168	0.117
	Incorrect	Incorrect	0.646	0.089	217	0.711	0.641

Table 3.9: Performance of the propensity-spline prediction estimator for ξ_2 in the spirit of Imai and van Dyk (2004) over 1,000 samples from the artificial population.

Sample size	T-model	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Correct	−0.014	0.204	−3	0.452	0.302
	Incorrect	0.679	0.532	93	0.996	0.713
$N = 1,000$	Correct	−0.001	0.028	−0	0.168	0.114
	Incorrect	0.707	0.089	236	0.765	0.700

also applied a varying-coefficient model that describes $\mathrm{E}(Y_i(t) \mid \hat{\pi}_i)$ as a smooth function of $\hat{\pi}_i$ and t. Those methods resemble prediction based on a $\hat{\pi}_i$-spline with no additional covariates. The performance of prediction from a $\hat{\pi}_i$-spline alone under correct and incorrect T-models is shown in Table 3.9. Comparing these results to Table 3.8, we see that adding $\boldsymbol{X}_i$ to propensity-spline prediction can reduce bias and increase efficiency if the additional predictors are the correct ones (corresponding to a correct Y-model) but reduce efficiency if they are not. This suggests it may be fruitful to try an intermediate strategy, beginning with the $\hat{\pi}_i$ basis as the minimal set of predictors, and adding components of $\boldsymbol{X}_i$ only if they substantially improve the fit.

3.3.9 Strategies for additional bias reduction

With the exception of importance weighting, all of the methods we tried eliminated bias when the crucial modeling assumptions were met. Under more realistic conditions where the T- and Y-models were misspecified, any of the methods could remove about 90% of the bias of the *prima facie* estimator, but the last 10% was still enough to seriously impair inferences about the parameter of interest. Fortunately, more strategies for reducing bias are available. For example, when fitting a T-model, we might search for nonlinear effects and interactions among covariates. Propensity scores for binary treatments have been estimated by neural networks (King and Zeng, 2001), boosted regression trees (McCaffrey et al., 2004), and machine learning (Lee et al., 2010), and Zhu et al. (2015) applied boosted regression to a continuous treatment.

To see if similar strategies could work for our example, we revisited the condition where the T-model (and, where applicable, the Y-model) was incorrect, but for $N = 200$, we enriched the T-model by including all two-way interactions among $A_{i1}, \ldots, A_{i8}$; and for $N = 1,000$, we included two- and three-way interactions. Results are shown in Table 3.10. Importance weighting is still unstable, but the bias of the other methods has been drastically reduced, and in many cases it has fallen below the threshold (% Bias ≈ 50) where confidence intervals and hypothesis tests are no longer seriously impaired. Inverse second-moment weighting has the best RMSE for $N = 200$, and propensity-spline prediction is best for $N = 1,000$. Prediction with residual bias correction performs well at both sample sizes, and prediction from weighted regression is effective for $N = 1,000$ but unstable for $N = 200$.

Table 3.10: Performance of estimators for ξ_2 over 1,000 samples from the artificial population using incorrect Y-models (where applicable) and misspecified but rich T-models.

Sample size	Method	Bias	Var.	% Bias	RMSE	MAE
$N = 200$	Importance weighting	3.87	2.54	243	4.19	4.10
	Inverse second-moment wt.	0.225	0.504	32	0.745	0.491
	Pred. + resid. bias correction	0.268	0.496	38	0.753	0.493
	Pred. from weighted reg.	0.249	14.9	6	3.87	0.570
	Propensity-spline pred.	0.166	0.585	22	0.783	0.531
$N = 1,000$	Importance weighting	2.75	1.89	200	3.07	3.07
	Inverse second-moment wt.	0.144	0.083	50	0.322	0.215
	Pred. + resid. bias correction	0.158	0.082	55	0.327	0.225
	Pred. from weighted reg.	0.133	0.087	45	0.324	0.219
	Propensity-spline pred.	0.106	0.084	36	0.308	0.204

3.4 Discussion

It has become common practice, especially in epidemiology, to avoid assumptions about unit-level causal effects and define the parameters of interest as contrasts among average potential outcomes in a population (Maldonado and Greenland, 2002). In contrast, we mapped the $Y_i(t)$'s to random vectors in $\mathbb{R}^k$ and then averaged these vectors to obtain population-level marginal effects. This mapping leads to a new method of inverse second-moment weighting that improves upon the importance weights of Robins et al. (2000) when the treatment variable is continuous. The mapping also facilitates prediction-based and dual-modeling approaches that are described here for the first time.

Because this work is still at an early stage, we have eschewed applications with real data and focused on one simulated population where the ADRF is linear. In our simulations, all of the techniques from Sections 3.3.4–3.3.8 performed well when their assumed models were correct. When the models were misspecified, best performance came from methods that relied on incorrect but rich T-models (Table 3.10). In

another simulation study, Zhu et al. (2015) used a population of linear curves with varying intercepts and a common slope for all units. They fit marginal structural models to samples of $N = 500$, estimating the importance-weight denominators $P(T_i \mid \boldsymbol{X}_i)$ by boosted regression. We replicated some of their simulations and found that our methods from Sections 3.3.4–3.3.8 outperformed theirs. In their Scenario (A) of Table 2, where the treatment model had $R^2 \approx 0.4$, propensity-spline prediction with a rich T-model achieved an MSE about 50% lower than their best method. In Scenarios (B) and (C), where the treatment models were weaker ($R^2 \approx 0.2$ or less), efficiency gains were less dramatic but still substantial.

We noted in Section 3.3.4 that, if we add basis functions to $\boldsymbol{b}(t)$ to fit a nonlinear ADRF, inverse second-moment weighting becomes less attractive, because it requires the treatment model to correctly describe higher moments of T_i. Propensity-spline prediction is more promising, because it depends on $P(T_i \mid \boldsymbol{X}_i)$ only through the estimated propensity functions $\hat{\boldsymbol{\pi}}_i$. However, as the number of basis functions and the pool of available covariates grows, prediction-based estimators may become unstable unless the Y-model is trimmed. Instead of using the full set of predictors $\boldsymbol{Z}_i = (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)$, we may need to simplify the model by omitting some of the interactions between $\boldsymbol{X}_i$ and the higher-order terms in $\boldsymbol{B}_i$. Strategies for choosing a suitable parsimonious Y-model for estimating nonlinear ADRF's is an important topic for future research.

Throughout this chapter, we have supposed the response variable is continuous. To adapt our approach to a discrete response, one could assume $g(\, E(Y_i(t) \mid \boldsymbol{\theta}_i)\,) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t)$, where g is a monotonic link function (e.g. logistic for a binary outcome). The population ADRF will no longer be determined solely by $\mathrm{E}(\boldsymbol{\theta}_i)$, but will require averaging $g^{-1}(\boldsymbol{\theta}_i^\top \boldsymbol{b}(t))$ over the distribution of $\boldsymbol{\theta}_i$. After averaging, the ADRF will not necessarily follow $g(\mu(t)) = \boldsymbol{\xi}^\top \boldsymbol{b}(t)$ for some $\boldsymbol{\xi} \in \mathbb{R}^k$. Prediction estimators that apply the same link function at the unit and population levels will

have a built-in incoherence that should not matter much in practice (in a real application, the assumed link function will not exactly hold at either level) but this discrepancy does require us to be careful about how we define the inferential target.

3.5 Appendix A

This section provides more detail to the main example in Chapter 3 and provides steps to derive the equations for the conditional mean and variance of the outcome given treatment. The goal is to find the conditional distribution of $(\theta_{i2}|T_i, Y_i)$ and of $(\theta_{i1}|\theta_{i2}, T_i, Y_i)$ which are the random coefficients that describe $Y_i = \theta_{i1} + \theta_{i2}T_i$. To derive these conditional distributions, start with the distribution of

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ T_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi_2 \\ \xi_2 \\ \kappa \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{12} & \omega_{22} & \omega_{23} \\ \omega_{13} & \omega_{23} & \omega_{33} \end{pmatrix} \right). \tag{3.9}$$

The potential outcome path for unit i is $Y_i(t) = \theta_{i1} + \theta_{i2}t$, with

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi_2 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix} \right)$$

which means that $\mathrm{E}[Y_i(t)] = \xi_1 + \xi_2 t$, where ξ_2 is the population average treatment effect and θ_{i2} is the unit-level treatment effect for unit i.

The conditional distribution of $\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}$ given (T_i, Y_i) is a degenerate (singular) bivariate normal, which can be written as $P(\theta_{i2}|T_i, Y_i) \times P(\theta_{i1}|\theta_{i2}, T_i, Y_i)$ where $P(\theta_{i1}|\theta_{i2}, T_i, Y_i)$ is a point mass at $\theta_{i1} = Y_i - \theta_{i2}T_i$ and $P(\theta_{i2}|T_i, Y_i)$ is normally distributed. The distribution of $P(\theta_{i2}|T_i, Y_i)$ follows below.

1. The distribution of

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} |T_i$$

is given by

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} |T_i \sim \mathcal{N}\left(\begin{pmatrix} \xi_2 + \frac{\omega_{13}}{\omega_{33}}(T-\kappa) \\ \xi_2 + \frac{\omega_{23}}{\omega_{33}}(T-\kappa) \end{pmatrix}, \begin{pmatrix} \omega_{11} - \frac{\omega_{13}^2}{\omega_{33}} & \omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}} \\ \omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}} & \omega_{22} - \frac{\omega_{23}^2}{\omega_{33}} \end{pmatrix} \right).$$

2. Transform the previous step to the distribution of

$$\begin{pmatrix} \theta_{i2} \\ Y_i \end{pmatrix} |T_i$$

by taking

$$\begin{pmatrix} \theta_{i2} \\ Y_i \end{pmatrix} |T_i = \begin{pmatrix} 0 & 1 \\ 1 & T_i \end{pmatrix} \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix},$$

$$\begin{pmatrix} \theta_{i2} \\ Y_i \end{pmatrix} |T_i \sim \mathcal{N}\left(\begin{pmatrix} \xi_2 + \frac{\omega_{23}}{\omega_{33}}(T-\kappa) \\ \mathrm{E}[Y|T] \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) & \mathrm{Cov}(\theta_{i2}, Y|T) \\ \mathrm{Cov}(\theta_{i2}, Y|T) & \mathrm{Cov}(Y, Y|T) \end{pmatrix} \right),$$

$$\begin{aligned} \mathrm{E}[Y|T] &= \mathrm{E}[\theta_{i1} + \theta_{i2}T|T] \\ &= \mathrm{E}[\theta_{i1}|T] + T\mathrm{E}[\theta_{i2}|T] \\ &= \xi_2 + \frac{\omega_{13}}{\omega_{33}}(T-\kappa) + T(\xi_2 + \frac{\omega_{23}}{\omega_{33}}(T-\kappa)), \end{aligned}$$

$$\mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) = \omega_{22} - \frac{\omega_{23}^2}{\omega_{33}},$$

$$\begin{aligned}\mathrm{Cov}(\theta_{i2}, Y|T) &= \mathrm{Cov}(\theta_{i2}, \theta_{i1} + \theta_{i2}T|T) \\ &= \mathrm{Cov}(\theta_{i2}, \theta_{i1}|T) + T\,\mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) \\ &= \omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}} + T\,(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(Y, Y|T) &= \mathrm{Cov}(\theta_{i1} + \theta_{i2}T, \theta_{i1} + \theta_{i2}T|T) \\ &= \mathrm{Cov}(\theta_{i1}, \theta_{i1}|T) + 2T\mathrm{Cov}(\theta_{i1}, \theta_{i2}|T) + T^2\mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) \\ &= \omega_{11} - \frac{\omega_{13}^2}{\omega_{33}} + 2T(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}}) + T^2(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}).\end{aligned}$$

3. Find the distribution of $\theta_{i2}|T_i, Y_i$ by using the previous step and conditioning on Y_i. In other words, the previous step conditioned on T_i, and this step conditions on Y_i in addition to T_i. The result is

$$\theta_{i2}|T_i, Y_i \sim \mathcal{N}\left(\mathrm{E}(\theta_{i2}|T_i, Y_i), \mathrm{Var}(\theta_{i2}|T_i, Y_i)\right),$$

where

$$\mathrm{E}(\theta_{i2}|T_i, Y_i) = \mathrm{E}[\theta_{i2}|T] + \frac{\mathrm{Cov}(\theta_{i2}, Y|T)}{\mathrm{Cov}(Y, Y|T)}(Y - \mu_Y),$$

$$\mathrm{Var}(\theta_{i2}|T_i, Y_i) = \mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) - \mathrm{Cov}(\theta_{i2}, Y|T)\mathrm{Cov}(Y, Y|T)\mathrm{Cov}(Y, \theta_{i2}|T),$$

$$\begin{aligned}\mathrm{E}[\theta_{i2}|T] + \frac{\mathrm{Cov}(\theta_{i2}, Y|T)}{\mathrm{Cov}(Y, Y|T)}(Y - \mu_Y) = &\left(\xi_2 + \frac{\omega_{23}}{\omega_{33}}(T - \kappa)\right) + \\ &\left(\frac{\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}} + T(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}})}{\omega_{11} - \frac{\omega_{13}^2}{\omega_{33}} + 2T(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}}) + T^2(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}})}\right)(T - \kappa),\end{aligned}$$

and

$$\mathrm{Cov}(\theta_{i2}, \theta_{i2}|T) - \mathrm{Cov}(\theta_{i2}, Y|T)\mathrm{Cov}(Y, Y|T)^{-1}\mathrm{Cov}(Y, \theta_{i2}|T) =$$
$$\left(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}\right) - \left(\frac{\left(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}} + T\left(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}\right)\right)^2}{\left(\omega_{11} - \frac{\omega_{13}^2}{\omega_{33}}\right) + 2T\left(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}}\right) + T^2\left(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}\right)}\right).$$

4. $\theta_{i1}|\theta_{i2}, T_i, Y_i = (Y_i - \theta_{i2}T_i)|\theta_{i2}, T_i, Y_i$ which is a point mass at $Y_i - \theta_{i2}T_i$.

A special case occurs when all units have the same treatment effect. Under this homogeneous treatment-effect condition, $\omega_{22} = 0 \implies \omega_{12} = 0$. In the set-up above, this can be visualized as each unit having a DRF with parallel lines but random intercepts. Suppose

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ T_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \\ T_i \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{12} & \omega_{22} & \omega_{23} \\ \omega_{13} & \omega_{23} & \omega_{33} \end{pmatrix}\right).$$

With homogeneous treatment effects, we have

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ T_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \\ T_i \end{pmatrix}, \begin{pmatrix} \omega_{11} & 0 & \omega_{13} \\ 0 & 0 & 0 \\ \omega_{13} & 0 & \omega_{33} \end{pmatrix}\right).$$

Also,

$$\begin{pmatrix} Y_i(t) \\ T_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi_1 + \xi_2 t \\ \kappa \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YT} \\ \sigma_{YT} & \omega_{33} \end{pmatrix}\right),$$

where

$$\begin{aligned}\sigma_Y^2 &= \mathrm{Var}(\theta_{i1} + \theta_{i2}t) \\ &= \mathrm{Var}(\theta_{i1}) + t^2\mathrm{Var}(\theta_{i2}) + 2\mathrm{Cov}(\theta_{i1}, \theta_{i2}t) \\ &= \omega_{11} + t^2\omega_{22} + 2t\omega_{12}\end{aligned}$$

and

$$\begin{aligned}\sigma_{YT} &= \mathrm{Cov}(Y_i(t), T_i) \\ &= \mathrm{Cov}(\theta_{i1} + \theta_{i2}t, T_i) \\ &= \mathrm{Cov}(\theta_{i1}, T_i) + t\mathrm{Cov}(\theta_{i2}, T_i) \\ &= \sigma_{\theta_{i1},T} + t\sigma_{\theta_{i2},T}\end{aligned}$$

This implies that for any $t \in \mathrm{T}$, the distribution of $Y_i(t)|T_i$ is normal with

$$\mathrm{E}[Y_i(t)|T_i] = \xi_1 + \xi_2 t + \frac{\omega_{13} + t\omega_{23}}{\omega_{33}}(T_i - \kappa)$$

and

$$\mathrm{Var}(Y_i(t)|T_i) = (\omega_{11} + \omega_{22} + 2\omega_{12}^2) - \frac{(\omega_{13} + t\omega_{23})^2}{\omega_{33}},$$

which follows from well-known properties of bivariate normal distributions.

In the special case of $T_i = t$, this becomes

$$\begin{aligned}\mathrm{E}(Y_i(t)|T_i = t) &= (\xi_1 + \xi_2\, t) + \left(\frac{\omega_{13} + \omega_{23}\, t}{\omega_{33}}\right)(t - \kappa) \\ &= \left(\xi_1 - \frac{\omega_{13}}{\omega_{33}}\,\kappa\right) + \left(\xi_2 + \frac{\omega_{13}}{\omega_{33}} - \frac{\omega_{23}}{\omega_{33}}\,\kappa\right)\, t + \left(\frac{\omega_{22}}{\omega_{33}}\right)\, t^2,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[Y_i(t)|T_i = t] &= \omega_{11} + 2\,\omega_{12}\,t + \omega_{22}\,t^2 - \frac{(\omega_{13} + \omega_{23}\,t)^2}{\omega_{33}} \\ &= \left(\omega_{11} - \frac{\omega_{13}^2}{\omega_{33}}\right) + 2\left(\omega_{12} - \frac{\omega_{13}\omega_{23}}{\omega_{33}}\right)\, t + \left(\omega_{22} - \frac{\omega_{23}^2}{\omega_{33}}\right)\, t^2,\end{aligned}$$

which implies that the joint distribution of (Y_i, T_i) is

$$\begin{aligned} P(T_i, Y_i) &= P(T_i) \times P(Y_i|T_i) \\ &= \mathcal{N}(\kappa, \omega_{33}) \times \mathcal{N}(\mathrm{E}(Y_i|T_i), \mathrm{Var}(Y_i|T_i)) \end{aligned}$$

The joint distribution will not necessarily be bivariate normal, because $\mathrm{E}(Y_i|T_i)$ and $\mathrm{Var}(Y_i|T_i)$ are each quadratic in T_i.

The prima facie estimate or the regression of Y_i on T_i estimates $\mathrm{E}(Y_i|T_i)$. If we happen to have a homogeneous treatment effect, then

$$\mathrm{E}[Y_i(t)|T_i = t] = \left(\xi_1 - \frac{\kappa\,\omega_{13}}{\omega_{33}}\right) + \left(\xi_2 + \frac{\omega_{13}}{\omega_{33}}\right) t,$$

which is not the same as

$$\mathrm{E}[Y_i(t)] = \xi_1 + \xi_2\, t,$$

which means the slope of the prima facie regression line is biased by $\left(\frac{\omega_{13}}{\omega_{33}}\right)$.

3.6 Appendix B

The following shows that inverse second moment weighting estimator is a generalization of the IPTW estimator. Suppose $T_i \in \{0, 1\}$. Recall that $U(\xi) = B_i\,(Y_i - B_i^\top \xi) = B_i\,B_i^\top\,(\theta_i - \xi)$. Suppose the estimating equation is modified to $U_{mod}(\xi) = w_i\,B_i\,(Y_i - B_i^\top \xi)$, where

$$
\begin{aligned}
w_i &= E\left[\begin{pmatrix} T_i & 0 \\ 0 & 1-T_i \end{pmatrix}\middle| X_i\right]^{-1} \\
&= \begin{bmatrix} P(T_i|X_i) & 0 \\ 0 & 1-P(T_i|X_i) \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \frac{1}{P(T_i|X_i)} & 0 \\ 0 & \frac{1}{1-P(T_i|X_i)} \end{bmatrix}.
\end{aligned}
$$

Solving this set of estimating equations gives the same estimates as the IPTW method. To see this, note that

$$
\begin{aligned}
U_{mod}(\xi) &= \begin{bmatrix} \frac{1}{P(T_i|X_i)} & 0 \\ 0 & \frac{1}{1-P(T_i|X_i)} \end{bmatrix} \begin{bmatrix} T_i \\ 1-T_i \end{bmatrix} (Y_i - (T_i E(Y_i(1)) + (1-T_i)E(Y_i(0))) \\
&= \begin{bmatrix} \frac{T_i}{P(T_i|X_i)} \\ \frac{1-T_i}{1-P(T_i|X_i)} \end{bmatrix} (Y_i - (T_i\, E(Y_i(1)) + (1-T_i)\, E(Y_i(0))) \\
&= \begin{bmatrix} \frac{T_i Y_i}{P(T_i|X_i)} - \frac{T_i^2\, E(Y_i(1))}{P(T_i|X_i)} - \frac{T_i\,(1-T_i)E(Y_i(0)))}{P(T_i|X_i)} \\ \frac{(1-T_i)Y_i}{1-P(T_i|X_i)} - \frac{(1-T_i)T_i\, E(Y_i(1))}{1-P(T_i|X_i)} - \frac{(1-T_i)^2\, E(Y_i(0)))}{1-P(T_i|X_i)} \end{bmatrix}.
\end{aligned}
$$

Setting $U_{mod}(\xi)$ to 0 and summing over the dataset gives

$$
\begin{bmatrix} \sum_{i=1}^N \frac{T_i Y_i}{P(T_i|X_i)} \\ \sum_{i=1}^N \frac{(1-T_i)Y_i}{1-P(T_i|X_i)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N \frac{T_i^2\, E(Y_i(1))}{P(T_i|X_i)} + \sum_{i=1}^N \frac{T_i\,(1-T_i)E(Y_i(0)))}{P(T_i|X_i)} \\ \sum_{i=1}^N \frac{(1-T_i)T_i\, E(Y_i(1))}{1-P(T_i|X_i)} + \sum_{i=1}^N \frac{(1-T_i)^2\, E(Y_i(0)))}{1-P(T_i|X_i)} \end{bmatrix}.
$$

Because T_i either takes a value of zero or one, we have the identities that $T_i^2 = T_i$ and $(1-T_i)^2 = 1-T_i$ and $(1-T_i)\,T_i = 0$ for each i. This means the previous

equation can simplify to

$$\begin{bmatrix} \sum_{i=1}^{N} \frac{T_i Y_i}{P(T_i|X_i)} \\ \sum_{i=1}^{N} \frac{(1-T_i) Y_i}{1-P(T_i|X_i)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \frac{T_i \, E(Y_i(1))}{P(T_i|X_i)} \\ \sum_{i=1}^{N} \frac{(1-T_i) \, E(Y_i(0)))}{1-P(T_i|X_i)} \end{bmatrix}$$
$$= \begin{bmatrix} E(Y_i(1)) \sum_{i=1}^{N} \frac{T_i}{P(T_i|X_i)} \\ E(Y_i(0)) \sum_{i=1}^{N} \frac{1-T_i}{1-P(T_i|X_i)} \end{bmatrix},$$

which leads to

$$\begin{bmatrix} \widehat{E(Y_i(1))} \\ \widehat{E(Y_i(0))} \end{bmatrix} = \begin{bmatrix} \left(\sum_{i=1}^{N} \frac{T_i Y_i}{P(T_i|X_i)}\right) \Big/ \left(\sum_{i=1}^{N} \frac{T_i}{P(T_i|X_i)}\right) \\ \left(\sum_{i=1}^{N} \frac{(1-T_i) Y_i}{1-P(T_i|X_i)}\right) \Big/ \left(\sum_{i=1}^{N} \frac{1-T_i}{1-P(T_i|X_i)}\right) \end{bmatrix},$$

which is the standard IPTW estimator. This shows that the inverse second-moment weighting estimator is analogous to IPTW when using the estimating equations in the binary treatment setting.

3.7 Appendix C

3.7.1 Standard errors for the regression prediction with propensity-spline method

We derive standard errors and tests by using a sandwich estimator. We start with an estimating function that is a stacked function with two subcomponents, $\boldsymbol{S}_i$ and $\boldsymbol{U}_i$,

$$\boldsymbol{\Psi}_i = \boldsymbol{\Psi}_i(\boldsymbol{\Gamma}^*, \boldsymbol{\xi}) = \begin{bmatrix} \boldsymbol{S}_i \\ \boldsymbol{U}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_i(\boldsymbol{\Gamma}^*) \\ \boldsymbol{U}_i(\boldsymbol{\Gamma}^*, \boldsymbol{\xi}) \end{bmatrix},$$

where

$$\begin{aligned}\boldsymbol{\Gamma} &= [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_k], \\ \boldsymbol{\nu} &= (\nu_1, \nu_2, \ldots, \nu_k)^\top,\end{aligned}$$

$$\begin{aligned}\boldsymbol{S}_i &= \boldsymbol{Z}_i(Y_i - \boldsymbol{B}_i^\top \boldsymbol{\Gamma}^{*\top} \boldsymbol{X}_i^*) \\ &= (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)(Y_i - (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)^\top \text{vec}(\boldsymbol{\Gamma}^*)), \\ \boldsymbol{U}_i &= E(\boldsymbol{\theta}_i | \boldsymbol{X}_i; \boldsymbol{\Gamma}^*) - \boldsymbol{\xi}, \\ (\boldsymbol{\Gamma}^*)^\top \boldsymbol{X}_i^* &= \boldsymbol{\nu} + \boldsymbol{\Gamma}^\top \boldsymbol{X}_i \\ &= \begin{bmatrix} \nu_1 + \boldsymbol{\gamma}_1^\top \boldsymbol{X}_i \\ \vdots \\ \nu_k + \boldsymbol{\gamma}_k^\top \boldsymbol{X}_i \end{bmatrix},\end{aligned}$$

$\boldsymbol{B}_i = \left(1, T_i, T_i^2, \ldots, T_i^k\right)^\top$, and length$(\boldsymbol{B}_i) = k$. Let length$(\boldsymbol{X}_i^*) = r$. To obtain the parameter estimates, solve $\sum_{i=1}^N \boldsymbol{\Psi}_i = 0$. This is a two-step process:

1. Solve $\sum_{i=1}^N \boldsymbol{S}_i(\boldsymbol{\Gamma}^*) = 0$ to obtain $\hat{\boldsymbol{\Gamma}}^*$.

2. Plug $\boldsymbol{\Gamma}^* = \hat{\boldsymbol{\Gamma}}^*$ into the set of equations $\sum_{i=1}^N \boldsymbol{U}_i(\boldsymbol{\xi}, \hat{\boldsymbol{\Gamma}}^*)$ to get $\hat{\boldsymbol{\xi}}$.

3.7.2 Background on sandwich estimator

Taylor series are used to derive the standard error of the parameters. A first-order approximation to the estimating equations is

$$\begin{aligned}\sum_{i=1}^N \boldsymbol{\Psi}_i(\hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\xi}}) &\approx \sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_0^*, \boldsymbol{\xi}_0) + \sum_{i=1}^N \frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_*^*, \boldsymbol{\xi}_*)}{\partial \boldsymbol{\phi}^\top} \left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \\ &= 0,\end{aligned}$$

where

$$\boldsymbol{\phi} = \begin{bmatrix} \text{vec}(\boldsymbol{\Gamma}^*) \\ \boldsymbol{\xi} \end{bmatrix}.$$

Rearranging some terms and using the law of large numbers gives

$$\begin{aligned}
\sqrt{N}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) &\approx -\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_*^*, \boldsymbol{\xi}_*)}{\partial \boldsymbol{\phi}^\top}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_0^*, \boldsymbol{\xi}_0) \\
&\approx \boldsymbol{A}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_0^*, \boldsymbol{\xi}_0) \\
&\approx \boldsymbol{A}^{-1}\mathcal{N}(\mathbf{0}, \boldsymbol{B}) \\
&\overset{D}{\approx} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{A} &= -E\left(\frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\Gamma}^*, \boldsymbol{\xi})}{\partial \boldsymbol{\phi}^\top}\right), \\
\boldsymbol{B} &= E\left(\boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_0^*, \boldsymbol{\xi}_0)(\boldsymbol{\Psi}_i(\boldsymbol{\Gamma}_0^*, \boldsymbol{\xi}_0))^\top\right).
\end{aligned}$$

To get the standard errors, we approximate $\boldsymbol{A}$ and $\boldsymbol{B}$ using the law of large numbers and plug in the estimated parameters in (3.10) and (3.11),

$$\widehat{\boldsymbol{A}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\Psi}_i(\hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\xi}})}{\partial \boldsymbol{\phi}^\top}\right), \tag{3.10}$$

$$\widehat{\boldsymbol{B}} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\Psi}_i(\hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\xi}})(\boldsymbol{\Psi}_i(\hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\xi}}))^\top\right). \tag{3.11}$$

To calculate $\widehat{\boldsymbol{A}}$, use

$$\left(\frac{\partial \boldsymbol{\Psi}_i(\hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\xi}})}{\partial \boldsymbol{\phi}^\top}\right) = \left[\begin{array}{c|c} \frac{\partial \boldsymbol{S}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{\xi}^\top} \\ \hline \frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\xi}^\top} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \frac{\partial \boldsymbol{S}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \mathbf{0} \\ \hline \frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\xi}^\top} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \frac{\partial (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)(Y_i - (\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*)^\top \mathrm{vec}(\boldsymbol{\Gamma}^*))}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \mathbf{0} \\ \hline \frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\xi}^\top} \end{array}\right]$$

$$= \left[\begin{array}{c|c} -\left(\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*\right)\left(\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*\right)^\top \mathbf{I}_{(k*r)\times(k*r)} & \mathbf{0} \\ \hline \frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \mathbf{I}_{k\times k} \end{array}\right]$$

$$= \left[\begin{array}{c|c} -\left(\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*\right)\left(\boldsymbol{B}_i \otimes \boldsymbol{X}_i^*\right)^\top & \mathbf{0} \\ \hline \frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} & \mathbf{I}_{k\times k} \end{array}\right],$$

where

$$\frac{\partial \boldsymbol{U}_i}{\partial (\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^*))^\top} = \begin{bmatrix} (\boldsymbol{X}_i^*)^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_i^*)^\top & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (\boldsymbol{X}_i^*)^\top \end{bmatrix}.$$

To calculate $\widehat{\boldsymbol{B}}$, plug in the parameter estimates. The sandwich estimator

becomes

$$\frac{1}{N}\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\left(\widehat{\boldsymbol{A}}^{-1}\right)^{\top} \approx V\begin{bmatrix}\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^{*})\\ \hat{\boldsymbol{\xi}}\end{bmatrix}$$
$$= \left[\begin{array}{c|c} V(\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^{*})) & \mathrm{Cov}(\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^{*}),\hat{\boldsymbol{\xi}}) \\ \hline \mathrm{Cov}(\hat{\boldsymbol{\xi}},\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^{*})) & V(\hat{\boldsymbol{\xi}}) \end{array}\right].$$

After the estimated covariance matrix is calculated, it is possible to test hypotheses about single parameters by using the Wald test. For example, to test if $\hat{\xi}_j = 0$, calculate the test statistic $\hat{\xi}_j/\mathrm{se}(\hat{\xi}_j)$ and compare it to a standard normal distribution. If the magnitude of the estimate is much bigger than the standard error, then it is unlikely that the null hypothesis is true.

We can also test hypotheses about multiple parameters simultaneously. For example, to test the null hypothesis

$$\boldsymbol{A}\boldsymbol{\xi} = \mathbf{c},$$

use

$$(\boldsymbol{A}\boldsymbol{\xi} - \mathbf{c})^{\top}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top})^{-1}(\boldsymbol{A}\boldsymbol{\xi} - \mathbf{c}) \sim \boldsymbol{\chi}^2_r,$$

where r is the number of degrees of freedom or the number of parameters being tested.

Chapter 4: Comparing propensity score methods with a continuous treatment: revisiting the lottery example

4.1 Introduction

Winning the lottery is a life-changing experience and can affect work habits, spending, relationships, and other aspects of life. Economists are interested in learning more about various aspects of this phenomenon, including how the amount of winnings affects future earned income. By studying lottery winners in this way, economists hope to understand how the amount of unearned income affects earned income not just for this lottery example, but for other types of unearned income such as gifts, inheritance, unemployment benefits, and other sources.

Imbens et al. (2001) obtained data from winners and players of the Massachusetts Megabucks lottery between 1984 and 1988 to study the long-term (5+ years) effects of winning the lottery. They concluded that unearned income reduces labor earning. The data included information about lottery winners such as their salaries before and after winning the lottery, their winning prize amount, education, savings, spending, and demographic variables including gender and age.

In this analysis, we focus only on people who won the lottery, and we consider how lottery prize amount affect their earned income six years after winning. We are interested in seeing whether the conclusions by Hirano and Imbens (2004) can be replicated using different methods. Hirano and Imbens (2004) found that those with lower earnings were much more sensitive to income changes than those with

higher earnings.

4.2 Goal of this analysis

The problem of estimating the average dose response function (ADRF) relies on novel causal inference methods because the amount of unearned income (lottery prize) is a continuous variable. Recently, there has been more interest in estimating causal effects when the treatment takes on continuous values (Flores et al., 2012; Kluve et al., 2012).

In this analysis, our goal is to estimate the ADRF of lottery winnings on earned salary. The ADRF is $\mu(t) = E[Y_i(t)]$, where $Y_i(t)$ is the salary obtained by winner i after winning prize amount t. The size of the lottery prize is strongly correlated with some background covariates. There is a high amount of unit and item nonresponse in the original survey, and it was shown that higher prize amounts are linked to lower probabilities of responding (Hirano and Imbens, 2004). Prize amount is related to some background covariates and the potential outcomes which means that adjustments need to be made in order to decrease confounding bias.

4.3 Data background

Our data set has 237 observations and is restricted to winners of the lottery. Table 4.1 gives summary statistics for the covariates. We removed observations with missing outcomes or in the upper 2% of the treatment value distribution (to avoid results driven by outliers) to retain 197 observations.

Table 4.1: Summary statistics and parameter estimates of generalized propensity score of lottery data set with 197 observations. Labor earnings are in thousands.

	Mean	S.D.	Corr w/Prize	t-stat ($\rho = 0$)	GPS Est.	GPS SE
Intercept					2.21	0.52
Age of winner	47.14	14.00	0.15	2.11	0.01	0.01
Years of high school	3.59	1.09	0.00	0.00	0.04	0.06
Years of college	1.36	1.58	0.05	0.75	0.03	0.04
Male	0.58	0.49	0.25	3.53	0.37	0.14
Tickets bought	4.60	3.28	0.00	0.06	-0.02	0.02
Working at time of win	0.79	0.41	0.04	0.51	-0.01	0.18
Year won after 1980	6.08	1.29	-0.07	-0.97	0.01	0.05
Earnings 1 yr. before win	14.67	13.81	0.16	2.33	0.00	0.01
Earnings 2 yrs. before win	13.45	13.03	0.19	2.75	-0.01	0.02
Earnings 3 yrs. before win	12.75	12.85	0.23	3.29	0.01	0.02
Earnings 4 yrs. before win	11.92	11.97	0.23	3.36	0.03	0.02
Earnings 5 yrs. before win	11.92	12.43	0.15	2.12	-0.02	0.02
Earnings 6 yrs. before win	11.70	12.41	0.14	1.95	-0.00	0.01

Table 4.2: Summary statistics with means and standard errors of complete cases with 197 observations. Prize and earnings are in thousands. The variable log($prize$) is the natural log of prize.

	mean	std. error
Earnings 6 yrs. after win	11.68	14.42
Prize amount	50.46	46.99
log(prize)	3.54	0.89

4.4 Model background

4.4.1 Potential outcomes and notation

The potential outcomes framework takes ideas from randomized experiments and applies them to observational data (see Chapter 2 for more details). In the continuous treatment setting, the goal is to understand the behavior of the same set of units under different treatment levels. Let the units in our sample be indexed by $i = 1, \ldots, N$. Let $Y_i(t)$ be the potential outcome for unit i for $t \in \mathcal{T}$ where $\mathcal{T}$ is the domain of possible treatments. Our goal is to estimate the ADRF which is defined as $\mu(t) = E[Y_i(t)]$. Let $Y_i = Y_i(T_i)$ denote the observed outcome, T_i denote the observed treatment, and $\boldsymbol{X}_i$ denote the covariates.

4.4.2 Assumptions

To make sure that estimates of causal effects are estimable, three standard assumptions are made: the stable unit treatment value assumption (SUTVA), no unmeasured confounding (which means that the observed covariates contain all information that is necessary to remove the confounding bias), and that each unit has a positive probability of receiving each treatment with probability $P(T_i = t|\boldsymbol{X}_i) > 0$ for every $\boldsymbol{x} \in \boldsymbol{X}_i$ in the population and all t with positive measure.

The weak unconfoundedness assumption means that adjusting for differences in pretreatment covariates will remove biases in units with different treatment values. It can be expressed as

$$Y_i(t) \perp\!\!\!\perp T_i|\boldsymbol{X}_i. \tag{4.1}$$

This assumption states that the conditional distribution of the potential outcomes are independent of the observed treatment given the covariates. The assumption is not directly testable, but including a rich set of covariates $\boldsymbol{X}_i$ makes it more

plausible.

4.4.3 Propensity score methodology

Two quantities used to adjust for the confounding bias in the continuous treatment setting are the generalized propensity score (GPS) (Hirano and Imbens, 2004) and the propensity function (PF) component (Imai and van Dyk, 2004). The GPS is a conditional density,

$$r(t, \boldsymbol{x}) = f_{T|\boldsymbol{X}}(t|\boldsymbol{X} = \boldsymbol{x}). \tag{4.2}$$

Let $R_i = r(T_i, \boldsymbol{X}_i)$, be the conditional density at the treatment actually received, and let $R_i^t = r(t, \boldsymbol{X}_i)$ denote the family of random variables indexed by t. For units with $T_i = t$, we have $R_i = R_i^t$. The PF component is a function of the covariates that uniquely parameterizes the GPS. For example, if $g(T)|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}^T\boldsymbol{\beta}, \sigma^2)$, then the PF component would be $\boldsymbol{X}^T\boldsymbol{\beta}$, if σ^2 is known. In some applications, the PF is estimated by the predicted treatment value conditioned on $\boldsymbol{X}$. On the other hand, methods that rely on the GPS include weighting methods which are analogous to using the propensity score in the binary treatment setting. Other examples that use the GPS are the partial mean approach (Hirano and Imbens, 2004; Flores et al., 2012; Bia et al., 2014).

The top row of Figure 4.1 shows the distribution of the lottery prize and the natural log-transformed version. The bottom row of Figure 4.1 shows the QQ-plot of residuals after regressing $\log(prize)$ on $\boldsymbol{X}$, and a plot of the residuals versus fitted values, respectively.

The treatment model is fit using a linear regression with log(treatment) regressed on main effects (no interactions) of all covariates. The QQ plot and residual plot show that a linear regression is an adequate fit.

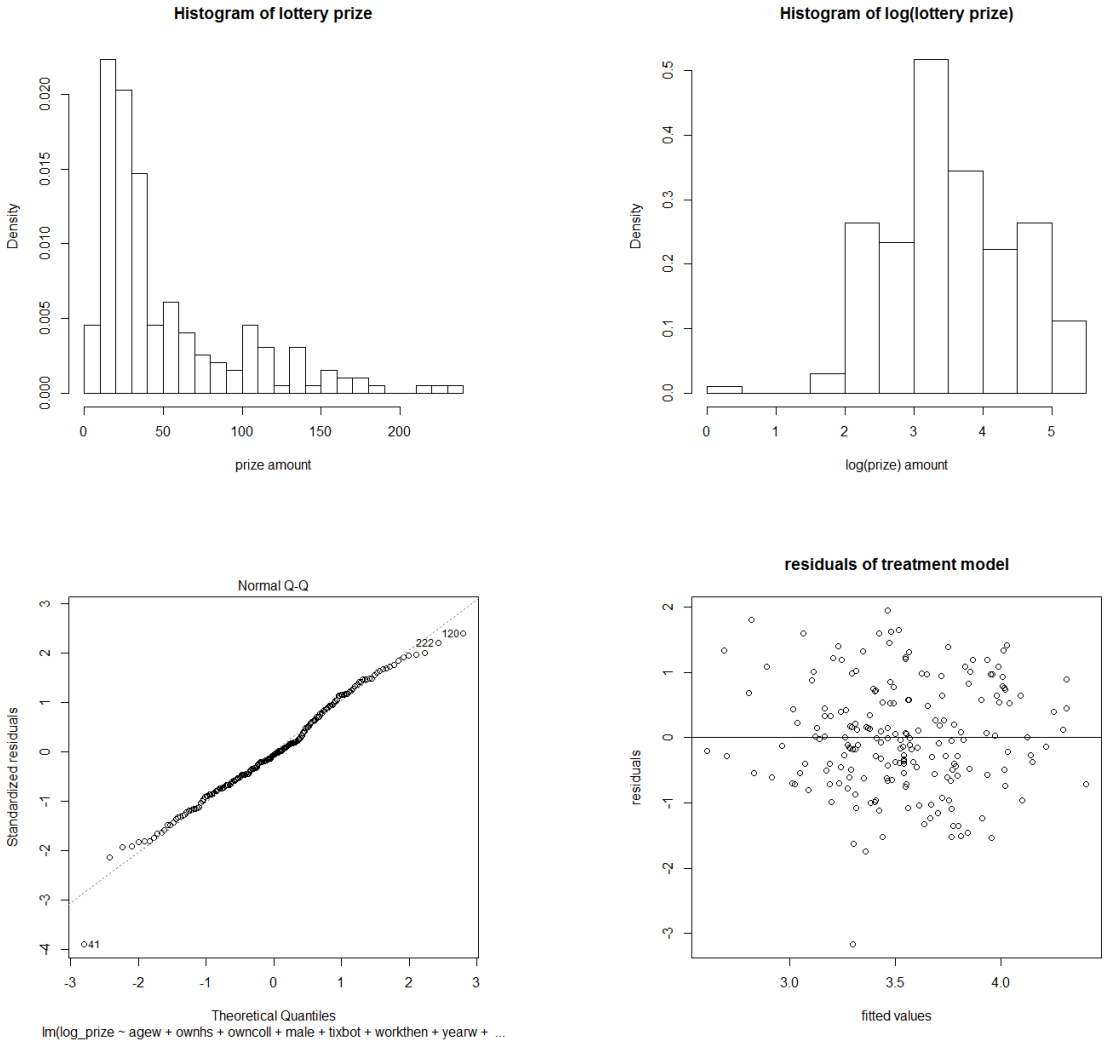

Figure 4.1: Histogram of lottery prizes, histogram of log(prize), QQ plot of residuals after regressing log(prize) on $\boldsymbol{X}$, and residuals of log(prize) versus fitted values for the lottery dataset with 197 observations.

4.4.4 Common support and covariate balance

The treatment model regresses the log of the prize amount on the covariates. We assess the overlap of the support for units with different treatment levels to minimize the amount of extrapolation and interpolation. A main goal of adjusting for the GPS is to remove confounding bias between different groups such as winners of small prizes versus big prizes. For example, in the binary treatment setting, big winners vs. small winners may have different distributions of their covariates. This lack of balance of covariates may bias the treatment-effect estimate. The GPS and propensity function (PF) are used to adjust the differences in covariates among different treatment levels. When adjusting for the GPS, the relevant balancing property is

$$[\boldsymbol{X}_i \perp\!\!\!\perp \mathbb{1}(T_i = t)] \,|r(t, \boldsymbol{X}_i), \tag{4.3}$$

which means that, conditioning on the GPS at a fixed treatment value, the distribution of the covariates is identical for those who received treatment level t and those who did not.

To enforce a common support in the continuous treatment setting, the treatment levels are first binned, and then GPS values are checked for each treatment level (Bia et al., 2014). For example, if the treatment is binned into K classes (based on treatment level), then for the first treatment class (the bin with the smallest treatments) we plug the median treatment value of the first bin into the treatment model to get predicted GPS values for each unit. The predicted GPS values of units in the first bin are compared with the predicted GPS values of units not in the first bin. Units that fall in the overlapping region are kept and the others are discarded. This process is repeated for each treatment class and the units that are not discarded are in the common support and are used for the analysis. More treatment classes tends to leave fewer units in the common support region.

We make the weak unconfoundedness assumption in Equation 4.1. The outcome is the earned income six years after winning the lottery and the pre-treatment variables are the following: age, gender, years in high school, years in college, winning year, number of tickets bought, working status at the time of winning, and earnings before winning the lottery for one year before through six years before.

From Bia et al. (2014), we use the formula

$$CS = \cap_{q=1}^{K}\{i : \hat{R}_i^q \in [\max\{\min_{\{j:Q_j=q\}} \hat{R}_j^q, \min_{\{j:Q_j\neq q\}} \hat{R}_j^q\}, \min\{\max_{\{j:Q_j=q\}} \hat{R}_j^q, \max_{\{j:Q_j\neq q\}} \hat{R}_j^q\}]\} \tag{4.4}$$

to get the common support. The sample is split into K intervals based on the distribution of the treatment variable, cutting equally at the quantiles of the observed distribution of treatments. The intervals are denoted by q_i and let Q_i denote the interval unit i belongs to: $T_i \in Q_i$. For each interval q_i, let $\hat{R}_i^k$ be the GPS evaluated at the median level of the treatment in that particular interval for unit i; $\hat{R}_i^k$ is calculated for all units. The common support region is calculated by comparing the support of the distribution of $\hat{R}_i^k$ for the units with $Q_i = q_k$ to the $\hat{R}_i^k$ values of units with $Q_i \neq q_k$. The common support subsample is CS and given in Equation 4.4. For three subclasses, the sample is reduced to 185 winners in the common support.

Figure 4.2 shows the GPS values before and after applying the common support restriction. We split the data into terciles based on treatment value and evaluate the units at the median value of the chosen tercile. For example, (a) compares the GPS values for units in the first tercile versus the units in the other terciles. The shaded bars represent units not in the chosen tercile while the white bars represent units in the chosen tercile. In histogram (a), we see that there are some units in the second and third tercile that have low GPS values that do not overlap with units in the first tercile. After restricting the sample to the common support region,

we see in (b) that the overlap of the GPS values is more in sync. The histograms on the left side of Figure 4.2 show the probabilities before applying the common support restriction and the ones on the right side show the histograms after the common support restriction. Histograms (c) and (d) compare units in the second tercile versus those in other terciles, and histgrams (e) and (f) compare units in the third terciles versus those in the first two terciles. The common support restriction removes non-overlapping units, which can be observed in Figure 4.2.

4.4.5 Checking covariate balance

To check the balancing property of the GPS there are four main methods. Flores et al. (2012) compare restricted and unrestricted models by using likelihood ratio tests. Imai and van Dyk (2004) compare regression coefficients before and after conditioning on the PF. Kluve et al. (2012) compare regression coefficients before and after conditioning on the GPS. Hirano and Imbens (2004) block on treatment and GPS values and compare mean differences for each covariate in blocked subgroups.

For the method of Flores et al. (2012), the unrestricted model fits treatment on the covariates and functions of the GPS values (GPS, GPS^2, GPS^3) and the restricted models sets either the covariates to zero or functions of the GPS values to zero. Comparing the unrestricted model with a restricted model that sets the covariates to zero shows that we cannot reject the null hypothesis that coefficients of the covariates in the treatment model are zero when conditioning on the GPS. This suggests that conditioning on the GPS adequately balances the covariates.

Imai and van Dyk (2004) compare coefficients with and without conditioning on the PF. Table 4.3 shows the coefficient estimates of each covariate x when regressing $\log(prize)$ on x (for units in the common support) with and without adjusting for $E[\log(prize)|\boldsymbol{X}_i]$. The predictors *age of winner*, *male*, *earnings 2 yrs. before win*, and *earnings 4 yrs. before win* are significant in the unconditional regressions,

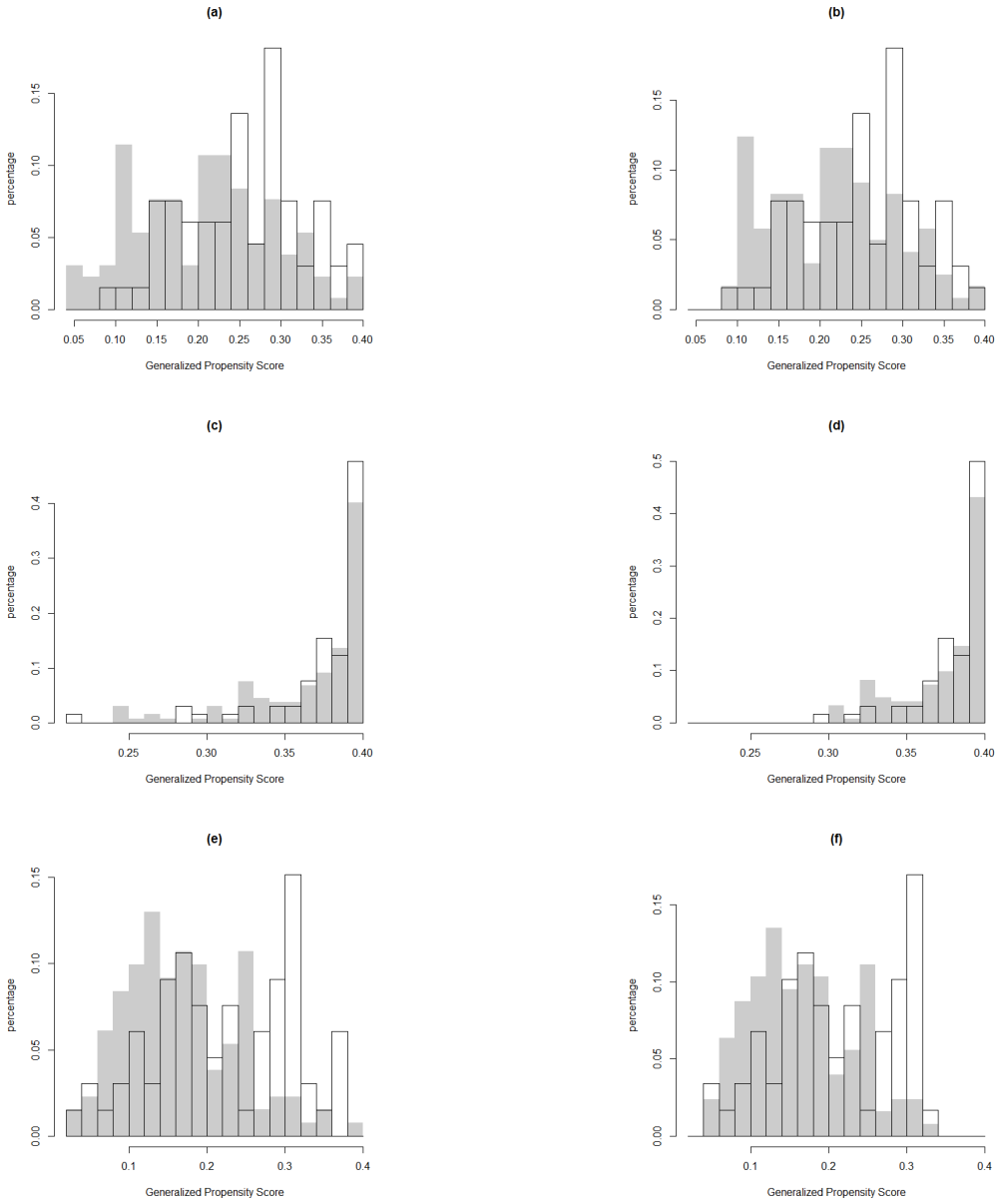

Figure 4.2: Common support restriction. Shaded bars represent units not in tercile, while white bars represent units in the tercile. (a) compares group 1 vs others before deleting non-overlapping units. (b) compares group 1 vs others after deleting non-overlapping units. (c) compares group 2 vs others before deleting non-overlapping units. (d) compares group 2 vs others after deleting non-overlapping units. (e) compares group 3 vs others before deleting non-overlapping units. (f) compares group 3 vs others after deleting non-overlapping units.

but when conditioning on $E[\log(prize)|\boldsymbol{X}_i]$, the coefficients for *age of winner* and *male* are not significant. This indicates that the covariates are more balanced given $E[\log(prize)|\boldsymbol{X}_i]$.

Kluve et al. (2012) compare coefficients with and without conditioning on the

GPS with the GPS evaluated at various potential values of the treatment level. In our case, we condition on $\widehat{gps}$ which is the estimated GPS at the observed treatment value and gps_{med} which is the GPS evaluated at the median treatment value. Table 4.4 shows the coefficient estimates of each covariate x when regressing *prize* on x (for units in the common support) with and without adjusting for $\widehat{gps}$ and gps_{med}. Variables *age of winner*, *male*, and earnings before the win are either significant or close to significant in the unconditional regressions, but when conditioning on $\widehat{gps}$ and gps_{med}, none of the coefficients are significant. Conditioning on $\widehat{gps}$ and gps_{med} makes the covariates more balanced.

4.5 Estimating the ADRF

When estimating the ADRF, we use polynomial-based methods from Chapter 3 and other methods described in Bia et al. (2014), Flores et al. (2012), Imai and van Dyk (2004), and Hirano and Imbens (2004). We calculate 95% pointwise standard errors using bootstrapping with 1000 samples. The bootstrapping takes into account the whole estimation process from estimating the GPS to estimating the ADRF.

For all methods, the problem of variable selection arises. For the treatment model, all covariates are used as main effects with no interactions. The treatment model produces residuals that appear normally distributed and diagnostics confirm an adequate fit.

4.5.1 Polynomial-based methods

The polynomial-based class of estimators consist of the prima facie, inverse probability of treatment weighting, inverse second moment weighting, regression, augmented inverse probability weighted estimating equations (aipwee), weighted regression, scalar weighted regression, propensity-spline prediction, and Imai-van-

Table 4.3: Covariate balance with and without adjustment (using the 185 observations within the common support). Unconditional effect of $\log(prize)$ compared to the effect of $\log(prize)$ conditional on $E[\log(prize)|\boldsymbol{X}_i]$.

	Unconditional effect of $\log(prize)$				*Effect of* $\log(prize)$ conditional on $E[\log(prize)\|\boldsymbol{X}_i]$			
	Est.	SE	t value	Pr(>|t|)	Est.	SE	t value	Pr(>|t|)
Age of winner	2.22	1.11	2.00	0.05	0.28	1.10	0.25	0.80
Years of high school	0.08	0.09	0.87	0.38	0.01	0.10	0.15	0.88
Years of college	0.03	0.13	0.25	0.80	-0.02	0.14	-0.14	0.89
Male	0.11	0.04	2.85	0.00	-0.00	0.03	-0.01	0.99
Tickets bought	-0.09	0.27	-0.33	0.74	-0.04	0.29	-0.13	0.90
Working then	0.02	0.03	0.47	0.64	0.01	0.04	0.14	0.89
Year won after 1980	-0.03	0.11	-0.30	0.76	-0.03	0.11	-0.24	0.81
Earnings 1 yr. before win	1.65	1.10	1.50	0.13	-0.21	1.09	-0.20	0.84
Earnings 2 yrs. before win	1.67	1.02	1.64	0.10	-0.32	0.99	-0.33	0.74
Earnings 3 yrs. before win	2.07	0.98	2.11	0.04	-0.20	0.92	-0.22	0.83
Earnings 4 yrs. before win	2.20	0.92	2.40	0.02	-0.13	0.83	-0.16	0.88
Earnings 5 yrs. before win	1.84	0.98	1.87	0.06	0.04	0.97	0.04	0.97
Earnings 6 yrs. before win	1.69	0.93	1.82	0.07	0.10	0.92	0.10	0.92

Table 4.4: Covariate balance with and without adjustment conditional on GPS (using the 185 observations within the common support). The estimates and Std. Errors are multiplied by 100.

	Unconditional effect of prize				*Effect of prize* conditional on $\widehat{gps}$ and gps_{med}			
	Est.	SE	t value	Pr(>|t|)	Est.	SE	t value	Pr(>|t|)
Age of winner	3.90	2.12	1.84	0.07	4.14	2.43	1.70	0.09
Years of high school	-0.02	0.18	-0.11	0.91	0.12	0.20	0.58	0.56
Years of college	0.04	0.25	0.14	0.89	-0.11	0.28	-0.41	0.69
Male	0.24	0.08	3.10	0.00	0.17	0.09	2.01	0.05
Tickets bought	0.09	0.52	0.17	0.86	-0.16	0.60	-0.27	0.79
Working then	0.03	0.06	0.46	0.64	-0.00	0.07	-0.01	0.99
Year won after 1980	-0.30	0.20	-1.49	0.14	-0.06	0.23	-0.27	0.79
Earnings 1 yr. before win	3.77	2.09	1.80	0.07	1.20	2.31	0.52	0.60
Earnings 2 yrs. before win	3.74	1.94	1.93	0.06	1.67	2.14	0.78	0.43
Earnings 3 yrs. before win	4.54	1.87	2.42	0.02	2.33	2.04	1.14	0.25
Earnings 4 yrs. before win	4.76	1.74	2.73	0.01	2.48	1.87	1.33	0.19
Earnings 5 yrs. before win	3.38	1.88	1.80	0.07	1.28	2.06	0.62	0.53
Earnings 6 yrs. before win	3.23	1.77	1.83	0.07	1.43	1.95	0.73	0.46

Dyk methods. See Chapter 3 for more details.

Table 4.5 summarizes the regressions of the outcome on differing functions of treatment values using polynomial functions of treatment with and without other covariates. The marginal effect shows a negative slope for prize. After controlling for covariates, the R^2 values are much higher.

Table 4.6 contains estimated slopes and standard errors when using linear polynomial based models. The estimated slopes are all negative. This can be interpreted as higher lottery winnings lead to less earned income in year 6.

Table 4.7 gives quadratic polynomial-based estimates of the slope, intercept, and quadratic coefficient. We also see Figures 4.3 and 4.4 show a negative trend as prize amount increases. This is in line with the first-degree estimators that suggest that earned income in year 6 decreases as lottery prize winnings increases.

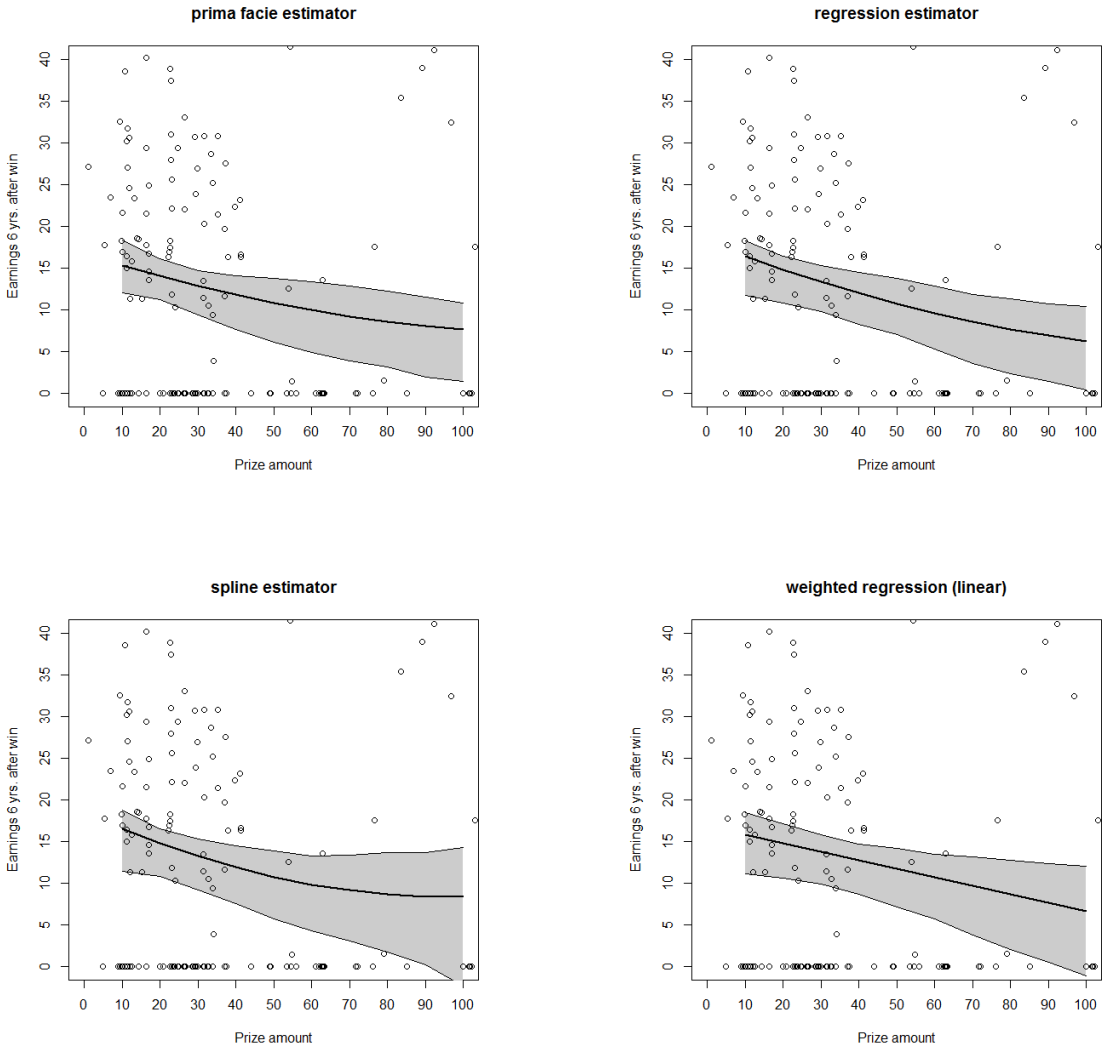

Figure 4.3: Estimated dose-response functions using 3 quadratic and 1 linear polynomial-based methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process.

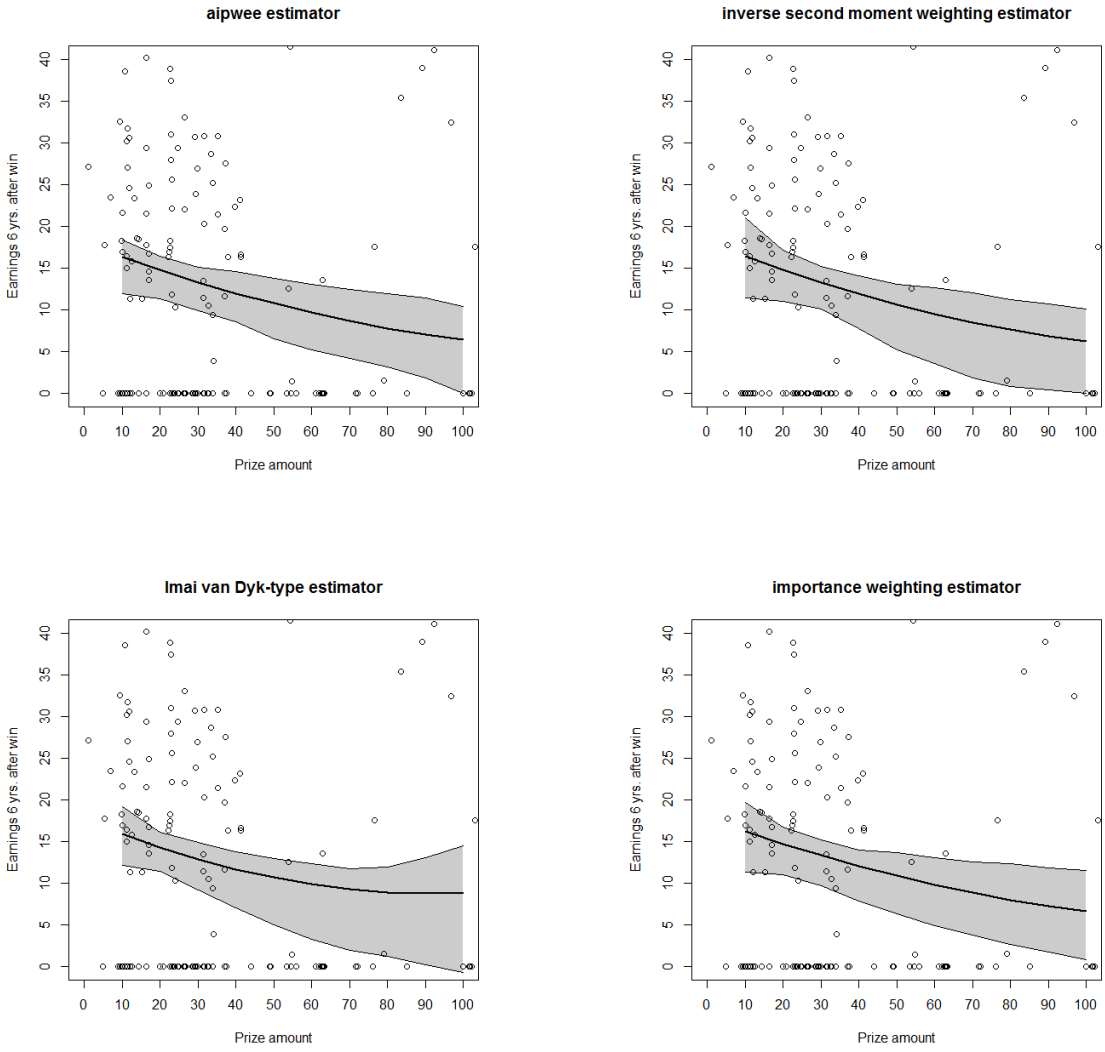

Figure 4.4: Estimated dose-response functions using 4 different quadratic polynomial-based methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process.

Table 4.5: Effect of prize on earned income in year 6 - estimates from linear regression model.

Variable	(1) marginal effect (SE)	(2) marginal effect (SE)	(3) marginal effect (SE)	(4) marginal effect (SE)
(a) Only control for prize				
$Prize$	-0.0491	-0.146	-0.132	-0.453
	(0.0226)	(0.0719)	(0.166)	(0.32)
$Prize^2/1000$		0.548	0.367	7.662
		(0.3863)	(1.925)	(6.51)
$Prize^3/1000^3$			0.578	-54.797
			(6.008)	(47.604)
$Prize^4/1000^4$				129.740
				(110.63)
R^2	0.0251	0.0358	0.0359	0.0432
Adjusted R^2	0.0198	0.0252	0.0199	0.0219
Number of obs.	185	185	185	185
(b) Control for prize and other variables				
$Prize$	-0.0757	-0.176	-0.0918	-0.524
	(0.0182)	(0.0561)	(0.128)	(0.244)
$Prize^2/1000$		0.573	-0.4962	9.403
		(0.3028)	(1.487)	(5.011)
$Prize^3/1000^3$			3.4060	-72.223
			(4.637)	(36.880)
$Prize^4/1000^4$				177.914
				(86.083)
R^2	0.501	0.511	0.513	0.525
Adjusted R^2	0.459	0.468	0.466	0.476
Number of obs.	185	185	185	185

Table 4.6: Estimated slopes from linear polynomial-based model. All numbers are multiplied by 10.

	mean	se
prima facie	-0.70	0.24
importance sampling	-0.77	0.24
inverse second moment weighting	-0.88	0.29
regression	-0.84	0.26
aipwee	-0.87	0.27
weighted reg.	-0.97	0.38
scalar weighted reg.	-0.81	0.25
propensity spline	-0.85	0.36
Imai van Dyk	-0.81	0.37

Table 4.7: Estimated coefficients from the quadratic polynomial-based models with means and standard errors.

	Intercept (SE)	slope (SE)	(quad coef) $\times$ 1000 (SE)
prima facie			
	16.68	-0.16	0.53
	2.19	0.08	0.44
importance sampling			
	17.03	-0.17	0.56
	2.51	0.09	0.48
inverse second moment weighting			
	18.34	-0.21	0.82
	3.47	0.13	0.78
regression			
	16.99	-0.15	0.41
	2.03	0.07	0.39
aipwee			
	17.06	-0.16	0.43
	1.95	0.07	0.39
scalar weighted reg.			
	16.83	-0.15	0.39
	2.18	0.08	0.43
propensity spline			
	17.39	-0.18	0.79
	2.46	0.10	0.91
Imai van Dyk			
	18.26	-0.24	1.27
	2.40	0.11	1.03

4.5.2 Other methods

The other estimators of the ADRF consist of the additive spline-based semi-parametric, BART, generalized additive model, Hirano-Imbens, inverse weighting kernel, and Nadaraya-Watson methods.

The Nadaraya-Watson estimate is given by

$$\hat{\mu}(t)_{NW} = \frac{\sum_{i=1}^{N} \tilde{K}_{h,\boldsymbol{X}}(T_i - t)Y_i}{\sum_{i=1}^{N} \tilde{K}_{h,\boldsymbol{X}}(T_i - t)}, \tag{4.5}$$

where $\sum_{i=1}^{N} \tilde{K}_{h,\boldsymbol{X}} = K_h(T_i - t)/\hat{R}_i^t$. This is a locally constant regression, but with each unit's kernel weight divided by the GPS at treatment level t. The kernel $K_h(t)$ is chosen to be Gaussian with the bandwidth selected by using the Sheather-Jones bandwidth selection method (Sheather and Jones, 1991).

The inverse weighting kernel estimate is

$$\hat{\mu}(t)_{IW} = \frac{D_0(t)S_2(t) - D_1(t)S_1(t)}{S_0(t)S_2(t) - S_1^2(t)}, \tag{4.6}$$

where $S_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,\boldsymbol{X}}(T_i - t)(T_i - t)^j$ and $D_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,\boldsymbol{X}}(T_i - t)(T_i - t)^j Y_i$. For more details, see Flores et al. (2012).

The Hirano-Imbens method uses the following steps (Graham et al., 2014):

1. Estimate a model for the treatment model $T_i|\boldsymbol{X}_i$.

2. Use the estimated parameters from from Step 1, with the assumed density function to calculate $\hat{R}_i = f_{T_i|bX_i}(T_i|\boldsymbol{X}_i = \boldsymbol{x}_i)$ and $\hat{R}_i = f_{T_i|bX_i}(t|\boldsymbol{X}_i = \boldsymbol{x}_i)$ for all t of interest.

3. Use $\hat{R}_i$ to find a region of common support and check covariate balance.

4. Estimate the parameters from the conditional outcome given the treatment

and the estimated GPS values by

$$E[Y_i|T_i, \hat{R}_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 \hat{R}_i + \alpha_4 \hat{R}_i^2 + \alpha_5 T_i \hat{R}_i. \tag{4.7}$$

5. Estimate the ADRF over grid points t by averaging over the distribution of $\hat{R}_i^t$ to obtain

$$\hat{\mu}_{HI}(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{R}_i^t + \hat{\alpha}_4 \hat{R}_i^{t2} + \hat{\alpha}_5 t \hat{R}_i^t \right]. \tag{4.8}$$

6. Use a bootstrap resampling scheme for the previous steps to estimate the variance.

We restrict our fit to quadratic terms, but higher order terms can also be included.

The generalized additive model method is analogous to the Hirano-Imbens method, but fits a generalized additive model instead of the flexible polynomial regression given in Equation 4.7. The generalized additive model fits the outcome on the treatment and GPS using $Y_i \sim s(T_i) + s(gps_i)$ and then averages over the grid points t analogous to Equation 4.8.

The additive spline-based semi-parametric method is also analogous to the Hirano-Imbens method, but uses additive splines to fit the outcome model in Equation 4.7 and average the estimates over all the units analogous to Equation 4.8.

The method using Bayesian additive regression trees (BART) is described by Hill (2011). This method fits a response surface to the data nonparametrically and makes few assumptions. It does not model the treatment, but only focuses on optimal prediction of the outcome surface. To predict the ADRF, first fit the response surface, then predict the potential outcome of each unit at a specified treatment value (using the fitted response surface), then find the average outcome at the specified treatment value by averaging over all the units, and finally repeat

this process over many grid values. In our example, we fit 10 treatment values for $t \in \{10, 20, \ldots, 100\}$.

These methods also show a similar downward trend as those in the polynomial methods. Table 4.8 gives parameter estimates and standard errors for the HI method. The different methods in Figure 4.5 show that higher lottery winnings leads to lower earned income in year six.

Table 4.8: Parameter estimates of conditional distribution of prize given covariates for overlapping data with 185 observations.

	Estimate	Std. Error
$(Intercept)$	11.07	5.62
$prize$	-0.08	0.08
$prize^2/1000$	1.05	0.52
$\log(gps)$	-5.09	5.00
$\log(gps)^2$	-0.36	0.72
$\log(gps) * prize$	0.08	0.04

4.6 Results

Results are shown in Figures 4.3, 4.4, 4.5, which display the estimated ADRFs and their bootstrap standard errors. The figures display a general trend of decreasing future income as prize amount increases from $\$10,000$ to $\$100,000$. The generalized additive model method in Figure 4.5 shows some oscillations in its standard error of the future income as the prize amount increases, but the overall decreasing trend is preserved. In Table 4.6, the slopes are significant and negative. In Table 4.7, almost all the methods show the slope is significant, while the quadratic coefficient is not significant.

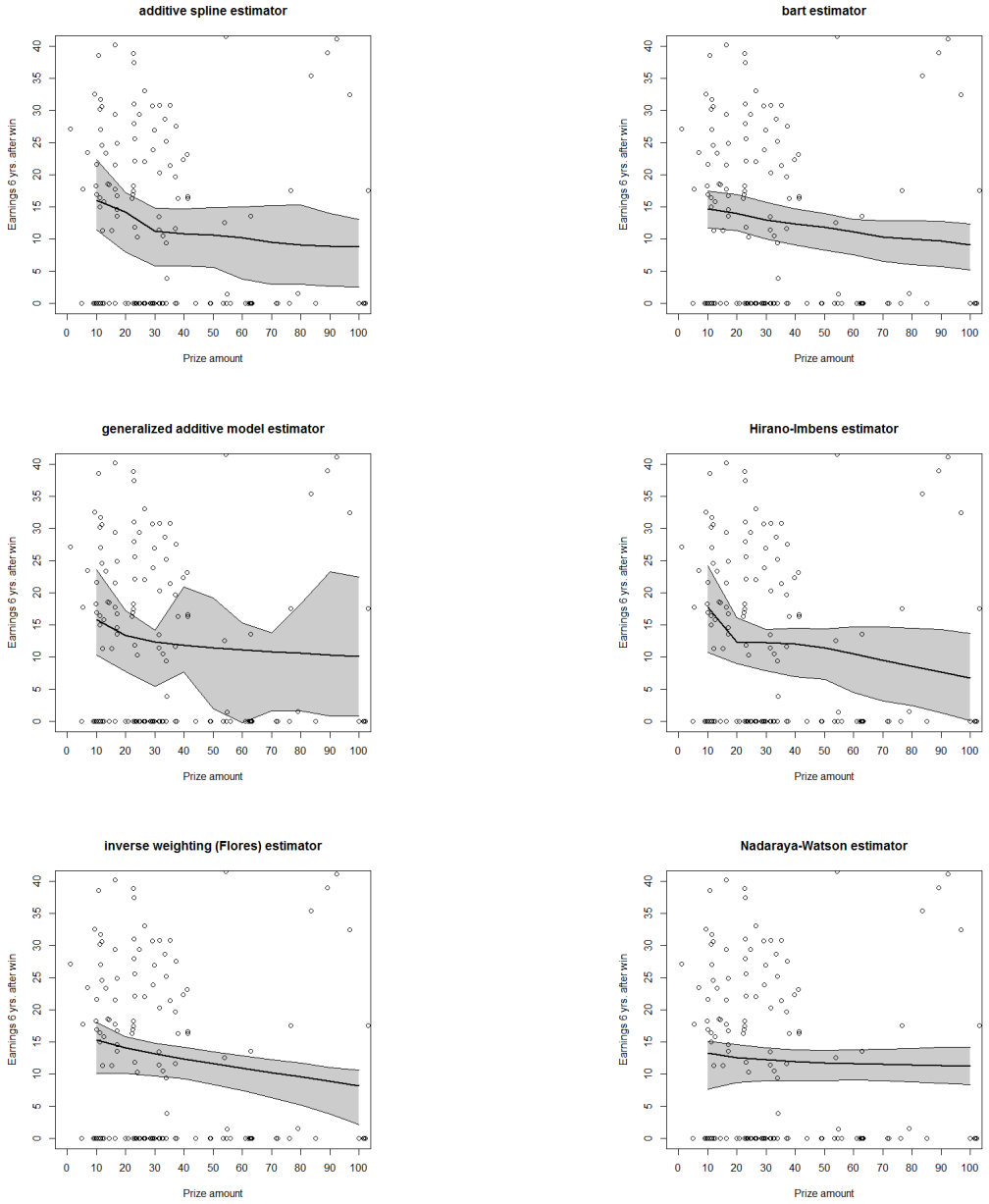

Figure 4.5: Estimated dose-response functions using six different methods with 95% pointwise standard errors.

4.7 Discussion

In this study, we have demonstrated that different statistical methods provide evidence that lottery winners decrease their earned income six years after winning the prize as the prize amount increases, similar to the trend described in Hirano and Imbens (2004) and Bia et al. (2014).

Estimating the ADRF using causal inference methods is an underdeveloped topic and one that needs more attention. Imai and van Dyk (2004) and Hirano and Imbens (2004) have introduced techniques to estimate the ADRF, and newer modifications such as Flores et al. (2012), Zhao et al. (2014), and we introduced other estimators in Chapter 3.

In the next chapter, we provide a new R package `causaldrf` that implements the different statistical methods for the continuous treatment setting. Bia et al. (2014) provide software in Stata for estimating average dose response functions, but there is not an R statistical language package that provides these methods. The R package `causaldrf` contains functions and data for estimating causal dose response functions to illustrate methods discussed in Chapter 3 and by Bia et al. (2014).

Chapter 5: Estimating average dose response functions using the R package `causaldrf`

5.1 Introduction

In this chapter, we provide examples to illustrate the flexibility and the ease of use of the `causaldrf` R package, which estimates the average dose response function (ADRF) when the treatment is continuous. The `causaldrf` R package also provides methods for estimating average potential outcomes when the treatment is binary or multi-valued. The user can compare different methods to understand the sensitivity of the estimates and a way to check robustness. The package contains new estimators based on a linear combination of a finite number of basis functions (Chapter 3). In addition, `causaldrf` includes functions useful for model diagnostics such as assessing common support and for checking covariate balance. This package fills a gap in the R package space and offers a range of existing and new estimators described in the statistics literature by Bia et al. (2014), Flores et al. (2012), Imai and van Dyk (2004), Hirano and Imbens (2004), and Robins et al. (2000).

The `causaldrf` R package is currently available on the Comprehensive R Archive Network (CRAN). The R package contains twelve functions for estimating the ADRF which are explained in more detail in Chapters 2, 3, and in the documentation files for the package (`https://cran.r-project.org/web/packages/causaldrf/index.html`). Users can choose which estimator to apply based on their particular problems and goals.

This chapter is organized as follows. In Section 5.2, we introduce a simulated dataset from Hirano and Imbens (2004) and Moodie and Stephens (2012) and apply functions from `causaldrf` to estimate the ADRF. In Section 5.3, we use data from the National Medical Expenditures Survey (NMES) to show the capabilities of `causaldrf` in analyzing a data set containing weights. Section 5.4 contains data from the Infant Health and Development Program (IHDP) and applies methods from `causaldrf` to the data. Conclusions are presented in Section 5.5.

5.2 An example based on simulated data

This section demonstrates the use of the `causaldrf` package by using simulated data from Hirano and Imbens (2004) and Moodie and Stephens (2012). This simulation constructs an ADRF with an easy to interpret functional form, and a means to clearly compare the performance of different estimation methods.

Let $Y_1(t)|X_1, X_2 \sim \mathcal{N}\left(t + (X_1 + X_2)e^{-t(X_1+X_2)}, 1\right)$ and let X_1, X_2 be unit exponentials, $T_1 \sim \exp(X_1 + X_2)$. The ADRF can be calculated by integrating out the covariates analytically (Moodie and Stephens, 2012),

$$\mu(t) = E(Y_i(t)) = t + \frac{2}{(1+t)^3}. \tag{5.1}$$

This example provides a setting to compare ADRF estimates with the true ADRF given in Equation 5.1. In this simulation, our goal is to demonstrate how to use the functions. We introduce a few of the estimators and show their plots.

To use the functions, first, install **causaldrf** and then load the package:

```
library (causaldrf)
```

The data are generated from:

```r
set.seed(301)
hi_sample <- function(N){
  X1 <- rexp(N)
  X2 <- rexp(N)
  T <- rexp(N, X1 + X2)
  gps <- (X1 + X2) * exp(-(X1 + X2) * T)
  Y <- T + gps + rnorm(N)
  hi_data <- data.frame(cbind(X1, X2, T, gps, Y))
  return(hi_data)
}


hi_sim_data <- hi_sample(1000)
head(hi_sim_data)
```

```
##          X1         X2           T        gps          Y
## 1 0.1942127 0.18045487 4.718463128 0.06395528  4.1426651
## 2 1.4441432 0.60652576 0.168123100 1.45266708  0.9888306
## 3 5.6393370 0.17758343 0.005784747 5.62444109  5.2284042
## 4 0.5079408 0.45976378 0.350261484 0.68950725 -0.3301777
## 5 0.2282938 0.71565806 0.431730712 0.62800127  1.8360819
## 6 1.1539278 0.09854209 0.786804283 0.46751158  1.4745739
```

Below is code for a few different estimators of the ADRF. The first is the additive spline estimator from Bia et al. (2014). This estimator fits a treatment model to estimate the GPS. Next, additive spline bases values are created for both the treatment and the GPS. The outcome is regressed on the treatment, GPS, treatment bases, and GPS bases. After the outcome model is estimated, each treatment grid value and set of covariates is plugged into the model which corresponds to imputed values for each unit at that particular treatment value. The imputed values are averaged to get the estimated ADRF at that treatment value. Repeating this process for many treatment values, `grid_val`, traces out the estimated ADRF.

The arguments are: `Y` for the name of the outcome variable, `treat` for the name of the treatment variable, `treat_formula` for the formula used to fit the treatment model, `data` for the name of the data set, `grid_val` for a vector in the domain of the treatment for where the outcome is estimated, `knot_num` for the number of knots for

the spline fit, and `treat_mod` for the treatment model that relates treatment with the covariates.

In this example we fit the correct treatment model so that the GPS is correctly specified with a gamma distribution.

```
add_spl_estimate <- add_spl_est(Y = Y,
                                treat = T,
                                treat_formula = T ~ X1 + X2,
                                data = hi_sim_data,
                                grid_val = quantile(hi_sim_data$T,
                                    probs = seq(0, .95, by = 0.01)),
                                knot_num = 3,
                                treat_mod = "Gamma",
                                link_function = "inverse")
```

The next estimator is based on the generalized additive model. This method requires a treatment formula and model to estimate the GPS. The estimated GPS values are used to fit an outcome regression. The outcome, `Y`, is regressed on two quantities: the treatment, `T`, and spline basis terms from the GPS fit.

```
gam_estimate <- gam_est(Y = Y,
                        treat = T,
                        treat_formula = T ~ X1 + X2,
                        data = hi_sim_data,
                        grid_val = quantile(hi_sim_data$T,
                                    probs = seq(0, .95, by = 0.01)),
                        treat_mod = "Gamma",
                        link_function = "inverse")
```

The Hirano-Imbens estimator also requires two models. The first model regresses the treatment, `T`, on a set of covariates to estimate the GPS values. The second step requires fitting the outcome, `Y`, on the observed treatment and fitted GPS values. The summary above shows the fit of both the treatment model and outcome model. Also shown is the estimated outcome values on the grid of treatment values, `quantile_grid`.

```
hi_estimate <- hi_est(Y = Y,
                      treat = T,
                      treat_formula = T ~ X1 + X2,
                      outcome_formula = Y ~ T + I(T^2) +
                        gps + I(gps^2) + T * gps,
                      data = hi_sim_data,
                      grid_val = quantile(hi_sim_data$T,
                                          probs = seq(0, .95, by = 0.01)),
                      treat_mod = "Gamma",
                      link_function = "inverse")
```

This last method, importance sampling, fits the treatment as a function of the covariates, then calculates GPS values. The GPS values are then used as inverse-probability weights in the regression of `Y` on `T` (Robins et al., 2000). The estimated parameters correspond to coefficients for a quadratic model of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$. In this example, the estimator is restricted to a quadratic fit.

```
iptw_estimate <- iptw_est(Y = Y,
                          treat = T,
                          treat_formula = T ~ X1 + X2,
                          numerator_formula = T ~ 1,
                          data = hi_sim_data,
                          degree = 2,
                          treat_mod = "Gamma",
                          link_function = "inverse")
```

The true ADRF and 4 estimates are plotted in Figure 5.1.

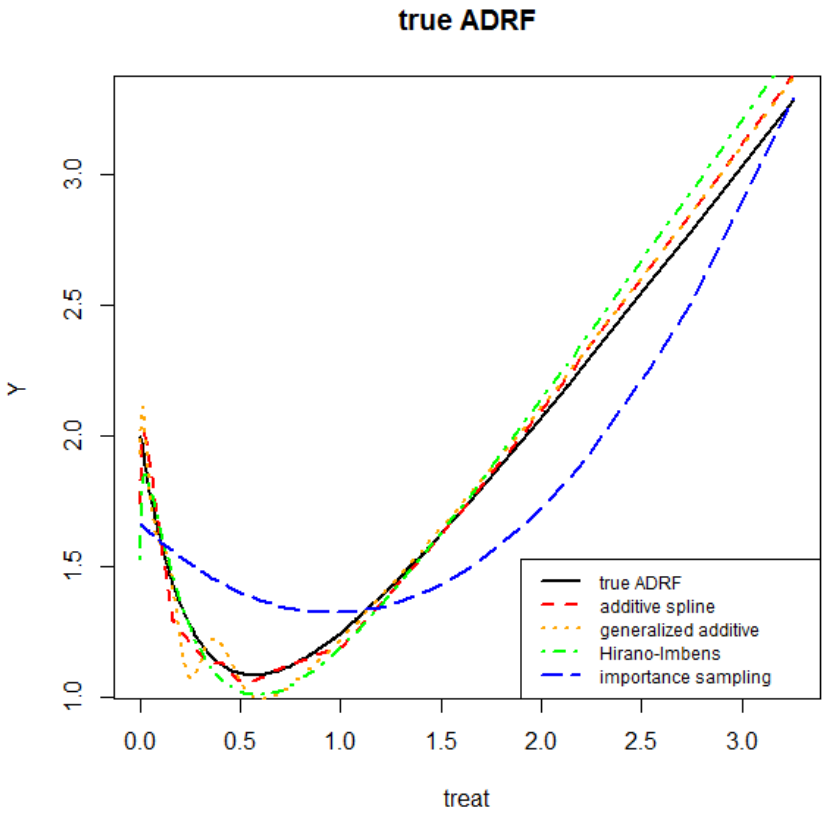

Figure 5.1: True ADRF along with estimated curves.

5.3 Analysis of the National Medical Expenditures Survey

5.3.1 Introduction

The 1987 National Medical Expenditures Survey (NMES) provides information about health status, behaviors, and medical expenditures for a representative sample of the U.S. civilian, non-institutionalized population (U.S. Department of Health and Human Services, Public Health service, 1987). The 1987 medical costs were verified by multiple interviews and other data from clinicians and hospitals.

Johnson et al. (2003) analyzed the NMES to estimate the fraction of disease cases and the fraction of the total medical expenditures attributable to smoking for two disease groups. Imai and van Dyk (2004) emulate the setting of Johnson et al. (2003) but estimated the effect of smoking amount on medical expenditures. Johnson et al. (2003) and Imai and van Dyk (2004) conducted a complete-case analysis by removing units containing missing values. Johnson et al. (2003) used multiple imputation techniques to deal with the missing values, but did not find significant differences between that analysis and the complete case analysis. Complete case analysis with propensity scores will lead to biased causal inference unless the data are missing completely at random (D'Agostino Jr and Rubin, 2000). Regardless of this drawback, the analysis in this section uses the complete-case data to illustrate the different statistical methods available for estimating the ADRF relating smoking amount and medical expenditures.

This example is analyzed in this section because the treatment variable, smoking amount, is a continuous variable. The data is restricted to that used in Imai and van Dyk (2004) with 9708 observations and 12 variables. For each person interviewed, the survey collected information on age at the time of the survey, age when the person started smoking, gender, race (white, black, other), marital status (married, widowed, divorced, separated, never married), education level (college

graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, or West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always) (Imai and van Dyk, 2004). The data are available in the `causaldrf` package.

Our goal is to understand how the amount of smoking affects the amount of medical expenditures. Johnson et al. (2003) use a measure of cumulative exposure to smoking that combines self-reported information about frequency and duration of smoking into a variable called *packyear*, defined as

$$packyear = \frac{\text{number of cigarettes per day}}{20} \times (\text{number of years smoked}). \quad (5.2)$$

One can also define *packyear* as the number of packs smoked per day multiplied by the number of years the person was a smoker. The total number of cigarettes per pack is normally 20.

The NMES oversampled subgroups of the population in order to reduce variances of the estimates. Oversampling reduces the variances of the estimates by increasing the sample size of the target sub-population disproportionately (Singh et al., 1994). The U.S. Department of Health and Human Services oversampled Blacks, Hispanics, the poor and near poor, and the elderly and persons with functional limitations (Cohen, 2000).

5.3.2 Data description

Load **nmes_data** into the workspace with

```
data("nmes_data")
dim (nmes_data)

## [1] 9708   12

summary(nmes_data)
```

```
##    packyears         AGESMOKE        LASTAGE          MALE
##  Min.   :  0.05   Min.   : 9.00   Min.   :19.0   Min.   :0.0000
##  1st Qu.:  6.60   1st Qu.:16.00   1st Qu.:32.0   1st Qu.:0.0000
##  Median : 17.25   Median :18.00   Median :45.0   Median :1.0000
##  Mean   : 24.48   Mean   :18.39   Mean   :47.1   Mean   :0.5159
##  3rd Qu.: 34.50   3rd Qu.:20.00   3rd Qu.:62.0   3rd Qu.:1.0000
##  Max.   :216.00   Max.   :70.00   Max.   :94.0   Max.   :1.0000
##  RACE3    beltuse  educate  marital  SREGION  POVSTALB
##  1: 633   1:2613   1:2047   1:6188   1:2047   1:1034
##  2:1496   2:2175   2:2451   2: 771   2:2451   2: 470
##  3:7579   3:4920   3:3386   3:1076   3:3386   3:1443
##                    4:1824   4: 333   4:1824   4:3273
##                             5:1340            5:3488
##
##     HSQACCWT        TOTALEXP
##  Min.   :  908   Min.   :     0.0
##  1st Qu.: 4975   1st Qu.:    90.0
##  Median : 7075   Median :   406.1
##  Mean   : 8072   Mean   :  2042.0
##  3rd Qu.:10980   3rd Qu.:  1350.3
##  Max.   :35172   Max.   :175096.0
```

The dataset **nmes_data** is a data frame with 9708 rows and 12 variables with summaries of the variables given above. Six of the variables are numeric and the other six are categorical. The outcome variable is the total amount of medical expenditures, `TOTALEXP` and the treatment is the amount of smoking, `packyears`. The data set contains weights, `HSQACCWT`, that can be used to upweight estimates to the population of interest which is the set of people who smoke and are above the age of 18. This analysis demonstrates the capability of `causaldrf` by estimating the ADRF with and without weights. In Figure 5.3, we plot the estimated ADRFs, their 95% confidence bands, and the 95% confidence bands without weigths.

5.3.3 Common support

The data set is restricted to observations that overlap and have a common support. Units outside of the common support are removed. See Figure 5.2. The

preliminary steps of analysis are omitted such as cleaning and making sure the data overlap.

From Bia et al. (2014), we use the formula

$$CS = \cap_{q=1}^{K}\{i : \hat{R}_i^q \in [\max\{\min_{\{j:Q_j=q\}} \hat{R}_j^q, \min_{\{j:Q_j\neq q\}} \hat{R}_j^q\}, \min\{\max_{\{j:Q_j=q\}} \hat{R}_j^q, \max_{\{j:Q_j\neq q\}} \hat{R}_j^q\}]\} \tag{5.3}$$

to get the common support. For 3 subclasses, the sample is reduced to 8732 units in the common support.

5.3.4 Checking covariate balance

One of the main goals of fitting a treatment model is to balance the covariates. The GPS or the PF provides a way to balance the covariates. Comparisons of the balance of the covariates before and after adjusting for the GPS or the PF are shown in the following results:

```
t(p_val_bal_cond)

##             Estimate  Std. Error    t value  Pr(>|t|)
## AGESMOKE 0.002649140 0.052968507 0.05001349 0.9601128
## LASTAGE  0.156568519 0.119806490 1.30684505 0.1912998
## MALE     0.006755654 0.005139165 1.31454310 0.1886980

t(p_val_bal_no_cond)

##             Estimate  Std. Error   t value      Pr(>|t|)
## AGESMOKE -0.64598664 0.047692883 -13.54472  2.225653e-41
## LASTAGE   5.11758526 0.140218443  36.49723 1.483453e-271
## MALE      0.05582729 0.004566978  12.22412  4.385267e-34
```

The last column displays the p-value of regressing each of the continuous covariates on the outcome variable, packyears, before and after conditioning on the PF. The first three rows show the p-values after conditioning on the PF, while the

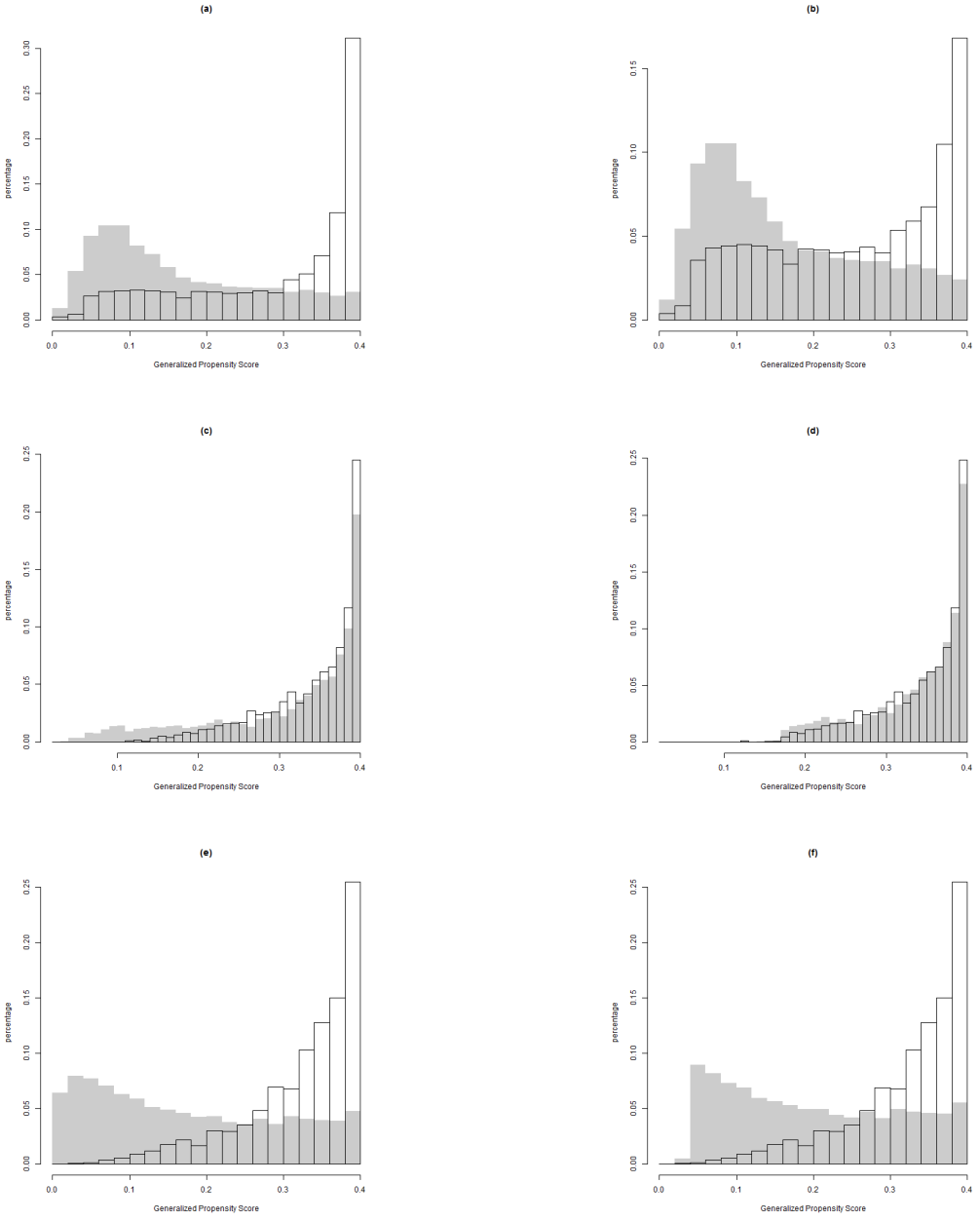

Figure 5.2: Common support restriction. Shaded bars represent units not in tercile, while white bars represent units in the tercile. (a) Compares group 1 vs others before deleting non-overlapping units. (b) Compares group 1 vs others after deleting non-overlapping units. (c) Compares group 2 vs others before deleting non-overlapping units. (d) Compares group 2 vs others after deleting non-overlapping units. (e) Compares group 3 vs others before deleting non-overlapping units. (f) Compares group 3 vs others after deleting non-overlapping units.

last three rows show the p-values when there is no conditioning.

5.3.5 Estimating the ADRF

The `causaldrf` R package contains a variety of estimators. Below is code for four other estimators that can account for weights. Although the true ADRF is not a polynomial, we will illustrate methods that are restricted to polynomial form of up to degree 2.

The prima facie estimator is a basic estimator that regresses the outcome `Y` on the treatment `T` without taking covariates into account. The prima facie estimator is unbiased if the data comes from a simple random sample; otherwise it will likely be biased. The model fit is $Y \sim \alpha_0 + \alpha_1 t + \alpha_2 t^2$.

```
pf_estimate <- reg_est(Y = TOTALEXP,
                       treat = packyears,
                       covar_formula = ~ 1,
                       data = full_data_orig,
                       degree = 2,
                       wt = full_data_orig$HSQACCWT,
                       method = "same")
pf_estimate

##
## Estimated values:
## [1] 1128.5947250   36.8409486   -0.1348346
```

The regression prediction method generalizes the prima facie estimator and takes the covariates into account (Galagate and Schafer, 2015a).

```
reg_estimate <- reg_est(Y = TOTALEXP,
                        treat = packyears,
                        covar_formula = ~ LASTAGE + LASTAGE2 +
                          AGESMOKE + AGESMOKE2 + MALE + beltuse +
                          educate + marital + POVSTALB + RACE3,
                        covar_lin_formula = ~ 1,
                        covar_sq_formula = ~ 1,
```

```
                          data = full_data_orig,
                          degree = 2,
                          wt = full_data_orig$HSQACCWT,
                          method = "different")
reg_estimate

##
## Estimated values:
## [1] 1619.329529   23.260395   -0.109507
```

The propensity spline prediction method adds spline basis terms to the regression prediction method. This method is similar to that of Little and An (2004) and Schafer and Kang (2008), but for the continuous treatment setting (Galagate and Schafer, 2015a).

```
spline_estimate <- prop_spline_est(Y = TOTALEXP,
                           treat = packyears,
                           covar_formula = ~ LASTAGE + LASTAGE2 +
                             AGESMOKE + AGESMOKE2 + MALE + beltuse +
                             educate + marital + POVSTALB + RACE3,
                           covar_lin_formula = ~ 1,
                           covar_sq_formula = ~ 1,
                           data = full_data_orig,
                           e_treat_1 = full_data_orig$est_treat,
                           degree = 2,
                           wt = full_data_orig$HSQACCWT,
                           method = "different",
                           spline_df = 5,
                           spline_const = 4,
                           spline_linear = 4,
                           spline_quad = 4)
spline_estimate

##
## Estimated values:
## [1] 1583.0374335   30.5793023   -0.1980041
```

This last method fits a spline basis to the estimated PF values and then regresses the outcome on both the basis terms and the treatment to estimate the ADRF. This is described in Imai and van Dyk (2004) and Galagate and Schafer

(2015a). The estimated parameters correspond to coefficients for a quadratic model of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$.

```
ivd_estimate <- prop_spline_est(Y = TOTALEXP,
                                treat = packyears,
                                covar_formula = ~ 1,
                                covar_lin_formula = ~ 1,
                                covar_sq_formula = ~ 1,
                                data = full_data_orig,
                                e_treat_1 = full_data_orig$est_treat,
                                degree = 2,
                                wt = full_data_orig$HSQACCWT,
                                method = "different",
                                spline_df = 5,
                                spline_const = 4,
                                spline_linear = 4,
                                spline_quad = 4)
ivd_estimate

##
## Estimated values:
## [1] 1487.99099309   24.89207005   -0.05530696
```

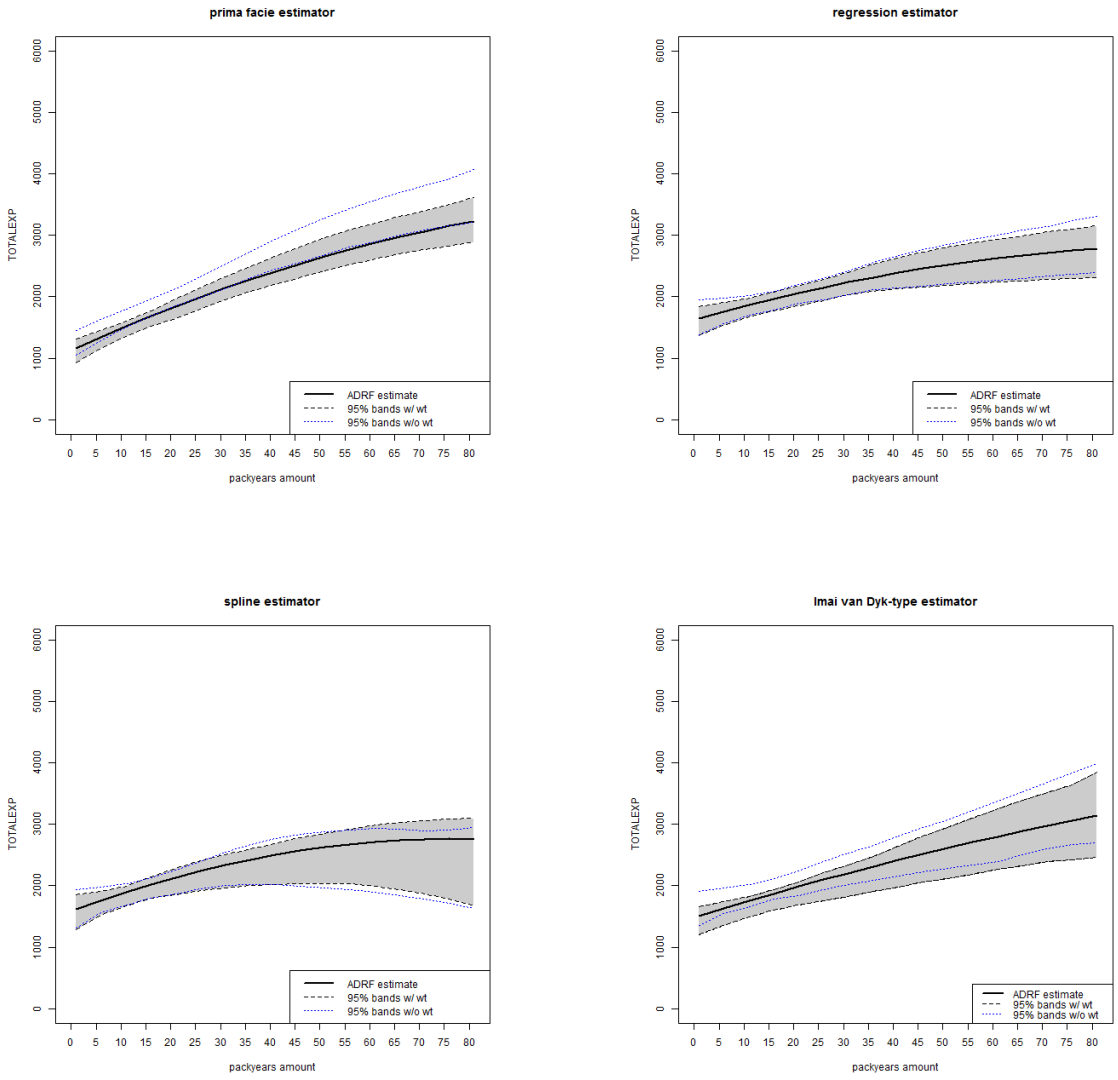

Figure 5.3: Estimated dose-response functions using 4 different methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process from the beginning.

5.3.6 Discussion

These four methods estimate the ADRF in a structured way and assume the true ADRF is a linear combination of a finite number of basis functions. Figure 5.3 shows an overall rising amount of `TOTALEXP` as `packyear` increases. Recall that in this example, the four estimators are restricted to fitting the ADRF as a polynomial of up to degree 2. Fitting more flexible models may give slightly different curves. The next section analyzes a different data set and will fit other flexible estimators such as BART, which allows for flexible response surfaces to estimate the ADRF.

5.4 Analysis of the Infant Health and Development Program

5.4.1 Introduction

The next example from the Infant Health and Development Program is described by Gross (1992):

> The Infant Health and Development Program (IHDP) was a collaborative, randomized, longitudinal, multisite clinical trial designed to evaluate the efficacy of comprehensive early intervention in reducing the developmental and health problems of low birth weight, premature infants. An intensive intervention extending from hospital discharge to 36 months corrected age was administered between 1985 and 1988 at eight different sites. The study sample of infants was stratified by birth weight (2,000 grams or less, 2,001-2,500 grams) and randomized to the Intervention Group or the Follow-Up Group.

The intervention (treatment) group received more support than the control group. In addition to the standard pediatric follow-up, the treatment group also received home visits and attendance at a special child development center. Although the

treatment was assigned randomly, families chosen for the intervention self-selected into different participation levels (Hill, 2011). Therefore, restricting our analysis to families in the intervention group and their participation levels leads to an observational setting.

In this section, even though families are randomly selected for intervention, we restrict our analysis to those selected for the treatment. These families choose the amount of days they attend the child development centers and this makes the data set, for practical purposes, an observational data set. We apply our methods on this subset of the data to estimate the ADRF for those who received the treatment.

We analyze this data set because the treatment variable, number of child development center days, is analyzed as a continuous variable. The data set we use was provided by Hill (2011).

5.4.2 Data description

Part of this data set is included in the supplement in Hill (2011), but does not include all the needed variables. The continuous treatment is available through the data repository at `icpsr.umich.edu`. To get the data, go to `http://www.icpsr.umich.edu/icpsrweb/HMCA/studies/9795?paging.startRow=51` and download DS141: Transport Format SAS Library Containing the 59 Evaluation Data Files - Download All Files (27.9 MB). After downloading the .zip file, extract the data file named "09795-0141-Data-card_image.xpt" to a folder and set the R working directory to this folder. The following instructions describe how to extract the continuous treatment variable.

Making sure the working directory contains "09795-0141-Data-card_image.xpt", the next step is to load the `Hmisc` package to read sas export files.

```
library(Hmisc)
mydata <- sasxport.get("09795-0141-Data-card_image.xpt")
data_58 <- mydata[[58]]
ihdp_raw <- data_58
# restricts data to treated cases
treated_raw <- ihdp_raw[which(ihdp_raw$tg == "I"),]
# continuous treatment variable
treat_value <- treated$cdays.t
```

The continuous treatment variable is merged with the data given in the supplement by Hill (2011) to create the data set for this section.

A few more steps are needed to clean and recode the data. We collect a subset of families eligible for the intervention and restrict the data set to families that use the child development centers at least once. The data set contains the outcome variable, `iqsb.36`, which is the measured iq of the child at 36 months. The treatment variable is the number of days the child attended the child development center divided by 100, `ncdctt` (i.e. `ncdctt` = 1.5 means 150 days in the child developement center). We select the covariates using a stepwise procedure to simplify the analysis.

5.4.3 Common support

```
overlap_temp <- overlap_fun(Y = iqsb.36,
                            treat = ncdctt,
                            treat_formula = t_formula,
                            data = data_set,
                            n_class = 3,
                            treat_mod = "Normal")

median_list <- overlap_temp[[2]]
overlap_orig <- overlap_temp[[1]]
overlap_3 <- overlap_temp[[3]]
fitted_values_overlap <- overlap_3$fitted.values
```

5.4.4 Estimating the ADRF

The BART estimator fits a rich outcome model on the treatment and covariates to create a flexible response surface (Hill, 2011). The flexible response surface imputes the missing potential outcomes. The estimated potential outcomes are averaged to get the estimated ADRF over a grid of treatment values.

```
bart_estimate <-  bart_est(Y = iqsb.36,
                           treat = ncdctt,
                           outcome_formula = iqsb.36 ~ ncdctt +
                           bw + female + mom.lths +
                           site1 + site7 + momblack +
                           workdur.imp,
                           data = full_data_orig,
                           grid_val = grid_treat)
```

The next method is described in Flores et al. (2012) and uses inverse weights to adjust for the covariates. First a treatment model is fit and GPS values are estimated. This is a method that uses weights to locally regress the outcome on nearby points. This is a local linear regression of the outcome, `iqsb.36`, on the treatment, `ncdctt`, with a weighted kernel. The weighted kernel is weighted by the reciprocal of the GPS values.

```
iw_estimate <- iw_est(Y = iqsb.36,
                      treat = ncdctt,
                      treat_formula = ncdctt ~ bw + female +
                       mom.lths + site1 + site7 +
                       momblack + workdur.imp,
                      data = full_data_orig,
                      grid_val = grid_treat,
                      bandw = 2 * bw.SJ(full_data_orig$ncdctt),
                      treat_mod = "Normal")
```

This next method, the Nadaraya-Watson based estimator, is similar to the inverse weighting method in the previous code chunk, but uses a locally constant regression.

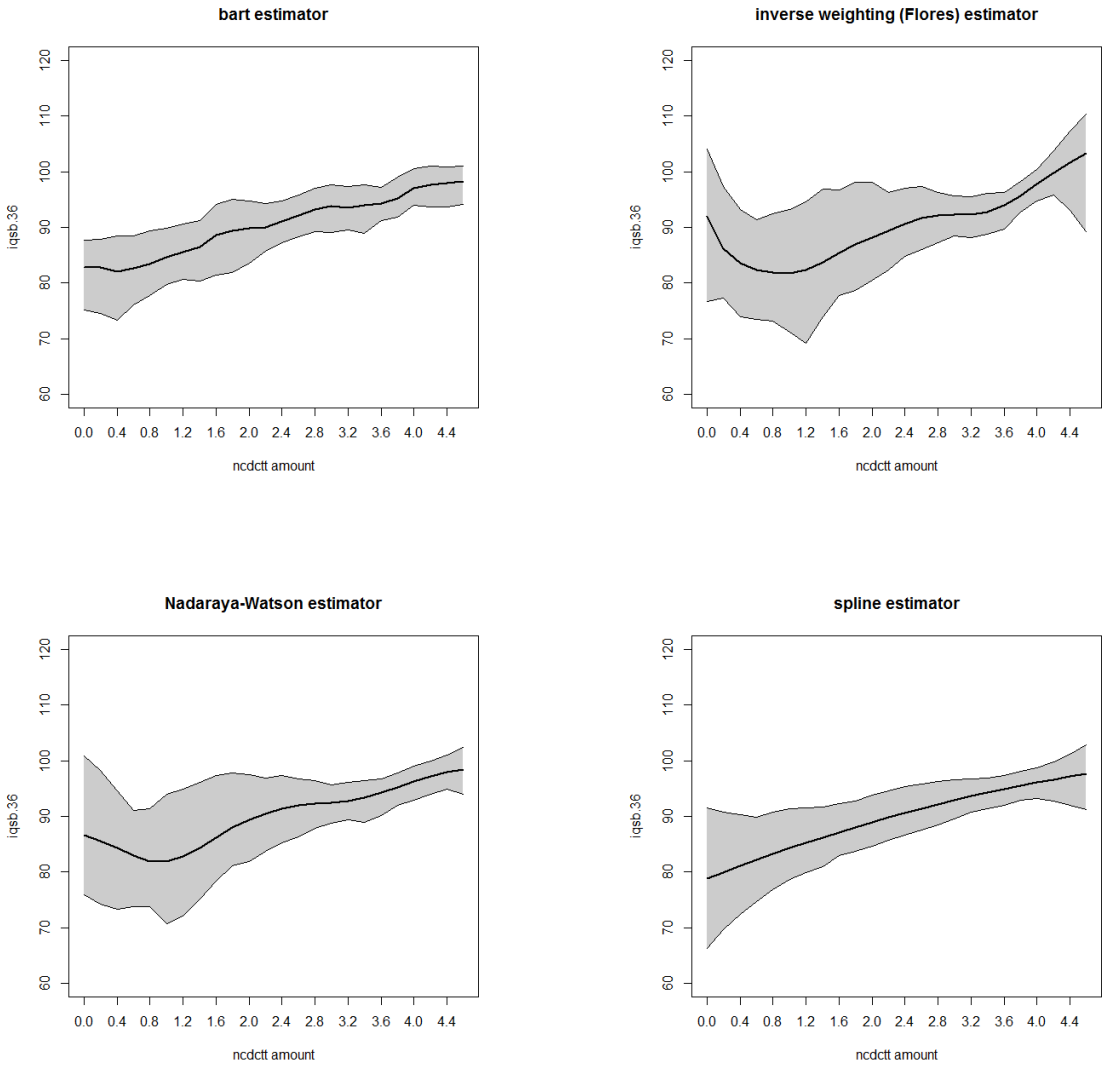

Figure 5.4: Estimated dose-response functions using four different methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process from the beginning.

```
nw_estimate <- nw_est(Y = iqsb.36,
                      treat = ncdctt,
                      treat_formula = ncdctt ~ bw + female +
                        mom.lths + site1 + site7 + momblack +
                        workdur.imp,
                      data = full_data_orig,
                      grid_val = grid_treat,
                      bandw = 2 * bw.SJ(full_data_orig$ncdctt),
                      treat_mod = "Normal")
```

The propensity spline estimator is a generalization of the prima facie and regression prediction method in Chapter 3. In this example, the estimator is restricted to a polynomial of up to degree 2 of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$.

```
spline_estimate <- prop_spline_est(Y = iqsb.36,
                                   treat = ncdctt,
                                   covar_formula = ~ bw + female +
                                     mom.lths + site1 + site7 +
                                     momblack + workdur.imp,
                                   covar_lin_formula = ~ 1,
                                   covar_sq_formula = ~ 1,
                                   data = full_data_orig,
                                   e_treat_1 = full_data_orig$est_treat,
                                   degree = 2,
                                   wt = NULL,
                                   method = "different",
                                   spline_df = 5,
                                   spline_const = 2,
                                   spline_linear = 2,
                                   spline_quad = 2)
```

5.4.5 Discussion

The plots in Figure 5.4 show the estimated relationship of IQ at 36 months, `iqsb.36`, on number of days in the child development care center, `ncdctt`. The inverse weighting and Nadaraya-Watson show a decreasing trend for `ncdctt` $\in (0, 0.8)$, but an increasing trend for `ncdctt` > 0.8. These estimators are jagged because of the

bandwidth selection. In this example, we use twice the Sheather-Jones bandwidth estimate to select the bandwidth. Picking a larger bandwidth will give smoother estimates. The BART and propensity spline estimators have a generally increasing trend.

5.5 Conclusion

In this chapter, we have demonstrated how to estimate ADRFs using different statistical techniques using the R package `causaldrf`, both for simulated and real data, by correcting for confounding variables. `causaldrf` can accommodate a wide array of treatment models, is user friendly, and does not require extensive programming. This contribution of the R package `causaldrf` will make ADRF estimation more accessible to applied researchers. In future updates of the package, the functions will be adapted to an even wider range of problems.

Chapter 6: Missing data and causal inference with a continuous outcome: an application to the National Growth and Health Study

6.1 Introduction

6.1.1 Problem setting

Obesity is currently a major health issue in the U.S. that is related to problems such as diabetes, cardiovascular disease, and other ailments (Mokdad et al., 2003; Hubert et al., 1983). Possible solutions for lowering the rate of obesity in the U.S. include encouraging people to increase their physical activity, to eat fewer calories, to watch less television, or to decrease the amount of sedentary activities (Sallis and Glanz, 2009). This paper applies causal inference methods to observational data from the National Growth and Health Survey (NGHS) to explore the effect of different levels of physical activity, diet and the amount of TV watched on obesity in adolescent girls (obesity is defined as a measure of body mass index (BMI)). Our primary analysis deals with physical activity and BMI. Additional analyses are included in the Appendix.

6.1.2 National Growth and Health Study data

The NGHS was a multicenter, 10-year longitudinal study of 2379 girls from the ages of 9-10 through 18-19. Data were collected yearly and includes variables

Table 6.1: Correlates of physical activity.

Category of Determinant	Determinants
Demographic and biological	age, BMI, skinfolds, single parent
Race	white, black
Psychological	self-efficacy, self-perception, enjoyment
Behavioral	watching TV, smoking
Social and Cultural	parent activity, parent support
Physical Environment	access to facilities

related to obesity development, but not all variables were collected each year. Year 1 of the NGHS refers to when the girls are 9-10 years old and year 10 when they are 18-19 years old. The purpose of the NGHS was to explore racial differences in diet, physical activity, demographic, psychological, and social factors associated with obesity development in young girls. The retention rate was 0.88 on average over the 10 year period and 58% completed all 10 annual visits. Black and white girls were almost equally split (51% to 49%). More details are given in Kimm et al. (2001).

In addition to the activity measures, the NGHS data set contains many other covariates: demographic, personal history, physical measurements, biochemical determinations, diet, physical activity, and psychosocial characteristics. Some examples of covariates are given in Table 6.1 (Van der Horst et al. (2007); Sallis et al. (2000)). Table 6.2 contains summary statistics on the households. Figures 6.1 and 6.2 display some longitudinal characteristics about the girls in the study.

6.2 Methods

6.2.1 Missing data

To deal with missing data, we use multiple imputation. We assume the data are missing at random. There are three main steps when conducting data analysis when using multiple imputation. First, multiple imputation is used to fill in missing

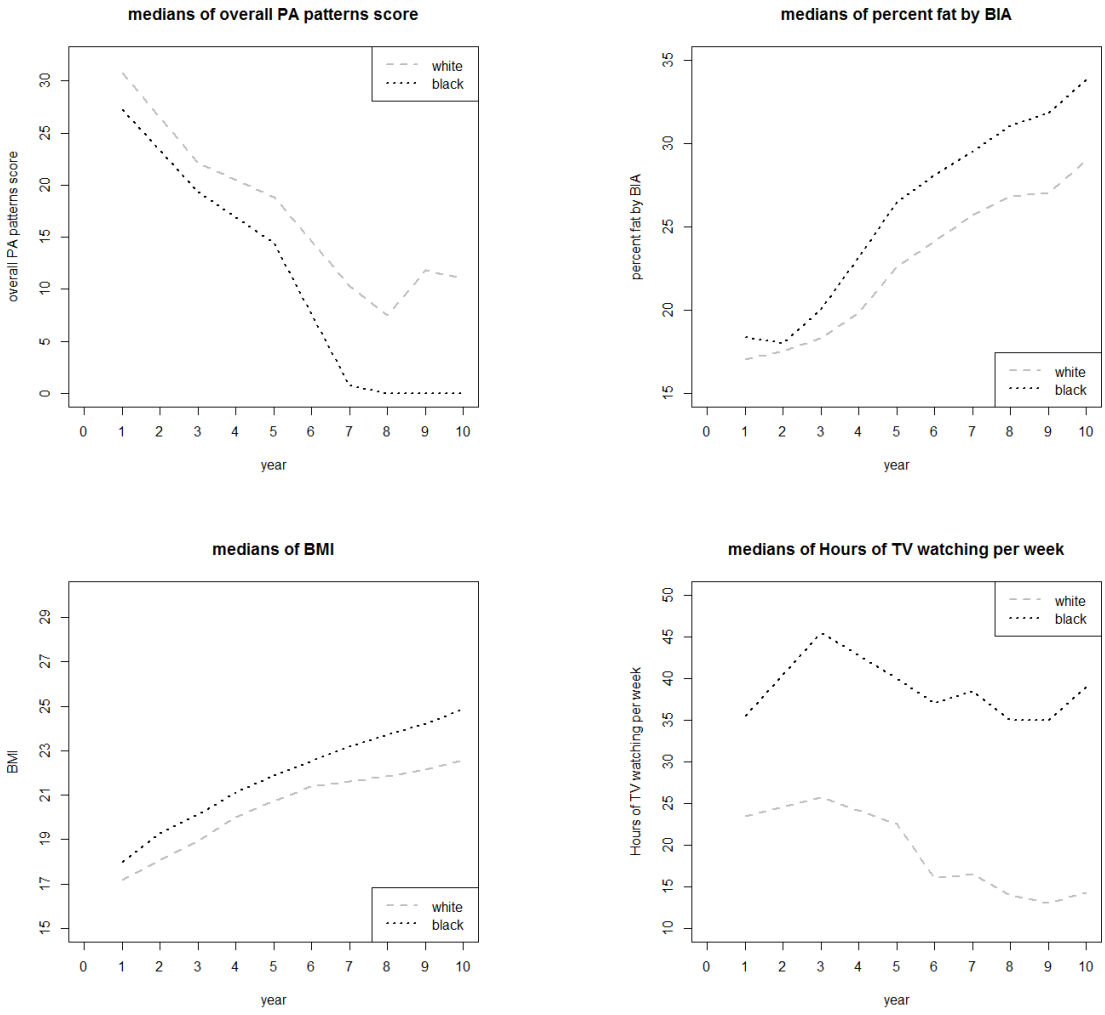

Figure 6.1: Median values of activity levels, body fat percentage, BMI, and TV viewing at different years.

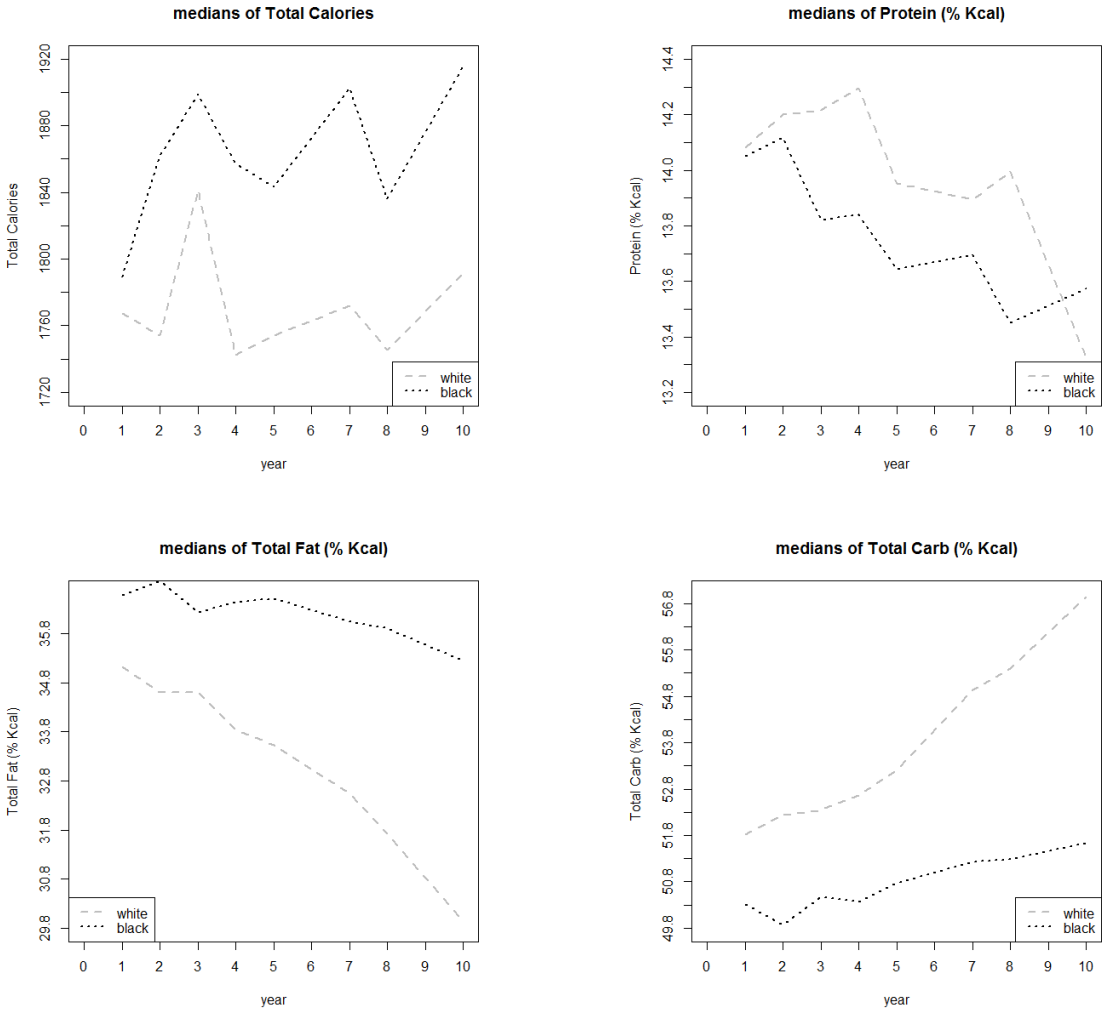

Figure 6.2: Median values of total calories, protein, fat, and carbohydrates at different years.

Table 6.2: Summary statistics by race in year 3.

	Black Girls (N = 1155)	**White Girls (N = 1073)**
Education (%)		
High School or Less	31.1	19.5
Some College	47.7	30.3
$\geq$ 4 yr college	21.2	50.2
Income (%)		
$< \$20,000$	46.4	16.3
$\$20,000 - \$39,999$	29.9	33.2
$\geq \$40,000$	23.7	50.5
No. of parents in HH (%)		
1	43.5	18.2
2	56.5	81.8

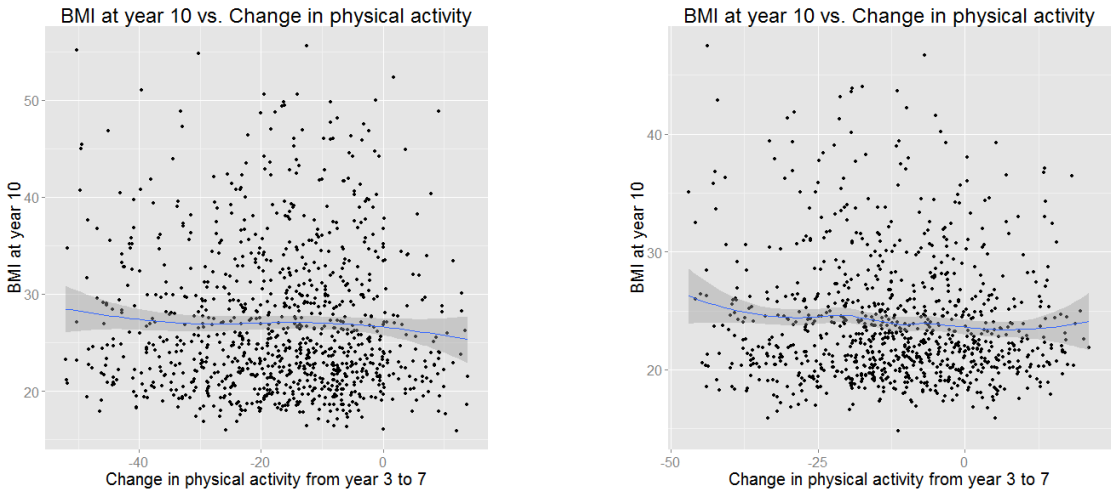

Figure 6.3: Scatterplots with overlapped smoothers for $\Delta PA_{7,3}$ vs. BMI. Plot representing black girls are on the left column and white girls on the right.

data by creating imputed datasets. Second, the m imputed datasets are created and analyzed individually. Third, the estimated parameters and their variances are pooled according to the rules for combining the results of multiply imputed datasets (Schafer, 1997; Little and Rubin, 2002).

We use the R package `Amelia` (Honaker et al., 2011) to implement the multiple imputation procedure. We impute $m = 5$ data sets, analyze each imputed data set separately, and then combine the estimated parameters and standard errors.

Table 6.3: Summary statistics of categorical variables in year 3. These come from a multiply imputed dataset.

		Black			*White*		
Variable	**Levels**	**n**	%	$\sum$%	**n**	%	$\sum$%
Max parental education (grouped)	1	264	28.8	28.8	146	17.9	17.9
	2	438	47.8	76.5	257	31.5	49.3
	3	215	23.4	100.0	414	50.7	100.0
	all	917	100.0		817	100.0	
Household income (grouped)	1	231	25.2	25.2	56	6.8	6.8
	2	179	19.5	44.7	74	9.1	15.9
	3	280	30.5	75.2	266	32.6	48.5
	4	227	24.8	100.0	421	51.5	100.0
	all	917	100.0		817	100.0	
Happy with body looks	1	300	32.7	32.7	119	14.6	14.6
	2	434	47.3	80.0	499	61.1	75.7
	3	142	15.5	95.5	158	19.3	95.0
	4	41	4.5	100.0	41	5.0	100.0
	all	917	100.0		817	100.0	

6.2.2 Estimating the ADRF

In order to adjust for confounding bias, we fit a treatment model for the propensity function (PF) or generalized propensity score (GPS). We assume strong ignorability defined in Chapter 2. We use the candidate variables associated with physical activity that have been identified by Van der Horst et al. (2007), Biddle

et al. (2005), and Sallis et al. (2000). These candidate variables are described in Tables 6.3 and 6.4 and include a spectrum of candidate covariates is rich and includes physical measurements, physical activity, socioeconomic status, family beliefs and attitudes, diet and many others. The treatment variable is the change in physical activity from year 3 to year 7 ($\Delta PA_{7,3}$), and the pool of candidate pre-treatment covariates come from year 3.

The ADRF is fit using different methods. We explore linear fits, quadratic fits, and non-parametric fits (Schafer and Kang, 2008). These different estimators are described in Chapters 2 and 3. Plots and tables are included in the later sections. We use the R package `causaldrf` from Chapter 5 to estimate the ADRFs. Table 6.6 shows results for estimating a linear fits for the NGHS data.

6.3 Results

The following plots show the estimated dose response functions using different methods. Figure 6.4 shows the ADRF estimates using semiparametric fits. A set of gridpoints are placed along the domain of $\Delta PA_{7,3}$ and 95% pointwise standard errors are calculated using 1000 bootstrap samples.

Table 6.6 shows negative slope estimates for black girls for all methods except for the prima facie estimate, but the slope estimate is not significant. The slope for white girls is negative for all methods, but are also not significant.

6.4 Discussion

We see that the prima facie estimator may show significant relationships between obesity and other factors, but adjusting for possible confounders attenuates the relationships. More study is needed to understand these relationships and either experiments or quasi-experiments can provide more evidence.

Table 6.4: Summary of continuous variables in year 3. Results come from a multiply imputed data set of black and white girls. Correlation with $\Delta PA_{7,3}$ is listed.

	Black Girls				*White Girls*			
	Mean	S.D.	Corr	T	Mean	S.D.	Corr	T
BMI	21.46	4.97	0.04	1.19	19.56	3.94	-0.04	-1.23
overall PA	22.29	14.00	-0.83	-45.47	25.18	14.88	-0.66	-24.82
TV amt.	94.53	39.96	-0.12	-3.68	55.14	33.15	-0.06	-1.58
PA diary	498.86	415.17	-0.18	-5.64	494.85	335.47	-0.15	-4.36
PA oth.	21.18	45.34	0.00	0.04	34.33	53.70	-0.09	-2.44
Cal avg.	275.45	180.58	-0.02	-0.71	303.22	155.30	-0.06	-1.86
schol.cmp.	2.97	0.66	-0.03	-0.78	3.01	0.66	0.00	0.09
ath.cmp.	2.77	0.68	-0.13	-3.83	2.77	0.74	-0.04	-1.18
self-worth	3.21	0.61	-0.05	-1.64	3.17	0.61	0.04	1.19
schol.imp.	3.45	0.71	-0.04	-1.18	3.38	0.68	-0.06	-1.80
ath.imp.	2.45	0.85	-0.04	-1.10	2.61	0.84	-0.07	-1.97
anxiety	11.11	6.16	-0.00	-0.03	10.75	6.63	-0.02	-0.58
calories	2054.86	794.64	-0.03	-0.81	1881.17	548.39	-0.02	-0.44
chol.	126.21	59.87	0.01	0.32	113.19	51.27	-0.07	-1.89
sucrose	57.12	38.96	-0.05	-1.47	49.94	29.60	0.01	0.34
sugars	130.25	63.47	-0.02	-0.47	122.76	50.99	0.03	0.79

Table 6.5: Summary statistics for outcome and treatment variables. BMI and overall PA are measured in year 10. These results are from a multiply imputed dataset.

	Variable	**n**	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**	**s**	**IQR**
Black										
	BMI	917	13.1	17.8	20.1	21.5	24.0	40.7	5.0	6.2
	PA	917	0.0	11.9	19.2	22.3	30.2	76.0	14.0	18.3
	$\Delta PA_{7,3}$	917	-57.1	-25.9	-14.9	-16.6	-6.5	22.2	14.7	19.4
White										
	BMI	817	12.9	16.6	18.8	19.6	21.5	39.3	3.9	4.9
	PA	817	0.0	14.2	22.2	25.2	33.2	100.6	14.9	19.0
	$\Delta PA_{7,3}$	817	-55.6	-21.2	-11.3	-12.2	-1.1	21.1	15.4	20.1

Table 6.6: Slope estimates and standard errors using different methods when using multiple imputation. All values are multiplied by 100.

	black		*white*	
	Slope	(SE)	Slope	(SE)
prima facie	0.55	(2.64)	-2.01	(1.42)
importance weighting	-8.88	(5.04)	-5.19	(9.88)
inverse second moment weighting	-1.76	(3.07)	-1.15	(1.38)
regression prediction	-1.76	(2.16)	-1.15	(1.20)
aipwee	-1.76	(2.23)	-1.13	(1.35)
scalar weighted regression	-0.60	(2.46)	-0.51	(1.60)
propensity spline	-1.73	(2.16)	-1.18	(1.20)
imai van dyk	-1.61	(3.53)	-0.34	(1.48)

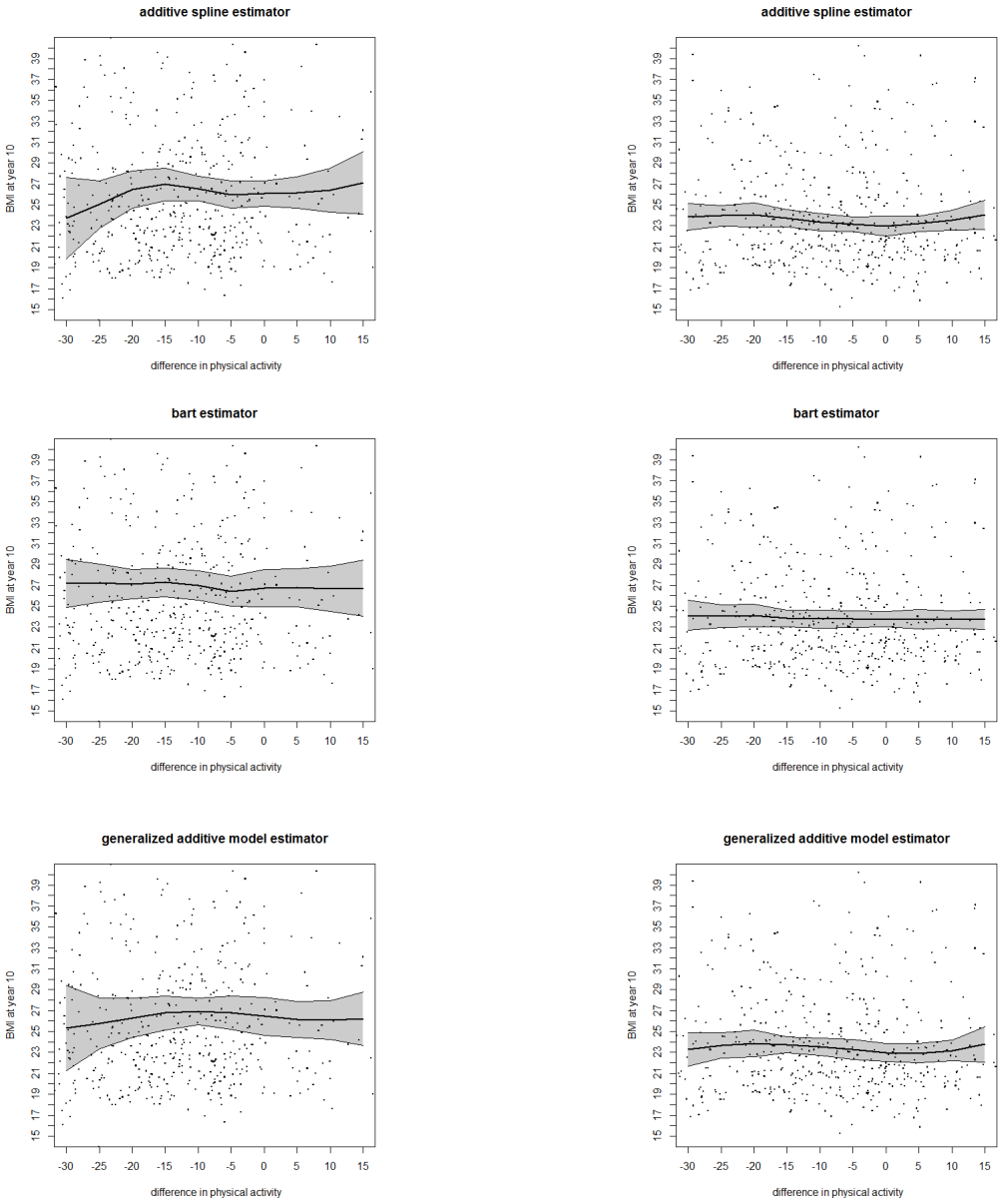

Figure 6.4: Estimated dose-response functions using 6 different methods with 95% pointwise standard errors using the multiply imputed data. Black girls are on the left and white girls are on the right. The standard errors are estimated by combining multiple imputation bootstrap standard errors.

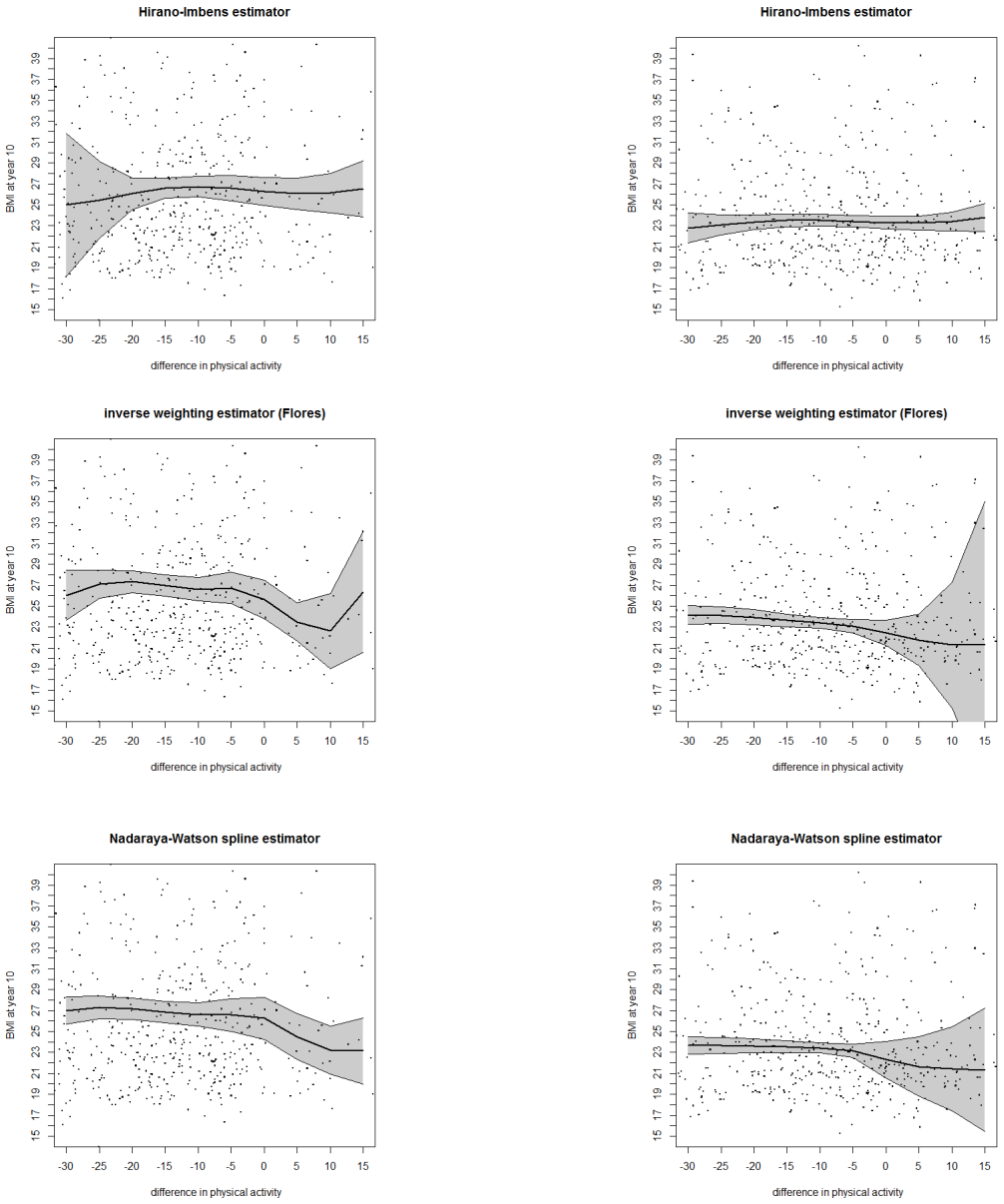

Figure 6.5: Estimated dose-response functions using 3 different methods with 95% pointwise standard errors using the multiply imputed data. Black girls are on the left and white girls are on the right. The standard errors are estimated by combining multiple imputation bootstrap standard errors.

Table 6.7: Slope estimates and standard errors using different methods when using multiple imputation. TV year 3. All values are multiplied by 100.

	black		*white*	
	Slope	(SE)	Slope	(SE)
prima facie	0.81	(0.62)	2.71	(0.66)
importance weighting	-0.26	(0.54)	0.11	(0.93)
inverse second moment weighting	-1.58	(0.71)	-1.08	(0.83)
regression prediction	-0.23	(0.39)	0.64	(0.46)
aipwee	-0.40	(0.40)	0.25	(0.49)
weighted regression	-0.16	(0.41)	0.79	(0.50)
scalar weighted regression	-0.19	(0.40)	0.59	(0.56)
propensity spline	-0.24	(0.39)	0.62	(0.46)
imai van dyk	-0.13	(0.50)	1.46	(0.59)

6.5 Appendix

We perform additional analyses to understand how other factors may affect BMI at year 10. The two factors we explore include the amount of TV viewed in year 3 and the percentage of protein consumed in their diets.

Table 6.7 shows estimated slopes for BMI at year 10 on TV units viewed in year 3. For black girls, the slopes are estimated negative for all methods except for the prima facie. For the white girls, the prima facie estimate is positive and significant, but all other methods except for the method by Imai and van Dyk are not significant.

Table 6.8 shows the estimated slopes of BMI at year 10 on % calories consumed from protein. The prima facie estimate is positive and significant, but after adjusting for possible confounders, the relationship is not significant.

The amount of TV viewed and percentage of protein consumed affect black and white girls differently, after adjusting for possible confounders.

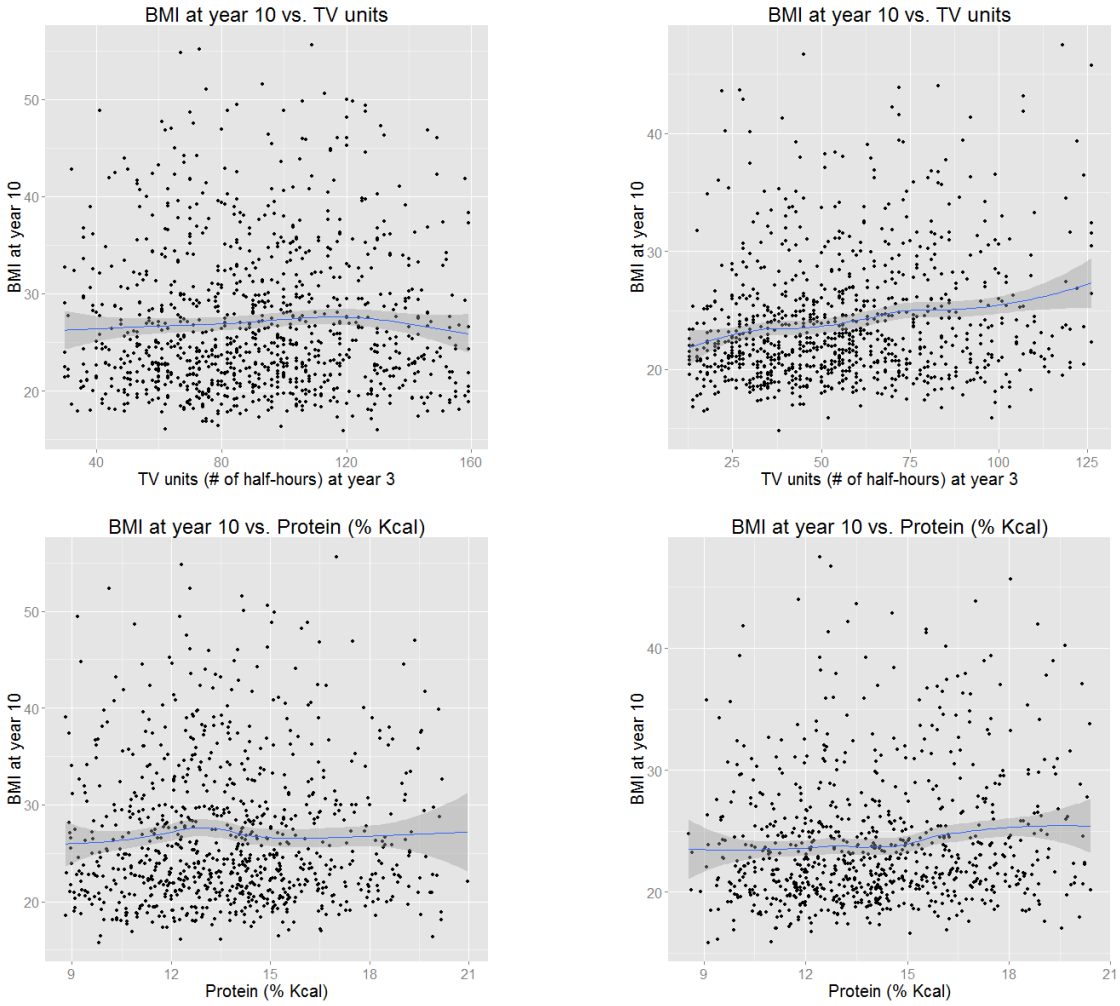

Figure 6.6: Scatterplots with overlapped smoothers for TV units and protein (% Kcal) vs. BMI. Plot representing black girls are on the left column and white girls on the right.

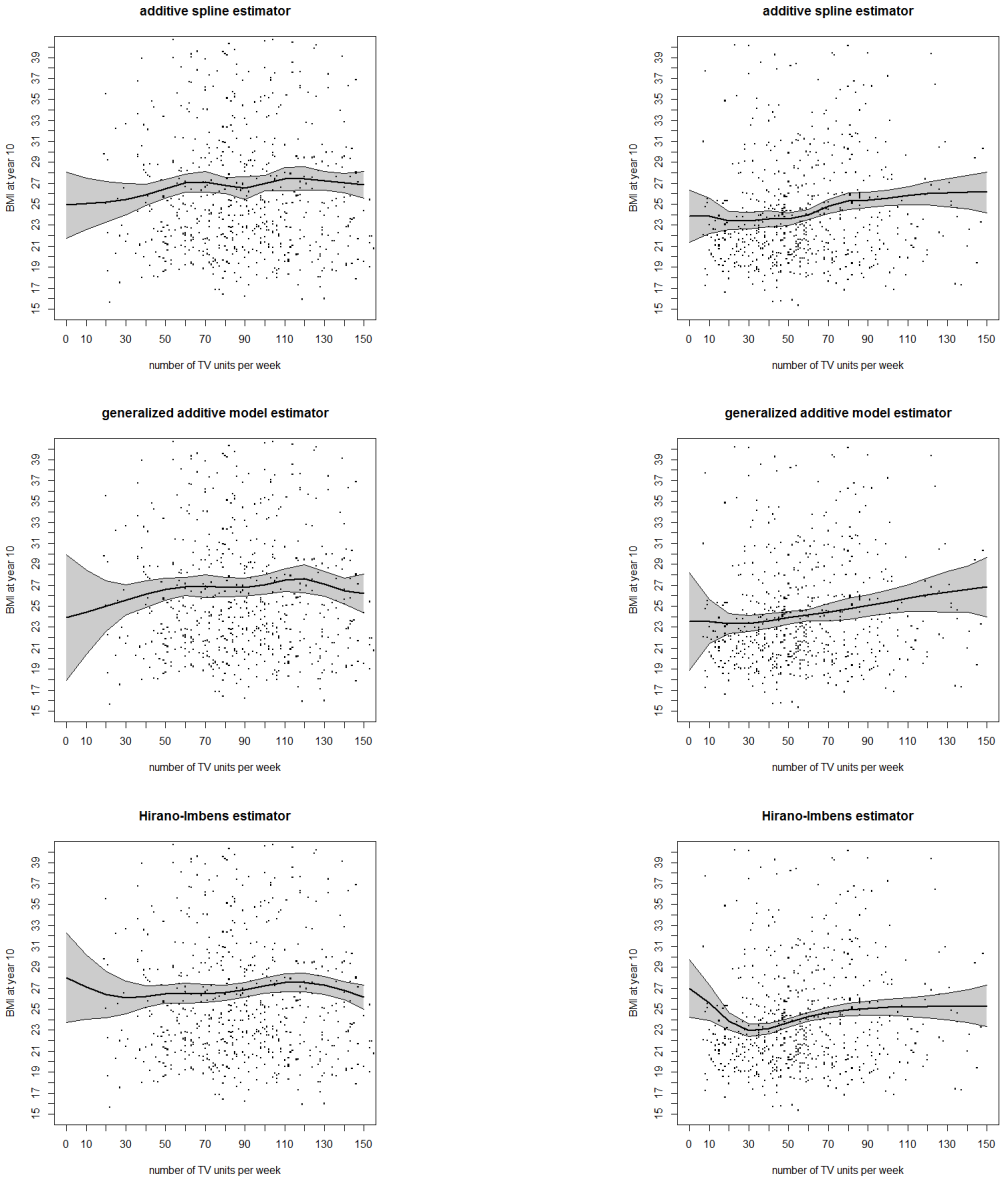

Figure 6.7: Estimated dose-response functions using 3 different methods with 95% pointwise standard errors using the multiply imputed data. Black girls are on the left and white girls are on the right. TV in year 3. The standard errors are estimated by combining multiple imputation bootstrap standard errors.

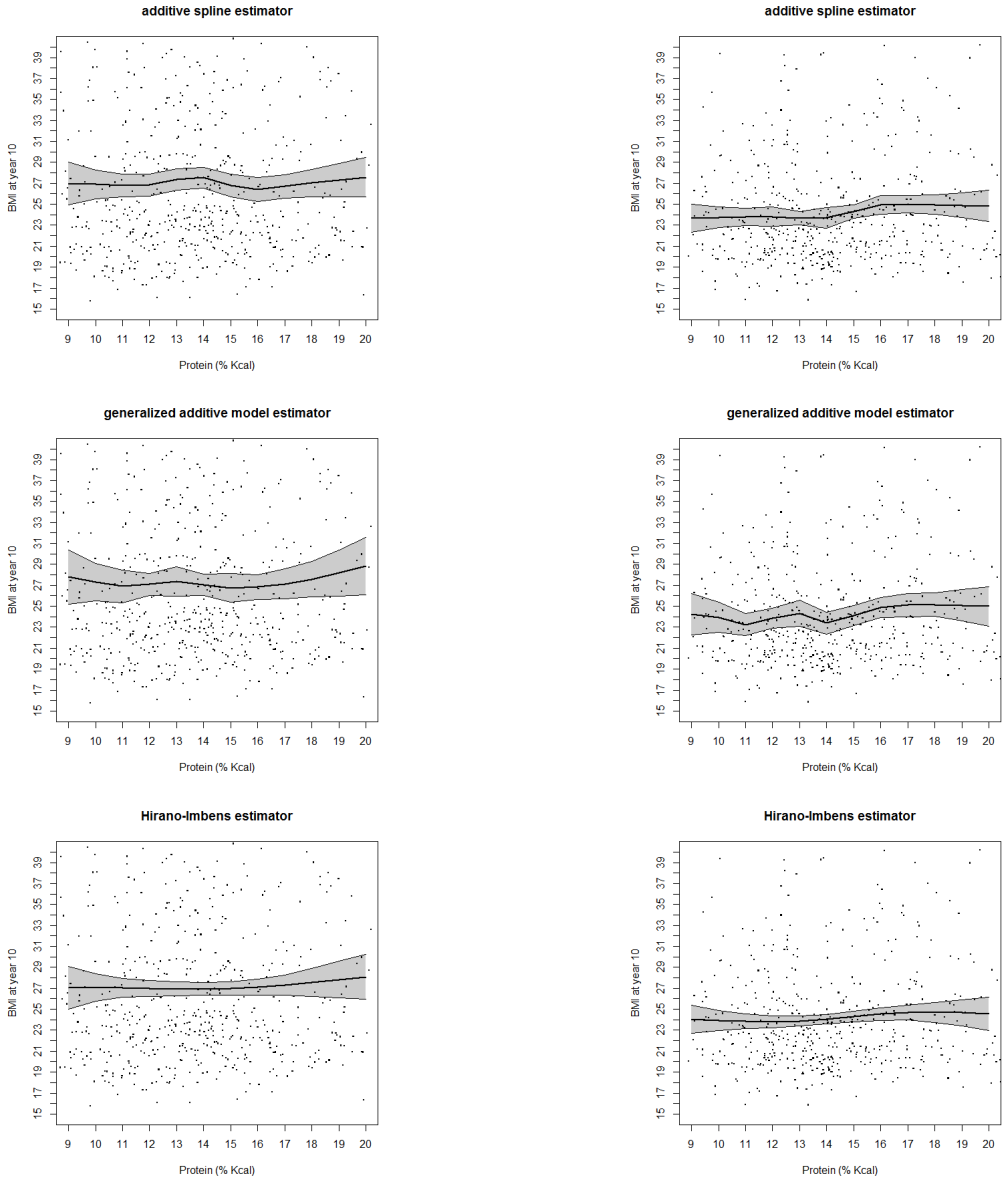

Figure 6.8: Estimated dose-response functions using 3 different methods with 95% pointwise standard errors using the multiply imputed data. Black girls are on the left and white girls are on the right. Protein in year 7. The standard errors are estimated by combining multiple imputation bootstrap standard errors.

Table 6.8: Slope estimates and standard errors using different methods when using multiple imputation. Protein (% Kcal) in year 7. All values are multiplied by 100.

	black		*white*	
	Slope	(SE)	Slope	(SE)
prima facie	7.06	(8.04)	12.98	(6.61)
importance weighting	-17.48	(36.69)	24.34	(23.39)
inverse second moment weighting	-0.93	(6.76)	7.18	(4.84)
regression prediction	-1.00	(4.33)	6.62	(4.17)
aipwee	-0.87	(6.87)	7.32	(5.14)
weighted regression	-2.22	(5.09)	8.13	(4.83)
scalar weighted regression	1.47	(7.69)	3.66	(7.06)
propensity spline	-1.47	(4.69)	6.55	(4.03)
imai van dyk	0.46	(8.74)	6.57	(4.66)

Chapter 7: Discussion and future work

7.1 Conclusions

This dissertation focused on the problem of estimating potential outcomes when the treatment and outcome are both continuous. The methods developed in Chapters 3 through 6 will help researchers analyze data with continuous treatments and outcomes to understand ADRFs.

The main findings of this dissertation include a set of new estimators by modifying estimating functions with weighting (Chapter 3). Our methods assume the true ADRF is represented by a linear combination of a finite set of basis functions. We modified some existing estimators in Chapters 4 through 6 (Flores et al., 2012), (Hirano and Imbens, 2004), (Imai and van Dyk, 2004), Hill (2011), and (Robins et al., 2000). In Chapter 5, we describe the R package `causaldrf` which contains a set of functions to estimate the ADRF in different settings (Galagate and Schafer, 2015b). We hope that these new methods and the software make causal inference with a continuous treatment more applicable to a broad group of researchers.

7.2 Future extensions

Some of the issues in this dissertation were motivated by U.S. Census Bureau applications, but have not been carried out due to the complexity of the problems. At the U.S. Census Bureau, causal inference methods can be applied to the challenges of reduced response rates and diminishing resources. The operational

questions of "what mode of data collection produces the highest response rates at the lowest cost?" or "what would happen to the accuracy of the unemployment rate if a modification is made on the data collection process?" are questions to which causal inference methods may be applied in the future.

The methods developed in this dissertation can be extended in a few general ways. Extensions include: allowing estimators to encompass longitudinal data, creating estimators that incorporate complex survey designs, and understanding missing data in this setting.

In this dissertation, the outcomes are assumed to be continuous, but more flexible outcome models are needed. Extensions to binary, discrete, and other outcomes is a future research topic. Some of the estimators require a parametric model for fitting the treatment. A potential future topic of research is to make the treatment model more flexible by allowing for empirical densities or nonparametric densities for the treatment model. Testing and loosening the common assumptions such as interference is an interesting topic for future work.

Variable selection is a general research topic in statistics and also applies to this setting. Deciding which variables to include in the treatment model, especially for methods presented in Chapter 3, should be explored.

Bibliography

Altshuler, B. (1981). Modeling of dose-response relationships. *Environmental Health Perspectives*, 42:23.

Bia, M., Flores, C. A. F., Flores-Lagunes, A., and Mattei, A. (2014). A stata package for the application of semiparametric estimators of dose-response functions. *The Stata Journal*, 14(3):580–604.

Biddle, S. J., Whitehead, S. H., O Donovan, T. M., and Nevill, M. E. (2005). Correlates of participation in physical activity for adolescent girls: a systematic review of recent literature. *Journal of Physical Activity & Health*, 2(4):423.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Cohen, S. B. (2000). *Sample Design of the 1997 Medical Expenditure Panel Survey, Household Component*. US Department of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality.

D'Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dominici, F., Daniels, M., Zeger, S. L., and Samet, J. M. (2002). Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, 97(457):100–111.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Review of Economics and Statistics*, 94(1):153–171.

Galagate, D. and Schafer, J. (2015a). Causal inference with a continuous treatment and outcome: alternative estimators for parametric dose-response models. *Manuscript in preparation*.

Galagate, D. and Schafer, J. (2015b). *causaldrf: Tools for Estimating Causal Dose Response Functions*. R package version 0.3.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2014). Quantifying causal effects of road network capacity expansions on traffic volume and density via a mixed model propensity score estimator. *Journal of the American Statistical Association*, 109(508):1440–1449.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419.

Gross, R. T. (1992). *Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight; Premature Infants in the United States, 1985-1988*. Inter-University Consortium for Political and Social Research.

Guo, S. and Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications*, volume 11. Sage Publications.

Hammersley, J. (2013). *Monte Carlo Methods*. Springer Science & Business Media.

Hernan, M. and Robins, J. (2016). *Causal Inference*. CRC, forthcoming.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1).

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164:73–84.

Ho, D., Imai, K., King, G., and Stuart, E. (2006). Matchit: Nonparametric preprocessing for parametric casual inference. *R package version*, 2:2–11.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

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(260):663–685.

Hubert, H. B., Feinleib, M., McNamara, P. M., and Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the framingham heart study. *Circulation*, 67(5):968–977.

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467).

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, G. W., Rubin, D. B., and Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, pages 778–794.

Jesus, J. and Chandler, R. E. (2011). Estimating functions and the generalized method of moments. *Interface focus*, 1(6):871–885.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333.

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics*, 112(1):135–151.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pages 523–539.

Kimm, S. Y., Barton, B. A., Obarzanek, E., McMahon, R. P., Sabry, Z. I., Waclawiw, M. A., Schreiber, G. B., Morrison, J. A., Similo, S., and Daniels, S. R. (2001). Racial divergence in adiposity during adolescence: the nhlbi growth and health study. *Pediatrics*, 107(3):e34–e34.

King, G. and Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, 53(04):623–658.

Kluve, J., Schneider, H., Uhlendorff, A., and Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2):587–617.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14:949–968.

Little, R. J., An, H., Johanns, J., and Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods*, 5(4):459.

Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21(1):121–145.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253.

Maldonado, G. and Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology*, 31(2):422–429.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA: The Journal of the American Medical Association*, 289(1):76–79.

Moodie, E. E. and Stephens, D. A. (2012). Estimation of dose–response functions for longitudinal data using the generalised propensity score. *Statistical Methods in Medical Research*, 21(2):149–166.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

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

Pattanayak, C. W., Rubin, D. B., and Zell, E. R. (2011). Propensity score methods for creating covariate balance in observational studies. *Revista Española de Cardiología (English Edition)*, 64(10):897–903.

Pearl, J. (2011). Graphical models, potential outcomes and causal inference: comment on linquist and sobel. *NeuroImage*, 58(3):770–771.

Pearl, J. (2012). The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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(387):516–524.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, 13(12):855–857.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469).

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

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

Sallis, J. F. and Glanz, K. (2009). Physical activity and food environments: solutions to the obesity epidemic. *Milbank Quarterly*, 87(1):123–154.

Sallis, J. F., Prochaska, J. J., Taylor, W. C., et al. (2000). A review of correlates of physical activity of children and adolescents. *Medicine and Science in Sports and Exercise*, 32(5):963–975.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, 13(4):279.

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(448):1096–1120.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690.

Singh, R. P., Petroni, R. J., Allen, T. M., et al. (1994). Oversampling in panel surveys. *Bureau of the Census*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Van der Horst, K., Paw, M., Twisk, J. W., and Van Mechelen, W. (2007). A brief review on correlates of physical activity and sedentariness in youth. *Medicine and Science in Sports and Exercise*, 39(8):1241.

Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

Wang, J. (2015). *Exposure–Response Modeling: Methods and Practical Implementation*. CRC press.

Zhao, S., van Dyk, D. A., and Imai, K. (2014). Propensity-score based methods for causal inference in observational studies with fixed non-binary treatments. *Manuscript*.

Zhu, Y., Coffman, D. L., and Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40.