Title
Methods for Estimating Causal Effects for Multivariate Continuous Exposure

Author
Williams, Justin Randall

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Methods for Estimating Causal Effects for Multivariate Continuous Exposure

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Justin Randall Williams

2020

ABSTRACT OF THE DISSERTATION

Methods for Estimating Causal Effects for Multivariate Continuous Exposure

by

Justin Randall Williams
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2020
Professor Catherine M. Crespi, Chair

The generalized propensity score (GPS) is an extension of the propensity score for use with quantitative or continuous exposures (e.g., dose of medication or years of education). Current GPS methods allow estimation of the dose-response relationship between a single continuous exposure and an outcome. However, in many real-world settings, there are multiple exposures occurring simultaneously that could be causally related to the outcome. We propose a multivariate GPS method (mvGPS) that allows estimation of a dose-response surface that relates the joint distribution of multiple continuous exposure variables to an outcome. The method involves generating weights under a multivariate normality assumption on the exposure variables. Focusing on scenarios with two exposure variables, we show via simulation that the mvGPS method can achieve balance across sets of confounders that may differ for different exposure variables and reduces bias of the treatment-effect estimates under a variety of data generating scenarios. We apply the mvGPS method to an analysis of the joint effect of two types of intervention strategies to reduce childhood obesity rates. The methods can be implemented using the mvGPS R package available on CRAN.

The dissertation of Justin Randall Williams is approved.

Hua Zhou

Tom Belin

May C. Wang

Catherine M. Crespi, Committee Chair

University of California, Los Angeles

2020

Thank you first and foremost to my family. My mom, Teresa, dad, Randy, and brother, Aaron, have always provided me with support throughout this journey. I cannot thank you enough for helping me cross this hurdle. Thanks also to my girlfriend, Veronica, for constantly being there for me when I needed someone to listen to my practice presentations, share my ideas with, or provide a word of encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to acknowledge my advisor, Dr. Crespi, who has helped guide me during my academic career and spent countless hours helping me refine my writing, challenging me to think critically, and pushing me to continually strive for more. Thank you to my mentors in the UCLA Semel Institute including Drs. Connie Kasari and Wendy Shih who both provided me with opportunities to grow not only in my statistical knowledge but in the ability to collaborate and communicate effectively. To all of my fellow students in the Department of Biostatistics that I had to pleasure to interact with during my time, I cannot thank you enough for your friendship and making this a truly remarkable experience.

VITA

2009-2013	B.A. (Mathematics), Boston College, Chestnut Hull, Massachusetts
2014-2016	M.S. (Biostatistics), University of California, Los Angeles (UCLA), Los Angeles, California
2015-2019	Graduate Student Researcher, UCLA Semel Institute for Neuroscience & Human Behavior, Connie Kasari Lab
2016	Teaching Assistant, UCLA Department of Biostatistics

PUBLICATIONS

Williams, J.R. & Crespi, C.M. (2020). "Causal inference for multiple continuous exposures via the generalized propensity score", *arXiv pre-print*, arXiv:2009.13767. Note: CRAN R package mvGPS.

Williams, J.R., Kim H., & Crespi, C.M. (2020). "Modeling observations with a detection limit using a truncated normal distribution with censoring", *BMC Med Res Methodol*, **20**:170. doi: 10.1186/s12874-020-01032-9. Note: CRAN R package tcensReg.

Williams, J, Bravo HC, Tom J & Paulson, JN. (2020). "microbiomeDASim: Simulating longitudinal differential abundance for microbiome data [version 2; peer review: 2 approved]", *F1000Research*, **8**:1769. doi: 10.12688/f1000research.20660.2. Note: Bioconductor R package microbiomeDASim.

Dean, M., Williams J, Kasari, C. & Orlich, O. (2020). "Adolescents with autism spectrum disorder and social skills groups at school: A randomized trial comparing intervention environment and peer composition", *School Psychology Review*, **49**(1):60-73. doi: 10.1080/2372966X.2020.1716636.

Gulsrud, A., Carr T., Williams J, Panganiban J., Jones F., Kimbrough J., Shih W., & Kasari, C. (2019). "Developmental screening and early intervention in a childcare setting for young children at risk for autism and other developmental delays: A feasibility trial", *Autism Research*, **12**(9):1423-1433. doi: 10.1002/aur.2160.

Locke, J., Williams J, Shih W., & Kasari, C. (2017). "Characteristics of socially successful elementary school-aged children with autism", *Journal of Child Psychology and Psychiatry*, **58**(1):94-102. doi: 10.1111/jcpp.12636.

CHAPTER 1

Introduction

The gold standard for estimating the causal relationship between an exposure and outcome is to directly manipulate the exposure levels that subjects receive through random assignment. Randomization can lead to unbiased estimates of the treatment effects by balancing comparison groups on both known and unknown confounders. However, it is often unrealistic or unethical to randomize treatment assignment. Interest often lies in estimating causal effects from observational studies where exposure levels are not assigned by the investigator. This creates challenges because the exposed population may systematically differ from the unexposed population on factors related to the outcome being measured, inducing confounding.

Regression adjustment is often used to correct for potential confounding. An alternative method that has become increasingly popular in fields such as economics, social science, policy evaluation, and many others is the propensity score method [Guo and Fraser, 2014]. The propensity score method to estimate causal effects in a non-randomized experiment was first introduced by Rosenbaum and Rubin [1983]. For a binary exposure, the propensity score is the probability that a subject receives exposure given their values of potential confounders. Using the data on exposure status and values of potential confounders, the propensity score can be estimated for both exposed and unexposed subjects. The estimated propensity score is then used to remove bias in estimation of the causal effect by comparing participants with similar propensities to receive exposure but different observed exposure values.

The first propensity score methods were developed for the setting in which there are

only two treatment levels, i.e., exposed versus control. Methods were subsequently extended to categorical, or multiple, treatments, which introduced the term "generalized propensity score" (GPS) [Imbens, 2000]. In the context of a categorical treatment variable, the GPS corresponds to the conditional probability of receiving a particular treatment given a set of confounders. Following the extension to categorical treatments, the GPS was adapted to the setting of continuous exposures via the use of conditional densities [Hirano and Imbens, 2004; Imai and Van Dyk, 2004].

Originally, the GPS for continuous exposures was estimated using Gaussian densities, with adjustment for confounding accomplished through either covariate regression [Hirano and Imbens, 2004] or stratification [Imai and Van Dyk, 2004]. In this setting, the GPS corresponds to the value of the probability density function (pdf) given the covariates. Several recent methods have aimed at increasing flexibility for estimating the GPS by using gradient boosting [Zhu et al., 2015], kernel smoothing [Flores et al., 2012; Kennedy et al., 2017], or ensemble algorithms [Kreif et al., 2015]. Other methods focus on simultaneously incorporating covariate balancing properties while estimating the GPS with a penalized likelihood or empirical likelihood approach [Fong et al., 2018] or constrained optimization on the entropy of weights constructed from the GPS [Tübbicke, 2020; Vegetabile et al., 2020]. All of these methods have maintained the assumption that the exposure is univariate, i.e., a single continuous treatment variable. Methods to accommodate multiple simultaneous treatment exposures have been only briefly mentioned in the literature [Imai and Van Dyk, 2004].

There are many situations in which evaluating the combined effect of multiple simultaneous exposures is critical to answering scientific questions. In medicine, combination therapies, which involve the patient taking several medications simultaneously, have been shown to be effective for treating many health conditions, such as Crohn's disease [Colombel et al., 2010], cancer [Jain, 2001], hypertension [Gradman et al., 2010], and HIV [Perelson et al., 1997]. Typical methods for estimation of the dose response surface for combination treatments requires careful design with repeated randomized experiments in order to esti-

mate the optimal combination of treatment doses [Khuri and Mukhopadhyay, 2010]. When such experimentation is not feasible, researchers may wish to use available data from observational or non-randomized studies to estimate the joint effects. For example, there is currently interest in studying potential combination therapies for COVID-19 using available data from non-randomized studies. However, determining a potentially beneficial dose of several medications may be complicated due to confounding by patient demographic characteristics, comorbities or other factors as well as due to potential interaction effects [Gautret et al., 2020a,b; Sanders et al., 2020; Stebbing et al., 2020].

We develop methods to estimate the causal effects of multiple continuous exposures occurring simultaneously, using data from a study in which treatment levels were not randomly assigned. We develop a general framework for estimating the causal effects of a multivariate exposure of arbitrary dimension, but focus on bivariate exposures in our simulations and motivating example. The primary objective is precision in estimation of the dose-response surface of the average outcome given a particular combination of exposure values. We propose methods for estimating weights using a multivariate generalized propensity score, which we call mvGPS, and use weighted regression to estimate the dose-response surface. Our methods rely on the assumption that the exposure variables have a multivariate normal distribution.

In Chapter 2 we introduce our motivating example, which involves assessing the joint effects of two types of intervention strategies for reducing childhood obesity rates in Los Angeles County. Chapter 3 summarizes the current framework and methods for causal inference with observational data using propensity scores when treatment is binary or categorical. Chapter 4 presents initial GPS methods along with an expanded discussion of recent extensions and adaptions. Chapter 5 develops the method of causal inference with multiple simultaneous continuous exposures using the multivariate generalized propensity score. Chapter 6 presents a simulation study designed to highlight strengths and limitations of the methodology. Chapter 7 applies the proposed methods to our motivating example involving the reduction in childhood obesity rates from macro and micro intervention strategies.

Finally Chapter 8 concludes with a discussion.

CHAPTER 2

Motivating Example

Obesity rates among low-income preschool-aged children in Los Angeles County were consistently higher than the national average for similar aged children in 2003-2009, with about 20% of such children classified as obese (BMI $\geq$ 30 kg/m^2) [PHFE WIC, 2010]. In response, several organizations, including Los Angeles County Department of Public Health, First 5 LA, Nemours, and the Special Supplemental Nutrition Program for Women, Infants and Children (WIC), implemented programs and policies aimed at reducing childhood obesity in the county. The interventions used a wide variety of different approaches and reflected a large investment of resources.

The Early Childhood Obesity Systems Science (ECOSyS) study, funded by National Institute of Health R01 HD072296, sought to evaluate the impact of these programs on childhood obesity prevalence. To this end, ECOSyS collected information on the nature, timing, location and reach of programs implemented in the county in 2003-2016. The research team also developed a method of calculating a "community intervention dose index" that aggregates exposure to childhood obesity interventions over multiple different programs [Wang et al., 2018]. The community intervention dose index is calculated using a multi-step procedure. Each program is coded to location and year of implementation, extent of reach into the target population, and which of nine different intervention strategies it used. The nine strategies are listed in Table 1, presented as part of a group of tables and figures beginning on page 57. The strategies are categorized as "micro" strategies, which target specific individuals, or "macro" strategies, which target a population at large. By aggregating over

the strategies implemented in a particular location during a particular year, strategy-specific as well as total micro and macro intervention dose indices can be calculated. Rather than estimating the joint effect of the nine individual intervention strategies, interest centered around understanding the joint effect of different levels of macro and micro interventions. It is sometimes hypothesized that micro and macro interventions can have synergistic effects; for example, nutrition education targeting individuals might be more effective when combined with programs increasing the availability of healthy foods in retail stores [Wang et al., 2018]. Collapsing to two exposure types also reduced the dimension of the exposures while protecting against low frequency strategies creating sparse high dimensional regions.

We focus on intervention exposures stemming from WIC programs. WIC serves low-income families and has seven agencies within Los Angeles County with approximately 90 clinics. In 2018, WIC served approximately half of all children under age 5 in Los Angeles County. While WIC offers many regular services, primarily food assistance and nutrition education that are uniform from clinic to clinic, WIC agencies can also receive additional funding to implement intervention programs. These programs are implemented non-randomly at clinics due to differences in community needs and other considerations. Our motivating example focuses on WIC intervention programs implemented in 2010-2016, given that a major change in the WIC food package occurred in 2009 that may have altered family behaviors and neighborhood food environments [Hillier et al., 2012; US Department of Agriculture, Food and Nutrition Service, 2014].

A total of 32 WIC intervention programs implemented in Los Angeles County from 2010-2016 were cataloged by the ECOSyS research team. Information about each program was obtained, including how it was implemented, the estimated reach in terms of client participation, which clinics participated, and how long the program was active. These exposures were mapped to census tracts where WIC participating children live by identifying the implementing WIC clinics and then defining a catchment area around the clinic intended to capture the exposed population. Catchment areas were defined for each clinic using records

of client attendance, and varied by strategy type (macro versus micro). Children living in census tracts that fell within a catchment area were potentially exposed. Exposure values at census tracts were then aggregated across programs by strategy and by year to obtain a single continuous dose for each of the nine intervention strategies. Strategy-specific doses were then summed into macro and micro intervention doses, which were log transformed due to skewness. For additional details about the process of constructing the continuous doses, see Appendix A.

Figure 1 on page 63 shows the resultant joint distribution of macro and micro intervention dose for the WIC intervention programs averaged over our defined intervention period, 2010-2016. Each point in the figure represents a census tract. WIC-participating children residing in a particular census tract were presumed to receive the calculated doses.

The outcome of interest was change in census tract-level childhood obesity prevalence. Childhood obesity prevalence was measured using administrative records from children ages 2-5 years who participated in WIC in Los Angeles County during 2007-2016, compiled by the WIC Data Mining Project, see https://lawicdata.org for more details. From these records, obesity prevalence by census tract and year was constructed for census tracts with at least 30 WIC-enrolled children. Census tracts used in the analysis were restricted to 8 regions within Los Angeles County that were targeted as part of the ECOSyS data collection effort. This resulted in a total of $n = 1079$ census tracts which serve as the units of analysis shown in Figure 2 on page 64. The outcome of interest, Y, was the difference in average obesity prevalence between post, 2012-2016, and pre, 2007-2009, intervention, i.e., $Y = \bar{p}_{post} - \bar{p}_{pre}$. The post intervention period was taken to start in 2012 rather than 2010 to account for potential lag in treatment effects.

We aimed to estimate the dose-response surface of Y, change in childhood obesity prevalence, associated with combinations of macro and micro intervention exposure doses, after removing bias due to non-random assignment of programs. Understanding the simultaneous effect of macro and micro intervention strategy exposures is important to help policy makers

make decisions about the allocation of scarce resources to various intervention strategies [Rosenkranz and Dzewaltowski, 2008].

CHAPTER 3

Background

3.1 Potential Outcomes Framework

In Chapter 1, the definition of treatment was left explicitly broad, but it is important to distinguish between conditions that are treatments, which allow for causal effect estimation, and those which are attributes. The distinction is formalized by Holland [1986], where attributes are defined as properties or characteristics that do not have the potential for exposure for each unit in the population, while treatments have this potential for exposure.

For variables like gender or race, it is not clear how units would be potentially exposed to different values. For example when testing the effect of gender on the probability of being promoted, would the gender of an applicant be changed from male to female on a resume, would two genetically identical clients who differ only in terms of gender be presented at an interview, or would this aim to test some other type of gender related difference? In Greiner and Rubin [2011], the authors propose an argument for re-framing causal questions of immutable traits to the *perception* of the trait rather than the actual trait itself. In this way investigators may construct well defined causal quantities and effects to answer questions such as whether a victim's race affects the jury's decision to impose the death penalty or life imprisonment when no member of a victim's family takes the witness stand. Clearly this type of causal inquiry for attributes like gender or race are not straightforward to define or test. On the other hand, treatments are typically thought of in the setting of classic scientific experiments that systematically manipulate treatment assignment and observe its

effect while maintaining all other variables constant to test a hypothesis of interest. For instance, investigators can randomize which mice receive a new drug to test for efficacy in an animal model. By using randomization to assign treatment status, each mouse has the potential to receive the test drug and therefore any differences observed are due to the drug itself.

Recently in Pearl and Mackenzie [2018], the authors have re-framed the argument of defining causal questions within a structure referred to as the ladder of causation. At the lowest rung are relationships that rely on seeing and observing that can be used to determining if two events are associated. Examples at this rung include questions like: "How likely is the pitcher to throw a fastball in an 0-2 count?" The middle rung the questions shift from *observing* to *doing* where some type of intervention must take place rather than passively collecting data. For instance if we ask: "How many more walks would a batter have if they never swung at the first pitch?" The final rung in the ladder requires imagining as we aim to compare the observed world to a fictitious alternative world. In this case we may ask: "If the batter had not swung at the first pitch, what is the probability they would have walked in that at bat?" While Holland made a clear distinction based on the potential of a unit to be exposed, Pearl and Mackenzie take this further by separating interventions and hypothetical experiments that aim to answer "what-if" questions. Both interventions and these hypothetical experiments must have units with the potential for exposure, but it is the latter which rely on causal methodology to frame and model that which is not observable. While advances in data mining and deep learning can more accurately answer questions in rung one and two, only causal models are able to encode knowledge about the structural relationship and maintain this ability when transferred to novel environments. The authors talk of the age of causal revolution beginning as we aim to move towards answering the questions of why events happen, but we must first begin by returning to Holland's initial distinction and clearly defining what it means to have the potential for exposure.

The idea of potential exposure is crucial in the Neyman-Rubin causal model. Neyman

[1923] is credited as the first author to introduce the potential outcome notation, albeit in the context of a completely randomized experiment for agricultural studies in his Ph.D. thesis. For a population of units, $k = 1, \ldots, m$, and treatments, $i = 1, \ldots, \nu$, Neyman used the double index notation to denote the potential crop yield for the k^{th} plot exposed to the i^{th} treatment as U_{ik}. By using this formulation, Neyman laid the groundwork for a novel probabilistic model that would be generalizable beyond the simple case of complete randomization. Recently, the double index notation for potential outcomes has been usurped by defining the potential outcome using parentheses, i.e., $Y_k(i)$ would represent the i^{th} potential outcome for unit k. Bridging the gap to the nonrandomized experiment, Rubin [1974] showed that while randomized experiments were superior for estimating causal effects, it was possible to use nonrandomized data to estimate causal effects. Holland [1986] discusses the philosophical implications for this combined framework which lead to the term Neyman-Rubin causal model, or sometimes simply the Rubin causal model.

We first introduce the Neyman-Rubin causal model in the case of a binary treatment. For a binary treatment, each unit has two potential outcomes, but only one potential outcome is observable. If unit i receives treatment then the outcome $Y_i(1)$ is observed, and $Y_i(0)$ is unobserved. Likewise, if unit i instead received no treatment then the outcome $Y_i(0)$ is observed and $Y_i(1)$ is unobserved. We can define the observed outcome Y_i using an indicator variable D_i where $D_i = 1$ if unit i received treatment and 0 otherwise,

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0). \tag{3.1.1}$$

In any setting we only observe the values Y_i and D_i. This inability to simultaneously observe both potential outcomes, $Y_i(0)$ and $Y_i(1)$, for unit i is what Holland deemed "the fundamental problem of causal inference".

To address this problem, we can first think of a randomized experiment. In a randomized experiment, treatment assignment is independent of the potential outcomes, i.e., $\{(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i\}$. Hence, the expected potential outcomes are the same regardless of

treatment assignment,

$$\mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i(0) \mid D_i = 1] \qquad \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i(1) \mid D_i = 0]. \tag{3.1.2}$$

This independence of potential outcomes and treatment assignment lays the framework for estimating causal effects. One causal outcome of interest is the population average treatment effect (PATE), defined as

$$\begin{aligned} \text{PATE} = \; & \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)], \\ & = \big(\pi\, \mathbb{E}[Y_i(1) \mid D_i = 1] + (1-\pi)\, \mathbb{E}[Y_i(1) \mid D_i = 0]\big) \\ & \quad - \big((1-\pi)\, \mathbb{E}[Y_i(0) \mid D_i = 0] + \pi\, \mathbb{E}[Y_i(0) \mid D_i = 1]\big), \\ & = \pi\big(\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1]\big) \\ & \quad + (1-\pi)\big(\mathbb{E}[Y_i(1) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 0]\big), \end{aligned} \tag{3.1.3}$$

where π is the proportion of the population assigned to the treatment group, i.e. $\Pr(D = 1)$. This quantity answers the question of what the difference in the outcome would be if the entire population received exposure versus no one receiving exposure. An alternative commonly used causal outcome is the population average treatment effect on the treated (PATT), defined as

$$\text{PATT} = \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1], \tag{3.1.4}$$

which requires only the first part of Equation 3.1.2 to hold as discussed in Guo and Fraser [2014]. The PATT summarizes the effect of taking away treatment from those who received it. The final potential quantity of interest is the population average treatment effect on the controls (PATC), defined as

$$\text{PATC} = \mathbb{E}[Y_i(1) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 0]. \tag{3.1.5}$$

The PATC described the effect of adding treatment for subjects who did not receive it. From Equation 3.1.3 we can see that the PATE is simply a weighted sum of the PATT and PATC.

Note that these quantities will sometimes be referred to as SATE, SATT, or SATC where "S" represent sample as they are interpreted conditional on the sample data. We shall focus on estimands for PATE, but similar results can be shown for PATT or PATC.

As a first choice in estimating PATE, define the average treatment effect, τ, as the difference between the average outcome of those who received treatment and those who did not,

$$\tau = \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]. \tag{3.1.6}$$

Independence of the exposure and outcome is a sufficient condition for τ to provide a consistent estimate of the PATE using Equation 3.1.2:

$$\begin{aligned}
\text{PATE} = \; & \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
& \big(\pi\, \mathbb{E}[Y_i(1) \mid D_i = 1] + (1-\pi)\, \mathbb{E}[Y_i(1) \mid D_i = 0]\big) \\
& - \big((1-\pi)\, \mathbb{E}[Y_i(0) \mid D_i = 0] + \pi\, \mathbb{E}[Y_i(0) \mid D_i = 1]\big), \\
= \; & \big(\pi\, \mathbb{E}[Y_i(1) \mid D_i = 1] + (1-\pi)\, \mathbb{E}[Y_i(1) \mid D_i = 1]\big) \\
& - \big((1-\pi)\, \mathbb{E}[Y_i(0) \mid D_i = 0] + \pi\, \mathbb{E}[Y_i(0) \mid D_i = 0]\big), \\
= \; & \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
= \; & \tau.
\end{aligned}$$

The sample average treatment effect (SATE),

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n_1} D_i y_i + \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - D_i) y_i, \tag{3.1.7}$$

is therefore a consistent and unbiased estimator of the PATE when independence holds. However, in nonrandomized experiments, Equation 3.1.2 is invalid. In the absence of randomization, treatment assignment may be associated with confounders for the outcome leading to inferential issues when interpreting the treatment effect. Is the observed difference between the treatment group due to the treatment alone or is this effect clouded by a confounding variable associated with treatment status and the outcome? To remedy this situation, we can use propensity score methods.

3.2 Propensity Score Methodology

In order to estimate the treatment effect in situations where treatment assignment is non-randomized, an investigator needs to address potential confounders that are associated with the treatment assignment and the outcome of interest. One solution is the propensity score methodology developed by Rosenbaum and Rubin [1983], which outlined how to estimate causal effects in the setting of a binary treatment assignment. The key concept underlying this method is to remove the bias of confounding variables using a single scalar value, the propensity score. The propensity score is used to remove this bias in a variety of ways, which include covariate adjustment, inverse probability weighting, subclassification, and matching. Each of these methods aims to compare groups within the population that have similar probability of receiving the treatment, but differ on their observed treatment assignment.

3.2.1 Binary Treatment Setting

The initial methodology of Rosenbaum and Rubin [1983] focused on binary treatments. Similar to the randomized setting described above, each unit, i, in the population has potential outcomes $Y_i(0)$ and $Y_i(1)$ corresponding to each treatment level $D_i \in \{0,\ 1\} = \mathcal{D}$. We further assume in the nonrandomized setting that we have a vector of covariates, $\mathbf{X}_i$, that are associated with the potential outcomes and the treatment, making them confounders. This confounding means that the direct estimate of τ using the SATE from Equation 3.1.7 is biased since the groups are not directly comparable due to imbalance in the confounders. To account for differences in confounders between treatment groups, Rosenbaum and Rubin's method presented three key pillars for bias removal using the propensity score: 1) stable unit treatment value assumption (SUTVA), 2) balancing score, and 3) strong ignorability.

The first pillar is an assumption implicit in the original propensity score method of Rosenbaum and Rubin. It assumes that the potential outcome of unit i under treatment d does not depend on the treatment given to unit i', where $i \neq i'$, and that there exists only

one version of each exposure. SUTVA posits that there must be a unique response, $Y_i(d)$, for unit i to treatment d [Rubin, 1980, 1986]. In other words, we assume that the potential outcome of the i^{th} unit depends only on the treatment assignment that it received, and is thus independent of the other treatment assignments. While this assumption may seem intuitive, we can think of examples where it is violated. An epidemiological example is the spread of contagious diseases, where the probability that you become infected may depend on the infection and immunity status of people in your immediate proximity. SUTVA is untestable due to the "fundamental problem of causal inference" since each sample is restricted to the observed ensemble of treatments and we cannot observe different potential outcomes under different treatment assignments.

The second key pillar of Rosenbaum and Rubin is the balancing score. The balancing score, $b(\mathbf{X})$, is a function of the observed covariates, $\mathbf{X}$, such that the conditional distribution of $\mathbf{X}$ given $b(\mathbf{X})$ is independent of treatment assignment, D,

$$\mathbf{X} \perp\!\!\!\perp D \mid b(\mathbf{X}). \tag{3.2.1}$$

Stated another way, within each level of the balancing score the treated and untreated groups have identical covariate distributions. A trivial example of a balancing score is $b(\mathbf{X}) = \mathbf{X}$, since $f\Big((D, \mathbf{X}) \mid \mathbf{X}\Big) = f(D)$. In the case where $\mathbf{X}$ includes only two categorical variables with R and C categories respectively, we would be conditioning on each cell in the $R \times C$ table and looking for the difference between individuals in that cell who were in the treated and untreated group. Even in this trivial example, with moderately sized values of R and C, say 6 each, we might expect certain cells in the 6×6 covariate space to be sparse and not contain both a treated and untreated individual. As the number of covariates becomes large, finding identical covariate patterns between treatment groups becomes difficult and makes conditioning on $\mathbf{X}$ impractical. The goal then is to find the coarsest balancing score that maps the high-dimensional covariates to a low-dimensional balancing score. The propensity score is such a balancing score. Rosenbaum and Rubin defined the propensity score as the conditional probability of receiving treatment given a set of potential confounding variables,

denoted by

$$e(\mathbf{X}) = \Pr(D = 1 \mid \mathbf{X}), \tag{3.2.2}$$

which was proved to be a balancing score in Cochran and Rubin [1973] for multivariate normal covariates $\mathbf{X}$. Traditional statistic methods for estimating probabilities for a binary outcome such as logit or probit regression were originally suggested to estimate the propensity score.

The third pillar outlined by Rosenbaum and Rubin is the strong ignorability assumption. Treatment assignment is strongly ignorable given a vector of covariates $\mathbf{X}$ if

$$\Big(Y(1), Y(0)\Big) \perp\!\!\!\perp D \mid \mathbf{X}, \text{ and } 0 < \Pr(D = 1 \mid \mathbf{X}) < 1. \tag{3.2.3}$$

for all $\mathbf{X}$. The second condition of Equation 3.2.3 is sometimes presented as a separate assumption called the positivity assumption, meaning that all units have the potential to receive treatment given the covariates $\mathbf{X}$. We can think of this assumption as saying that the observed covariates, $\mathbf{X}$, contain all the information about the potential confounding between the treatment and outcome so that by conditioning on these covariates we are in a situation analogous to a randomized experiment. In particular, this condition allows us to compare outcomes between individuals with identical observed covariates but different observed treatment assignment in an unbiased fashion.

However as mentioned earlier, as the number of covariates becomes large, it becomes difficult to find similar covariate patterns between treatment groups. The key insight was to recognize that instead of conditioning on the observed covariates, it was possible to condition on the propensity score. Using the three key pillars outlined above, Rosenbaum and Rubin demonstrated that if Equation 3.2.3 is true for the observed covariates $\mathbf{X}$, then treatment is also strongly ignorable given the propensity score, i.e.,

$$\Big(Y(1), Y(0)\Big) \perp\!\!\!\perp D \mid \mathbf{X} \implies \Big(Y(1), Y(0)\Big) \perp\!\!\!\perp D \mid e(\mathbf{X}). \tag{3.2.4}$$

Given strong ignorability of the propensity score, we have that

$$\mathbb{E}[Y_i(1) \mid D_i = 1, e(\mathbf{X}_i)] - \mathbb{E}[Y_i(0) \mid D_i = 0, e(\mathbf{X}_i)] = \mathbb{E}[Y_i(1) \mid e(\mathbf{X}_i)] - \mathbb{E}[Y_i(0) \mid e(\mathbf{X}_i)]. \tag{3.2.5}$$

Taking the expectation with respect to the distribution of $e(X)$, we have

$$\mathbb{E}\Big[\mathbb{E}[Y_i(1) \mid e(\mathbf{X}_i)] - \mathbb{E}[Y_i(0) \mid e(\mathbf{X}_i)]\Big] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)], \tag{3.2.6}$$

which is our PATE. By averaging outcomes from individuals who have similar propensity scores but different observed treatment value, and then averaging over the propensity scores, an unbiased estimate of the population average treatment effect is obtained. In practice the distribution of $e(\mathbf{X})$ is unknown and must be estimated.

Rosenbaum and Rubin proposed three methods for bias removal using the propensity score: 1) matching, 2) subclassification, and 3) covariate adjustment. Matching occurs in a two-step process. Treated units are first matched to untreated units using the value of the propensity score. Then, the expected difference between each matched pair is computed and averaged across all pairs to return an estimate equal to the population average treatment effect. There are many types of matching methods. Dehejia and Wahba [2002] compare the use of caliper, 1:1, or 1:K matching strategies. An alternative to matching is subclassification which instead of matching propensity scores exactly between treated and control units, breaks up the observed range of propensity scores into strata with at least one observations from each treatment group within a stratum. Within each stratum, the expected difference again equals the average treatment effect, and using the weighted average of these differences, where weights are defined based on the number of observations within each stratum, we obtain an unbiased estimate of the PATE. Subclassification was shown to produce unbiased estimates in Rosenbaum and Rubin [1984]. Finally, covariate adjustment incorporates the propensity score as a covariate in the outcome analysis with the assumption that the conditional expectation of the potential outcome given the propensity score is linear. For

example, we can define the conditional expectation as

$$\mathbb{E}[Y \mid D = d, e(\mathbf{X})] = \beta_0 + \beta_1 D + \beta_2 e(\mathbf{X}),$$

for $d = 0, 1$, with the estimate $\hat{\beta}_1$ used as an estimate for the PATE. While this result holds in theory, Hade and Lu [2014] showed that the covariate adjustment method is substantially biased when the true relationship between outcome and propensity score is nonlinear and that this bias is potentially present even if linearity holds when using estimated values for the propensity score. Hade and Lu suggest propensity score matching and subclassification as a robust alternative to covariate adjustment.

Another popular method for adjusting using the propensity score involve weighting [Hirano and Imbens, 2001; Hirano et al., 2003; Robins et al., 2000; Rosenbaum, 1987]. These methods are referred to as inverse probability weighting (IPW), inverse probability of treatment weighting (IPTW), or weighted regression. The idea behind the weighted approach is to re-weight the treated and control observations using the propensity score so that the weighted values are representative of the population. The approach is similar to the method of Horvitz and Thompson [1952] for survey sampling weights. First we note that conditional on $\mathbf{X} = \mathbf{x}$ we have that

$$\mathbb{E}\left[\frac{Y_i \times D_i}{e(\mathbf{X}_i)} \mid \mathbf{X}_i = \mathbf{x}_i\right] = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}_i].$$

Taking the iterated expected value with respect to $\mathbf{X}$, we have that the weighted expression is equal to $\mathbb{E}[Y_i(1)]$. This means that an unbiased estimator for the PATE with binary exposures is given by

$$\mathbb{E}\left[\frac{Y_i \times D_i}{e(\mathbf{X}_i)} - \frac{Y_i \times (1 - D_i)}{1 - e(\mathbf{X}_i)}\right] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)].$$

The same result can be obtained using weighted least squares estimation with a regression function

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i,$$

where weights are of the form

$$w(D, \mathbf{X}) = \frac{D}{e(\mathbf{X})} + \frac{1 - D}{1 - e(\mathbf{X})}.$$

As the true propensity scores is generally unknown, the values $e(\mathbf{X})$ are replaced with estimated values $\hat{e}(\mathbf{X})$. Again this method gives us the estimate of PATE via $\hat{\beta}_1$.

All of these various methods allow the researcher to incorporate the propensity score as a means for obtaining unbiased estimates of the true average treatment effect. When the treatment of interest is no longer binary, complications arise. We discuss such treatments next.

3.2.2 Discrete-Valued Treatment Setting

While the propensity score method was originally proposed to handle only binary exposures, it was quickly extended to handle discrete-valued treatments taking more than two values [Imbens, 2000; Lechner, 2001]. There are two types of discrete-valued treatments, nominal and ordinal. Nominal, also referred to as categorical, treatments have no intrinsic ordering. An example would be a three-arm study testing surgery, drug treatment, or neither. Unlike nominal treatments, ordinal treatments have an implicit ordering, such as the dose of a drug. In both cases, we can represent the treatments using integer values from a set of discrete values between 0 and K, i.e. $\mathcal{D} = \{0, 1, \ldots, K\}$.

To handle multi-valued treatments, Imbens first introduced the term generalized propensity score as a generalization of the bivariate propensity score. To avoid confusion between the generalized propensity score for discrete-valued treatments and the generalized propensity score for continuous treatments discussed in later chapters, we will refer to the discrete-valued version as the discrete propensity score and reserve the term generalized propensity score to refer only to the case with continuous exposure. Imbens defined the discrete propensity score as the conditional probability of receiving a particular level of treatment given the

pre-treatment variables, which we can express as

$$r(d, \mathbf{x}) = \Pr(D = d \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[\mathbb{1}_d \mid \mathbf{X} = \mathbf{x}], \tag{3.2.7}$$

where $\mathbb{1}_d$ is an indicator of receiving treatment level d,

$$\mathbb{1}_d = \begin{cases} 1 \text{ if } D_i = d \\ 0 \text{ otherwise.} \end{cases}$$

In order to use the discrete propensity score, Imbens altered the strong ignorability assumption in Equation 3.2.3 to a weaker version that relies on independence of the marginal potential outcomes rather than the joint distribution of potential outcomes. Assignment to treatment D is weakly ignorable, given pre-treatment variables X, if

$$Y(d) \perp\!\!\!\perp \mathbb{1}_d \mid \mathbf{X} \; \forall \; d \in \mathcal{D}. \tag{3.2.8}$$

Rather than requiring the treatment to be independent of the entire set of potential outcomes, weak ignorability requires only pairwise independence of treatment with each potential outcome. Imbens notes that weak ignorability restricts independence to the 'local' treatment level of interest, and in this way it is similar to the definition of 'missing at random' [Little and Rubin, 2014; Rubin, 1976].

It is important to pause and think about the potential outcomes, $Y(d)$, in the discrete-valued setting. In Section 3.2.1, there were only two potential outcomes representing the treated and non-treated groups, $Y_i(1)$ and $Y_i(0)$. Inference focused on the estimation of the PATE, τ, by using the propensity score to account for the differences in confounders. With discrete-valued treatments, the number of potential outcomes is now dependent on the set $\mathcal{D}$ and the number of possible pairwise comparisons is equal to $\binom{K+1}{2}$. Depending on the application, causal estimands of interest could be comparing the average difference between two treatment levels, $\mathbb{E}[Y_i(k) - Y_i(k')]$ for $k \neq k' \in \mathcal{D}$, or between other contrasts of treatment combinations,

$$\mathbb{E}[\mathbf{c}^T \mathbf{Y}_i(d)],$$

where $\sum_{k=1}^{K} c_k = 0$ and $\mathbf{Y}_i(d) = \Big(Y_i(1), Y_i(2), \ldots, Y_i(K)\Big)^T$. The investigator must think about the scientific question being studied, and construct the appropriate causal estimand to answer the research hypothesis.

Assuming that weak ignorability holds, then by the same argument as the binary treatment setting, we have that treatment assignment is weakly ignorable given the discrete propensity score: $\mathbb{1}_d \perp\!\!\!\perp Y(d) \mid r(d, \mathbf{X})$, for all $d \in \mathcal{D}$. With this result, we can use the discrete propensity score in place of $\mathbf{X}$ to estimate the conditional expectation of the potential outcome. Similar to the covariate balancing approach proposed by Rosenbaum and Rubin [1983], we can define the conditional expectation of the outcome as a function of the treatment and the discrete propensity score, $\beta(d, r)$. The average potential outcome for treatment level d, $\mathbb{E}[Y(d)]$, can then be found by averaging over the discrete propensity score. This means that if treatment assignment is weakly ignorable given X, then for all $d \in \mathcal{D}$,

(i) $\beta(d, r) = \mathbb{E}[Y(d) \mid r(d, \mathbf{X}) = r] = \mathbb{E}[Y \mid \mathbf{D} = \mathbf{d}, r(D, \mathbf{X}) = r]$,

(ii) $\mathbb{E}[Y(d)] = \mathbb{E}[\beta(d, r(d, \mathbf{X}))]$.

Note that in (ii), the expectation is taken by averaging the discrete propensity score at treatment level d rather than at the observed treatment level D.

Implementing the propensity score methodology for discrete-valued treatment via the discrete propensity score consists of three steps. In the first step, the discrete propensity score $r(d, \mathbf{x})$ is estimated. With a binary treatment, this is typically done using logistic regression. With discrete-valued treatments, there are different approaches depending on the type of treatment. For nominal treatments, a discrete choice model such as a multinomial or nested logit model could be used. These two types of models differ with respect to how changing one choice affects the alternative choices. In a multinomial model we assume independence of irrelevant alternatives (IIA), meaning that adding or deleting alternative outcomes does not affect the odds among the remaining outcomes [McFadden, 1973]. In a nested logit model,

this independence assumption is relaxed by creating groups of nested alternatives [McFadden, 1978]. By using a nested structure, choices in the same nest are treated as independent, but choices between nests are correlated. If the outcome is an ordinal treatment, an ordinal logistic regression could be used via a cumulative link model that assumes proportional odds between the respective levels. The second step is to estimate the conditional expectation $\beta(d, r)$. A variety of different methods may be used to model this conditional expectation depending on the outcome and the level of smoothness with respect to d. Finally, in the last step the average response at treatment level d is estimated as the average of the estimated conditional expectation, $\hat{\beta}(d, r(d, \mathbf{X}))$, averaged over the distribution of the pre-treatment variables. Again, it is important to mention that the conditional expectation $\beta(d, r)$ is evaluated at the treatment level of interest $r(d, \mathbf{X}_i)$, not at the observed level of treatment $r(D_i, \mathbf{X}_i)$.

This framework allows the researcher to estimate causal effects for different levels of the treatment. The estimation requires careful consideration of the sub-populations defined by the conditioning sets $r(d, \mathbf{X})$. This work was critical for extending the propensity score methodology beyond the case of binary treatments, and ultimately to the generalized propensity score with continuous treatment.

CHAPTER 4

Univariate Generalized Propensity Score

With the foundation from both the binary and nominal/ordinal exposures, the next logical step is to address continuous treatments. Following Hirano and Imbens [2004] we define the generalized propensity score (GPS) as the conditional probability of receiving exposure given a set of potential confounders, which is a generalization of the binary and categorical propensity scores. The major difference between these earlier methods for non-continuous exposures is the form of the conditional distribution of exposure, which must be adapted to handle exposures along a continuous interval rather than at discrete values within the domain. Likewise, the inferential target with continuous exposures differs from binary and categorical exposures, for which the aim is typically to compare pairwise values of the exposure; for continuous exposures, interest is in estimating the shape of the dose-response function along the entire domain of exposure.

In the literature, methodological work for the GPS has focused on several key components, including the choice of conditional distribution for estimation of the probability of receiving a given exposure level, evaluating the balancing property of the estimated GPS, and the manner in which the resultant GPS is used to remove bias in estimation of the dose-response function. Several recent methods for the GPS have focused on relaxing the parametric assumptions initially proposed for the GPS method such as using the SuperLearner ensemble algorithm [Kreif et al., 2015], kernel density estimation [Flores et al., 2012], and boosting algorithm [Zhu et al., 2015]. Alternatively, other proposed methods have focused on ensuring that the balancing property of the GPS is enforced either via penalization during GPS esti-

mation [Fong et al., 2018] or constrained optimization on the entropy of weights constructed from the GPS [Tübbicke, 2020; Vegetabile et al., 2020]. Finally, many different mechanisms for bias removal have also been proposed such as covariate adjustment [Hirano and Imbens, 2004], subclassification [Imai and Van Dyk, 2004], and inverse probability weighting [Robins et al., 2000]. It is therefore imperative as an analyst to evaluate the selection of each component when performing a causal analysis of continuous exposures using the GPS.

The GPS methods that are currently available have been used in a variety of applied settings. For instance, investigators have used GPS methods in the economics literature to estimate the expected difference in employment outcomes between those who spend different lengths of time in a job training program [Flores et al., 2012; Kluve et al., 2012] or to understand the effect of winning the lottery on subsequent labor earnings [Hirano and Imbens, 2004]. Other investigators have focused on using the GPS to quantify health related outcomes such as the relationship between the duration of breastfeeding and childhood obesity [Jiang and Foster, 2013], the effect of mothers' overall weight concern on daughters' dieting behavior [Zhu et al., 2015], or on the relationship between smoking intensity and medical expenditures [Imai and Van Dyk, 2004]. Another popular area of research that utilizes GPS methods includes educational policy and evaluation such as studying the effects of the number of credits taken on the transfer rate of students from community college to four-year institutions [Doyle, 2011]. The ability to utilize continuous exposure values rather than forcing investigators to discretize exposure makes the GPS a popular and powerful tool for causal inference.

While there have been many improvements to the GPS beyond the initial framework proposed by Hirano and Imbens [2004] and Imai and Van Dyk [2004], there are still several areas that remain unanswered, most notably the lack of methods to handle multiple simultaneous continuous exposures. All of these examples and methods mentioned have been focused on a single univariate continuous exposure. Multiple simultaneous exposures are noted only briefly in the development by Imai and Van Dyk [2004].

In this Chapter I will introduce the foundation for causal inference with continuous exposures and the relevant notation for a single continuous exposure in Section 4.1, highlight the initial parametric specifications of the GPS in Section 4.2, discuss recent extensions in Section 4.3, and conclude with a note on key limitations that motivate the new method proposed in Section 4.4.

4.1 Framework for Causal Inference with Continuous Exposures

Suppose we have a random sample of units, indexed by $i = 1, \ldots, N$. Each unit i has a set of potential outcomes, $Y_i(d)$, for exposure level $d \in \mathcal{D}$. In the binary case $\mathcal{D} = \{0,\ 1\}$ and for categorical exposures $\mathcal{D} = \{0, \ldots, K\}$. For the continuous case, $\mathcal{D}$ is assumed to be a continuous interval, $\mathcal{D} = [d_0,\ d_1]$. We define $Y_i(d)$ as the unit-level dose-response which is equal to the potential outcome for the i^{th} subject for exposure level d. For each unit i, we observe a p dimensional vector of covariates $\mathbf{X}_i$, the level of treatment received $D_i \in [d_0,\ d_1]$, and the observed outcome Y_i. It is important to note that we can equate the observed outcome with the corresponding unit-level dose response when SUTVA holds, as discussed in Section 3.2.1 for binary exposures, i.e., $Y_i = Y_i(d)$. Interest lies in the average dose-response function, $\mu(d) = \mathbb{E}[Y_i(d)]$. It is assumed that $\{Y_i(d)\}_{d \in \mathcal{D}}$, D_i, and $\mathbf{X}_i$ are defined on a common probability space, that D_i is continuously distributed with respect to Lebesgue measure on $\mathcal{D}$, and that $Y_i = Y_i(D_i)$ is a well defined random variable. Compared to binary and categorical treatments, one of the key differences with a continuous treatment domain is that the treatment effect is no longer an unknown scalar parameter or discrete set of parameters, but rather an unknown function, $\mu(d)$.

4.2 Parametric Specification

Methods to estimate the dose-response function for continuous treatments were first developed by Hirano and Imbens [2004] and Imai and Van Dyk [2004]. Both methods focused on applying parametric methods to estimate the GPS. Specifying the conditional distribution for the GPS using a parametric family of distributions was the first logical step as it allowed for the subsequent results to utilize the established parametric theory often with closed form estimation and low computational burden. While both methods shared similar assumptions about the form of the conditional densities, they advocated for different mechanisms of bias removal with respect to the GPS. In Hirano and Imbens [2004] a covariate adjustment strategy was used where the dose-response function was modeled as a flexible function of the treatment and GPS in a linear model, while in Imai and Van Dyk [2004] a subclassification method was proposed. The notation used in this section will follow that of Hirano and Imbens [2004].

Similar to the discrete-valued setting described in Section 3.2.2, adapting the propensity score methodology to a continuous setting focused on generalizations of the three key pillars of Rosenbaum and Rubin [1983]. The first pillar is strong ignorability, see Equation 3.2.3. In the case of a continuous treatment, strong ignorability places too much of a burden on the potential outcomes. If strong ignorability were maintained, it would require joint independence of all potential outcomes $\{Y(d)\}_{d\in[d_0,d_1]}$. Instead, a weaker version is adopted analogous to the weak ignorability with discrete-valued treatments where conditional independence is required at each value of the treatment rather than joint independence. We can write this form of weak ignorability as

$$Y(d) \perp\!\!\!\perp D \mid \mathbf{X} \; \forall \; d \in \mathcal{D}. \tag{4.2.1}$$

As noted by Imai and Van Dyk [2004], similar to the binary case, this result is difficult to implement directly in practice using the observed pre-treatment covariates because as the dimension of $\mathbf{X}$ increases and/or there are several continuous pre-treatment covariates,

matching and subclassification become impossible. To remedy this issue, Hirano and Imbens define the function $r(d, \mathbf{x})$, here-to referred to as the propensity function, as the conditional density of the treatment given the covariates:

$$r(d, \mathbf{x}) = f_{D|\mathbf{X}}(d \mid \mathbf{x}), \tag{4.2.2}$$

where $f(\cdot \mid \mathbf{x})$ parameterizes the distribution. The value of the propensity function for the observed treatment is called the generalized propensity score (GPS),

$$R = r(D, \mathbf{X}) = f_{D|\mathbf{X}}(D \mid \mathbf{X}). \tag{4.2.3}$$

The GPS is aptly named, as we can think of the propensity score with binary treatments from Section 3.2.1 and the discrete propensity score from Section 3.2.2 as special cases of the GPS. In Equation 3.2.2 the conditional distribution $f(\cdot \mid \mathbf{x})$ is the binomial distribution, which uses the logit$(\cdot)$ link to model the probability of receiving treatment as a linear equation of the pre-treatment covariates. Likewise in Equation 3.2.7, the parametric density for the discrete propensity score depends on the type of discrete-valued treatment, nominal or ordinal.

To use the GPS for causal inference, we need weak ignorability of the treatment assignment conditional on the propensity function, $r(d, \mathbf{X})$, rather than the distribution of pre-treatment covariates, $\mathbf{X}$. To do this we assume that Equation 4.2.1 is true, and we want to show that for every value of d,

$$f_D(d \mid r(d, X), Y(d)) = f_D(d \mid r(d, X)). \tag{4.2.4}$$

Hirano and Imbens [2004] prove this result using iterated expectation, and the fact that the propensity function is measurable with respect to the sigma-algebra generated by $\mathbf{X}$, implying that $f_D(d \mid \mathbf{X}, r(d, \mathbf{X})) = f_D(d \mid \mathbf{X})$).

While both Hirano and Imbens and Imai and Van Dyk had similar definitions of the GPS, one important difference between the methods is that Imai and Van Dyk further suppose that the propensity function can be uniquely parameterized. Assuming unique parameterization

means that the propensity function depends on $\mathbf{X}$ only through the parameter $\theta(\mathbf{X}) \in \Theta$, so the propensity function can be rewritten as $r(d, \theta) = f_{D|\theta}(d \mid \theta)$. Imai and Van Dyk use this result in order to match or subclassify on θ or any one-to-one function of θ. For a Gaussian propensity function, a natural choice would be $\theta(\mathbf{X}) = \mathbf{X}^T\beta$.

As mentioned in Rosenbaum and Rubin [1983], the second key properties of the propensity score is the balancing property, such that the distribution of the pre-treatment covariates is independent of treatment given the propensity score. This must also hold with the propensity function for continuous exposures although it becomes more difficult to assess directly. Conditioning on the value of $r(d, \mathbf{X})$, the probability that a unit received exposure level $D = d$ is independent of their pre-treatment covariates. We can write this as

$$\mathbb{1}_{D=d} \perp\!\!\!\perp \mathbf{X} \mid r(d, \mathbf{X}). \tag{4.2.5}$$

This is a testable result, and is an important step for investigators to check when using GPS methods. Hirano and Imbens [2004] propose testing this balancing property using t-statistics with a blocking strategy based on evaluating the propensity function at the median value of treatment tertiles. Imai and Van Dyk [2004] suggest constructing a linear model that predicts the effect of the observed treatment D on each covariate while controlling for the estimated propensity function. Balance can then be assessed using t-statistics for the coefficient of D in the model, although the authors note that this linear model may not detect all deviations from independence. Other methods for checking balance have been proposed such as using covariate-exposure correlation values as proposed by Zhu et al. [2015] with a guideline of absolute correlations less than 0.1 for sufficient balance. In a recent comparison of different methods for assessing covariate balance for continuous exposures, Austin [2019] suggest using covariate-exposure correlations. Note that these correlation statistics typically aim to balance on first order moments alone, but methods for balancing on higher order moments have been proposed by Fong et al. [2018]; Vegetabile et al. [2020]. In the methodology proposed in Chapter 5 we use the first order correlation coefficients as our balance assessment metric.

The final key pillar of the propensity score from Rosenbaum and Rubin [1983] to check in our development for continuous exposures is SUTVA. To maintain identifiability we assume that SUTVA holds, meaning that the potential outcomes $Y_i(d)$ for the i^{th} unit are uniquely different without multiple versions of the exposure or interference between units.

In addition to adapted versions of the three key pillars for use of propensity scores proposed by Rosenbaum and Rubin [1983], there are additional properties that need to be assessed with care for continuous exposures. Specifically, investigators need to check whether positivity holds for all units. We define positivity as

$$0 < f_{D|\mathbf{X}}(D = d \mid \mathbf{X} = \mathbf{x}) < 1 \ \forall \ d \in \mathcal{D}. \tag{4.2.6}$$

This expression states that all units have the potential to receive a particular level of exposure given any value of the confounders. Typically positivity is enforced by restricting the domain, $\mathcal{D}$, to only the observed exposure interval. However, for continuous exposures this can present a unique challenge especially when there are few exposure values or low density regions in the exposure domain. A common solution is to trim the exposure interval to ensure positivity and prevent outliers from having outsized influence on the shape of the dose-response function [Crump et al., 2009].

Having generated the propensity score and tested the necessary properties, the final choice is the method of bias removal for estimating the dose-response function. As mentioned earlier in this section, Hirano and Imbens [2004] approach bias removal via covariate adjustment similar to the method advocated by Imbens [2000]. To remove the bias associated with confounders modeled via the GPS, we define the conditional mean of the potential outcome for treatment value d as a function of the GPS and observed treatment, $\beta(d, r)$. The average dose-response corresponding to the treatment value, $\mu(d)$, is then obtained by averaging the propensity function over the observed covariates. The method of covariate adjustment relies on the result of weak ignorability when Equation 4.2.1 holds,

(i) $\beta(d, r) = \mathbb{E}[Y(d) \mid r(d, X) = r] = \mathbb{E}[Y \mid D = d, R = r]$ and

(ii) $\mu(d) = \mathbb{E}[\beta(d, r(d, X))]$.

Similar to the discrete-valued case, it is important to note that the expectation in (ii) is not with respect to the GPS, $r(D, X)$, but rather the propensity function evaluated at the treatment level of interest, $r(d, \mathbf{X})$. To implement this method, we estimate the conditional expectation of the outcome, which we term the dose-response, as a function of observed treatment, D, and GPS, R, and then average this conditional expectation over the propensity function at exposure level d. As noted by Hade and Lu [2014], there is potential for bias with misspecification of $\beta(d, r)$ when using covariate adjustment, so modeling this as a flexible function of D and R is crucial. Hirano and Imbens use second order polynomials for each of D and R along with a first order interaction term between the exposure level and the GPS. It is important to clarify that the coefficient estimates for D in the dose-response model should not be interpreted as the treatment effect of D, because $\beta(d, r)$ does not have a causal interpretation.

Alternatively, Imai and Van Dyk [2004] implement subclassification on the value of $r(d, \theta) = r(d, \mathbf{X}^T\beta)$ with a weighted average of the potential outcomes within subclass used to calculate the dose-response function. This is done through a series of steps. In the first step, the parameters of the propensity function are estimated. The estimated value $\hat{\theta}$ is computed based on the parametric form specified for the propensity function. Using $\hat{\theta}$, J subclasses are created of roughly equal size. Within each subclass, a parametric model is chosen for the dose-response, $f(Y(d) \mid D = d)$. Finally, the overall distribution of the dose-response function is computed as the weighted average of the within-subclass distributions for each treatment level d. The authors note that while the theory holds for the marginal distribution given only the treatment assignment, many times additional adjustment within each subclass can be implemented to further reduce the bias. An alternative to estimating each subclass separately is to use penalized regression splines to flexibly model the function across the different subclasses.

4.3 Generalized Propensity Score Extensions

As originally proposed by Hirano and Imbens [2004] and Imai and Van Dyk [2004], the propensity score methodology for continuous treatments focused on specifying the parametric form for the conditional density of the treatment given the covariates of interest, i.e., the GPS. Having used this parametric form to define the GPS, the investigator was free to use their preferred method for removing bias such as covariate adjustment [Hirano and Imbens, 2004] or subclassification [Imai and Van Dyk, 2004]. In the case of covariate adjustment, an additional parametric form for the conditional mean of the outcome given the treatment and GPS is assumed for the dose-response function. Recalling the work of Hade and Lu [2014], misspecification of the propensity score or covariate equation can lead to significantly biased estimates of the treatment effect in the case of a binary treatment. With this in mind, researchers sought out methods to model the GPS and/or dose-response non-parametrically or to introduce additional model flexibility in either the GPS estimation or dose-response model in an effort to reduce the potential bias associated with misspecification of the parametric distribution. Recent methodological developments for the GPS by Flores et al. [2012]; Kennedy et al. [2017], Zhu et al. [2015], and Kreif et al. [2015] have shown the flexibility of modeling the conditional density and/or the dose-response function using kernel estimation, boosting algorithms, and ensemble algorithms, respectively. In addition, alternative methods have been proposed which aim to ensure covariate balance balance of weights described by Robins et al. [2000] for eliminating bias such as the covariate balancing generalized propensity score [Fong et al., 2018] and entropy balancing approaches [Tübbicke, 2020; Vegetabile et al., 2020]. We will briefly discuss some of the major contributions of these works in extending the methodology for continuous exposures.

Both Flores et al. [2012] and Kennedy et al. [2017] advocated non-parametric kernel density approaches for inference with continuous exposures. Flores et al. [2012] use a parametric method to estimate the GPS via a generalized linear model such as the log normal

while, but specifies the dose-response as a local polynomial regression using non-parametric kernel estimators. Two different kernel estimators for the the dose-response were proposed including a partial mean model using the product normal kernel of the exposure and GPS and an inverse weight model using a normal kernel with the exposure alone. The bandwidth is selected global using the procedure described in Fan and Gijbels [1996]. When using this method, it is important to test the sensitivity of results to the choice of kernel function $K(\cdot)$ and bandwidth h. Conversely, Kennedy et al. [2017] propose a completely non-parametric approach to estimating the dose-response curve based on finding a doubly robust mapping from the observed data and two nuisance parameter functions, the conditional density of exposure, i.e., the GPS, and the outcome regression model, where the conditional expectation of the response given treatment is equal to the dose-response curve of interest. In this case the double robustness property ensures that the resultant estimate of the dose-response is unbiased if either one of the nuisance parameters is properly specified. Similar to Flores et al. [2012] though the hyperparameters of the kernel smoothing function must be estimated and the results are potentially sensitive to misspecification.

An alternative to kernel density estimation, both Kreif et al. [2015] and Zhu et al. [2015] model the conditional density for the GPS non-parametrically by leveraging machine learning algorithms. In a typical parametric linear GPS model we would model D as

$$D = \mathbf{X}^T\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

We can generalize this expression by replacing $\mathbf{X}^T\beta$ with a generic mean function $m(\mathbf{X})$. In Zhu et al. [2015] the mean function, $m(\mathbf{X})$, is estimated using a nonparametric boosting algorithm that automatically extracts important covariates, nonlinear terms, and interactions among covariates. The boosting algorithm fits an additive model where each component is a regression tree. In Zhu et al. [2015] the estimated GPS via boosting is then used to create inverse probability weights. The weights are then used as part of a regression spline function for the outcome model, selected using weighted AIC or weighted BIC criterion. Rather than using a single machine learning approach such as boosting, Kreif et al. [2015] employs a

machine learning method called the 'SuperLearner', which solves an optimization problem for a specified loss function given a list of potential regression estimators. This algorithm is also called a stacked prediction method, where the solution is a convex combination of the list of prediction algorithms. Unlike the methods of Flores et al. [2012] and Zhu et al. [2015], the SuperLearner can be used to estimate both the conditional density for the GPS and the mean outcome given the GPS. In Kreif et al. [2015], a wide variety of linear, quadratic, spline, GAM, and Bayesian GLM specifications are included as potential estimators for the outcome regression while the GPS estimates include both normal and gamma error distributions with various degrees of interactions and polynomials of the pre-treatment covariates.

Each of these extensions presented so far have focused on adaptions to increase the flexibility of modeling the generalized propensity score and/or dose-response function. Several recent methods have instead focused on achieving covariate balance as part of the estimation procedure. In Fong et al. [2018] the authors propose including explicit covariate balancing conditions when estimating the GPS. The authors propose both a parametric and non-parametric method. In the parametric method, parameters are estimated using method of moments with score conditions on the ratio of marginal exposure density to conditional exposure density and the weighted cross moment between exposure and covariates. In the non-parametric method, an empirical likelihood method is used that estimates the stabilized inverse generalized propensity score weights without directly estimating the GPS but instead using constrained optimization on the weights. The non-parametric method proposed by Fong et al. [2018] shares many similarities with the entropy balancing approaches of Tübbicke [2020] and Vegetabile et al. [2020]. Both entropy balancing methods aim to directly solve for the weights using constrained optimization similar to the approach of Hainmueller [2012] for binary exposure. The term entropy in these methods refers to the entropy metric, $h(w_i) = w_i \ln(w_i)$, introduced in Shannon [1948] and is included in each proposed method as part of the loss function. The authors use Lagrange-multipliers to then solve the constructed loss functions with additional constraints that the weights sum to the total number of ob-

servations and that the balance the correlation between the exposure and covariates. The two proposed methods differ primarily in that Vegetabile et al. [2020] includes additional higher order moment constraints of the marginal exposure and between the exposures and covariates.

4.4 Limitations of Current GPS Methodology

These extensions of the GPS provide added flexibility when faced with data that violate parametric assumptions. Each method has different strengths and weaknesses. For instance the boosting approach by Zhu et al. performs non-linear covariate selection. On the other hand, the boosting method is sensitive to initial model parameters such as the number of trees and nodes which may lead to overfitted models that are not generalizable and is computationally intensive. The SuperLearner approach of Kreif et al. is able to incorporate a wide variety of potential estimators as part of the ensemble prediction, and often the constraint on α produces results that are predominantly a mixture of only a few of the potential estimators. This is also a restriction though, as the method does not test estimators that are not listed, so it is possible that the true best estimator is left out when optimizing the loss function. Similarly, the kernel estimation methods of Flores et al. and Kennedy et al. are useful in that they smoothly estimate the dose-response function by using local polynomial regression, but they too are dependent on hyperparameters such as the choice of kernel and the bandwidth. Covariate balancing methods discussed by Fong et al., Tübbicke, and Vegetabile et al. make substantial improvements to ensure that the balancing property of the GPS is met. However, they may also lead to potential bias by sacrificing some precision to achieve balance.

Most importantly, all of these methods handle only the case of a single continuous exposure. There are unique challenges to extending any of these proposed methods to multiple exposures, including: appropriately defining the exposure domain, assessing balance across

multiple dimensions, handling varying degrees of overlapping confounding, and properly estimating a high dimensional dose-response function. In the following chapter we propose a new method to address these challenges and formalize the framework for causal inference with multiple continuous exposures with emphasis on bivariate exposures.

CHAPTER 5

Generalized Propensity Score for Multivariate Continuous Exposures

5.1 Notation

Our approach to developing the multivariate generalized propensity score follows the Neyman-Rubin causal model and uses the potential outcome notation introduced by Neyman [Neyman, 1923] and made popular by Rubin [Rubin, 1974]. Let Y_i denote the outcome of interest for unit i from a population of size n and $\mathbf{D}_i$ be a vector of length m providing the values for m continuous exposures for unit i. The confounders relevant to each exposure are allowed to be different. Let $C_i = \{\mathbf{C}_{i1}, \ldots, \mathbf{C}_{im}\}$ be a set of size m where each element in the set, $\mathbf{C}_{ij}$, $j = 1, \ldots, m$, is a p_j dimensional vector of baseline confounders associated with the j^{th} exposure and the outcome. We denote the value of the k^{th} confounder of the j^{th} exposure for the i^{th} individual as C_{ijk}, with $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 1, \ldots, p_j$. If all exposures have identical confounders, then $\mathbf{C}_{i1} = \cdots = \mathbf{C}_{im} = \mathbf{C}_i$ and $p_1 = \cdots = p_m = p$. The observed data for the i^{th} unit is represented as $(Y_i, D_{i1}, \ldots, D_{im}, C_{i11}, \ldots, C_{i1p_1}, \ldots, C_{im1}, \ldots, C_{imp_m})$. Further, we define the potential outcome $Y_i(\mathbf{d})$ as the outcome that the i^{th} subject would have if assigned the exposure vector $\mathbf{d} = (d_1, \ldots, d_m)$. We will use capital $\mathbf{D}$ to represent the multivariate random variable representing dose combinations, and lowercase $\mathbf{d}$ as a particular value in the multidimensional space. Estimation focuses on the average dose-response function defined as $\mu(\mathbf{d}) = \mathbb{E}[Y(\mathbf{d})]$, which is assumed to be well defined for any $\mathbf{d} \in \mathcal{D} \subseteq \mathbb{R}^m$. Note that with a bivariate exposure, i.e., $m = 2$, $\mu(\mathbf{d})$ is a dose-response

surface in 3-dimensional space.

5.2 Identification Assumptions

We make the following identifying assumptions: weak ignorability, positivity, and stable-unit treatment value. Weak ignorability, also known as selection on observables or unconfoundedness, states that exposure is conditionally independent of the potential outcomes given the appropriate set of confounders. We write this in the multivariate case as

$$Y_i(\mathbf{d}) \perp\!\!\!\perp \mathbf{D}_i \mid \mathbf{C}_{i1}, \ldots, \mathbf{C}_{im} \quad \forall \quad \mathbf{d} \in \mathcal{D}.$$

When this assumption holds, we can replace the high-dimensional conditioning set with a scalar value by means of the conditional density function of exposure [Rosenbaum and Rubin, 1983]. In our case the conditional density is defined as the multivariate generalized propensity score (mvGPS), which we denote $f_{\mathbf{D}|\mathbf{C}_1,\ldots,\mathbf{C}_m}$. Weak ignorability is often the most difficult assumption to rationalize as it requires perfect knowledge and collection of all possible confounders of the exposures and outcome in the set C. We assume that the set C is well defined and that there is no unmeasured confounding.

The second assumption, positivity, claims that all units have the potential to receive a particular level of exposure given any value of the confounders. In notation, we have

$$0 < f_{\mathbf{D}|\mathbf{C}_1,\ldots,\mathbf{C}_m}(\mathbf{D} = \mathbf{d} \mid \mathbf{C}_1, \ldots, \mathbf{C}_m) < 1 \quad \forall \quad \mathbf{d} \in \mathcal{D}.$$

This assumption requires that we carefully define $\mathcal{D}$ such that all units have the potential to receive any particular value in the domain. In the case of a univariate continuous exposure, positivity is often enforced by restricting estimation to either the observed range or a trimmed version [Crump et al., 2009]. For example, using the observed range we would define $\mathcal{D} = [d_0, d_1]$ where d_0 and d_1 correspond to the minimum and maximum observed exposure. In the case of a multivariate exposure, a natural inclination might be to extend this approach

to multiple dimensions by setting $\mathcal{D} = \mathcal{G}$ where $\mathcal{G}$ is defined as

$$\mathcal{G} = \prod_{j=1}^{m} [d_{0j}, d_{1j}] \subset \mathbb{R}^m,$$

where d_{0j} and d_{1j} are the minimum and maximum observed exposure, respectively, along dimension j. However, when exposure variables are correlated, i.e., $Cov(D_j, D_{j'}) \neq 0$ for $j \neq j'$, the region $\mathcal{G}$ may include areas with few or no observations. Instead, we propose defining the estimable region for multivariate exposures as $\mathcal{D} = \mathcal{H} \subset \mathcal{G}$, where $\mathcal{H}$ is defined as the convex hull of the multivariate exposure [Chazelle, 1993]. Using a convex hull ensures that inference is restricted to regions where data are observed and avoids extrapolating to sparse data regions in the multidimensional space. For the case of $m = 2$, Figure 3 on page 65 shows the difference between regions $\mathcal{G}$ and $\mathcal{H}$ when $Cov(D_1, D_2) = 0.5$. Additionally, similar to the univariate case, we can define trimmed versions of $\mathcal{G}$ or $\mathcal{H}$. By specifying a value $q \in [0.5, 1]$, we construct $\mathcal{G}_q$ using trimmed minimum and maximum values as

$$\mathcal{G}_q = \prod_{j=1}^{m} [d_{0j}^q, d_{1j}^q] \subset \mathcal{G},$$

where $d_{0j}^q = Q(\mathbf{d}_j, 1-q)$, $d_{1j}^q = Q(\mathbf{d}_j, q)$, and $Q(\cdot, q)$ is the sample quantile function. To create the trimmed convex hull, $\mathcal{H}_q$, we recalculate the convex hull using the subset of observations that falls within the trimmed minimum and maximum across all exposure dimensions.

The final assumption is the stable-unit treatment value assumption (SUTVA), which states that the potential outcome of each unit does not depend on the exposure that other units receive and that there exists only one version of each exposure [Rubin, 1980]. This assumption rules out potential interference between units or other errors in defining the potential outcomes caused by multiple versions of the exposure. Therefore the potential outcomes are well-defined for each unit and the observed outcome given exposure $\mathbf{D} = \mathbf{d}$ corresponds to the potential outcome, i.e., $Y_i(\mathbf{d}) = Y_i$. We discuss the tenability of this assumption to our data application in the Discussion.

5.3 Multivariate Generalized Propensity Score

Using the identifying assumptions above, there are a variety of different methods to estimate the dose-response function including covariate adjustment [Hirano and Imbens, 2004] or stratification [Imai and Van Dyk, 2004]. We focus on weighted estimation, originally proposed for binary treatments with marginal structural models [Robins et al., 2000] and motivated by weights used in survey sampling [Horvitz and Thompson, 1952]. We aim to construct a set of weights, w, that when applied to the observed data return a consistent estimate for the average dose-response function, i.e.,

$$\mathbb{E}[wY \mid \mathbf{D}] = \mathbb{E}[Y(\mathbf{d})]. \tag{5.3.1}$$

In the case of univariate continuous exposure, weights are constructed by either estimating the generalized propensity score [Fong et al., 2018; Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Kennedy et al., 2017; Zhu et al., 2015] or by direct optimization using an entropy loss function [Tübbicke, 2020; Vegetabile et al., 2020]. We choose to extend the generalized propensity score by using an appropriately defined multivariate conditional distribution, which we refer to as the multivariate generalized propensity score (mvGPS). The weights are thus constructed as the ratio of the multivariate marginal density to the conditional density

$$w = \frac{f(\mathbf{D})}{f(\mathbf{D}|\mathbf{C}_1, \ldots, \mathbf{C}_m)}, \tag{5.3.2}$$

where the numerator is the marginal density of the multivariate exposure and the denominator is the mvGPS. These weights are referred to in the literature as stabilized inverse probability of treatment weights (IPTW) [Robins et al., 2000]. To motivate the intuition behind constructing weights in this manner, we can note that $w = 1$ when the probability of exposure is independent of the confounding set C, i.e., $f(\mathbf{D} \mid \mathbf{C}_1, \ldots, \mathbf{C}_m) = f(\mathbf{D})$, which would hold in the case of a randomized experiment. For tractability, we propose using

multivariate normal models for both densities, i.e.,

$$\mathbf{D} \sim \mathrm{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \mathbf{D} \mid \mathbf{C}_1, \ldots, \mathbf{C}_m \sim \mathrm{N}_m\left(\begin{bmatrix} \boldsymbol{\beta}_1^T \mathbf{C}_1 \\ \vdots \\ \boldsymbol{\beta}_m^T \mathbf{C}_m \end{bmatrix}, \boldsymbol{\Omega} \right),$$

where each $\boldsymbol{\beta}_j^T$ is a row vector of length p_j corresponding to the effect of the set of confounders $\mathbf{C}_j$ on D_j. By factorizing both the numerator and denominator in Equation 5.3.2, we can compute w using full conditionals, i.e.,

$$\begin{aligned} w =& \frac{f(D_m \mid D_{m-1}, \ldots, D_1) \cdots f(D_1)}{f(D_m \mid \mathbf{C}_1, \ldots, \mathbf{C}_m, D_{m-1}, \ldots, D_1) \cdots f(D_1 \mid \mathbf{C}_1, \ldots, \mathbf{C}_m)}, \\ w =& \frac{f(D_m \mid D_{m-1}, \ldots, D_1) \cdots f(D_1)}{f(D_m \mid \mathbf{C}_m, D_{m-1}, \ldots, D_1) \cdots f(D_1 \mid \mathbf{C}_1)}, \end{aligned} \tag{5.3.3}$$

where each conditional expression is univariate normal. The second line is a result of the fact that the j^{th} exposure is independent of the confounders of other exposures given $\mathbf{C}_j$, i.e.,

$$D_j \perp\!\!\!\perp \mathbf{C}_{-j} \mid \mathbf{C}_j \; \forall \; j = 1, \ldots, m,$$

where $\mathbf{C}_{-j}$ represents the set of confounders excluding $\mathbf{C}_j$, i.e., $\mathbf{C}_{-j} = \{\mathbf{C}_1, \ldots, \mathbf{C}_m\} \setminus \mathbf{C}_j$. Evaluating only the conditional densities reduces computational burden by eliminating the need to directly estimate the covariance matrices, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$.

Let $\boldsymbol{\theta}$ be the collection of mean and variance parameters from all of the univariate normal densities in Equation 5.3.3. Estimation of the parameters to obtain $\hat{\boldsymbol{\theta}}$ proceeds by maximizing the corresponding conditional density via least squares. The weight for the i^{th} subject, w_i, is obtained by evaluating the densities using $\hat{\boldsymbol{\theta}}$ with the values of the observed exposures, $D_{i1}, \ldots, D_{im}$, and confounders, $\mathbf{C}_{i1}, \ldots, \mathbf{C}_{im}$.

When the weights are properly specified, the covariance between each exposure D_j and

confounder C_{jk} for $j = 1, \ldots, m$ and $k = 1, \ldots, p_j$ is zero:

$$\begin{aligned}
\mathbb{E}[w(D_j - \mu_{D_j})(C_{jk} - \mu_{C_{jk}})] &= \int_{\mathcal{D}} \int_{\mathcal{C}_1} \cdots \int_{\mathcal{C}_m} w(d_j - \mu_{D_j})(c_{jk} - \mu_{C_{jk}}) f(\mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{D}} \int_{\mathcal{C}} \frac{f(\mathbf{d})}{f(\mathbf{d}|\mathbf{c}_1, \ldots, \mathbf{c}_m)} (d_j - \mu_{D_j})(c_{jk} - \mu_{C_{jk}}) f(\mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{D}} \int_{\mathcal{C}} \frac{f(\mathbf{d}) f(\mathbf{c}_1, \ldots, \mathbf{c}_m)}{f(\mathbf{d}|\mathbf{c}_1, \ldots, \mathbf{c}_m) f(\mathbf{c}_1, \ldots, \mathbf{c}_m)} (d_j - \mu_{D_j})(c_{jk} - \mu_{C_{jk}}) f(\mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{D}} \int_{\mathcal{C}} (d_j - \mu_{D_j})(c_{jk} - \mu_{C_{jk}}) f(\mathbf{d}) f(\mathbf{c}_1, \ldots, \mathbf{c}_m) \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{D}} (d_j - \mu_{D_j}) f(\mathbf{d}) \partial \mathbf{d} \int_{\mathcal{C}} (c_{jk} - \mu_{C_{jk}}) f(\mathbf{c}_1, \ldots, \mathbf{c}_m) \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= 0.
\end{aligned} \tag{5.3.4}$$

This balancing property of the weights serves as an important diagnostic when using the mvGPS as part of a causal analysis [Austin, 2019]. Weights that do not reduce the exposure-confounder correlation suggest that the distributional assumptions are invalid, the propensity equations are misspecified, or that there are insufficient data as the balance is achieved asymptotically.

Further, it follows that these weights are already normalized, i.e, $\mathbb{E}[w] = 1$, and they maintain the marginal moments of $\mathbf{D}$ and $\mathbf{C}_j$, meaning $\mathbb{E}[wD_j] = \mathbb{E}[D_j]$ and $\mathbb{E}[w\mathbf{C}_j] = \mathbb{E}[\mathbf{C}_j]$ for $j = 1, \ldots, m$, where the expectations are taken with respect to the joint density $f(\mathbf{D}, \mathbf{C}_1, \ldots, \mathbf{C}_m)$.

It remains to show that the weights as constructed satisfy Equation 5.3.1. To do this we follow the logic proposed by Robins on using IPTW to correct for confounding [Robins, 2000]. We first note that the joint density of the potential outcome can be factorized as

$$\begin{aligned}
f(Y(\mathbf{d}), \mathbf{D}, \mathbf{C}_1, \ldots, \mathbf{C}_m) &= f(\mathbf{D} \mid Y(\mathbf{d}), \mathbf{C}_1, \ldots, \mathbf{C}_m) f(\mathbf{C}_1, \ldots, \mathbf{C}_m \mid Y(\mathbf{d})) f(Y(\mathbf{d})) \\
&= f(\mathbf{D} \mid \mathbf{C}_1, \ldots, \mathbf{C}_m) f(\mathbf{C}_1, \ldots, \mathbf{C}_m \mid Y(\mathbf{d})) f(Y(\mathbf{d})),
\end{aligned}$$

where the second line follows from the assumption of weak ignorability. We can then let $f(\mathbf{D})$ be a density for our multivariate exposure and construct a new joint density f^* where we

replace $f(\mathbf{D} \mid \mathbf{C}_1, \ldots, \mathbf{C}_m)$ with $f(\mathbf{D})$ as would be the case if the exposures were independent of the confounders. This new density is written as

$$f^*(Y(\mathbf{d}), \mathbf{D}, \mathbf{C}_1, \ldots, \mathbf{C}_m) = f(\mathbf{D}) f(\mathbf{C}_1, \ldots, \mathbf{C}_m \mid Y(\mathbf{d})) f(Y(\mathbf{d})),$$

where the marginal mean of the potential outcomes is equivalent under either joint density f or f^*, i.e., $\mathbb{E}^*[Y(\mathbf{d})] = \mathbb{E}[Y(\mathbf{d})]$. Using this new density we can write our dose response as

$$\mathbb{E}^*[Y(\mathbf{d})] = \mathbb{E}^*[Y(\mathbf{d}) \mid \mathbf{D} = \mathbf{d}] = \mathbb{E}^*[Y(\mathbf{D}) \mid \mathbf{D} = \mathbf{d}] = \mathbb{E}^*[Y \mid \mathbf{D} = \mathbf{d}],$$

using the SUTVA assumption. The resulting expression, $\mathbb{E}^*[Y \mid \mathbf{D} = \mathbf{d}]$, is equivalent to the mean expression in a linear regression of the observed exposures on outcome. Finally, we have

$$\begin{aligned}
\mathbb{E}^*[Y \mid \mathbf{D} = \mathbf{d}] &= \int_{\mathcal{Y}} \int_{\mathcal{D}} \int_{\mathcal{C}} y f^*(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial y \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{Y}} \int_{\mathcal{D}} \int_{\mathcal{C}} y \frac{f^*(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m)}{f(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m)} f(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial y \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{Y}} \int_{\mathcal{D}} \int_{\mathcal{C}} y \frac{f(\mathbf{d})}{f(\mathbf{d} \mid \mathbf{c}_1, \ldots, \mathbf{c}_m)} f(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial y \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \int_{\mathcal{Y}} \int_{\mathcal{D}} \int_{\mathcal{C}} w y f(y, \mathbf{d}, \mathbf{c}_1, \ldots, \mathbf{c}_m) \partial y \partial \mathbf{d} \partial \mathbf{c}_1, \ldots, \partial \mathbf{c}_m \\
&= \mathbb{E}[wY \mid \mathbf{D} = \mathbf{d}],
\end{aligned} \tag{5.3.5}$$

which gives us the result from Equation 5.3.1 that our weighted regression does indeed provide a consistent estimate of the dose-response function.

CHAPTER 6

Simulation

6.1 Design

We conducted a simulation study to demonstrate the performance of the mvGPS method under different scenarios of confounding and compare it to three commonly used univariate methods. The univariate methods were entropy balancing [Tübbicke, 2020], the covariate balanced generalized propensity score (CBGPS) [Fong et al., 2018], and the generalized linear propensity score (PS). The entropy balancing method uses non-parametric constrained optimization with an entropy loss function to solve for weights without specifying a propensity score model. CBGPS attempts to achieve propensity specification and covariate balance simultaneously by introducing a penalty term into the likelihood. The PS method uses univariate normal densities for the marginal distribution of exposure and the generalized propensity score without balance constraints. Although these univariate methods can handle only single exposure variables, we expected that they might perform adequately when the multiple exposure variables are highly correlated and have the same confounders. However, when exposure variables have separate sets of confounders and/or are only weakly correlated, we expected that the mvGPS method would outperform the univariate methods.

In our simulations we focus exclusively on a bivariate exposure, $m = 2$, similar to that found in our motivating example. For each simulated data scenario, each univariate method was applied twice, once to each exposure variable, with each such application yielding a set of weights that were used to assess balance on confounders and estimate the dose-response

function.

The first step of the simulation is to draw the vector of covariates $\mathbf{X}$ for each unit. We assume that there are a total of 10 covariates collected prior to exposure and that the covariates follow a normal distribution,

$$\mathbf{X} \sim N_{10}(\mathbf{0}, \mathbf{\Sigma}_X),$$

where the covariance matrix $\mathbf{\Sigma}_X$ is compound symmetric with variance 1 and covariance 0.2, to create a set of correlated covariates.

Realizations of the conditional distribution of the bivariate continuous exposure levels, $\mathbf{D} = (D_1, D_2)^T$ given $\mathbf{X}$, were then generated as bivariate normal,

$$\mathbf{D} \mid \mathbf{X} \sim N_2(\boldsymbol{\beta}\mathbf{X}, \mathbf{\Sigma}_{D|X}),$$

where $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \boldsymbol{\beta}_2^T \end{bmatrix}$ is a 2×10 matrix with row vectors $\boldsymbol{\beta}_1^T$ and $\boldsymbol{\beta}_2^T$ representing the effects of $\mathbf{X}$ on D_1 and D_2, respectively, and $\mathbf{\Sigma}_{D|X}$ is the 2×2 conditional covariance matrix. For all simulations the conditional standard deviation for each exposure was set to 2, while values of the conditional correlation $\rho_{D|X}$ were allowed to vary over $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

Note that the marginal covariance matrix of the exposures, $\mathbf{\Sigma}_D$, is equal to $\mathbf{\Sigma}_D = \mathbf{\Sigma}_{D|X} + \boldsymbol{\beta}\mathbf{\Sigma}_X\boldsymbol{\beta}^T$. This means that the marginal correlation of the two exposure variables, ρ_D, depends on their conditional correlation $\rho_{D|X}$, the covariance of $\mathbf{X}$ and the degree of overlap of covariates. The degree of overlap is reflected in the number of non-zero elements that are common between $\boldsymbol{\beta}_1^T$ and $\boldsymbol{\beta}_2^T$. As the degree of overlap increases, the marginal correlation also increases. Since $\mathbf{\Sigma}_X$ is compound symmetric with constant covariance of 0.2, the marginal correlation is guaranteed to be non-zero even with zero overlap and zero conditional correlation.

Finally, the outcome Y was sampled from a univariate normal distribution conditional

on $\mathbf{D}$ and $\mathbf{X}$ as

$$Y \mid \mathbf{D}, \mathbf{X} \sim N\left(\boldsymbol{\alpha}^T \begin{bmatrix} \mathbf{X} \\ \mathbf{D} \end{bmatrix}, \sigma_Y^2\right) = N(\boldsymbol{\alpha}_{\mathbf{X}}^T \mathbf{X} + \boldsymbol{\alpha}_{\mathbf{D}}^T \mathbf{D}, \sigma_Y^2),$$

where $\boldsymbol{\alpha}^T = \begin{bmatrix} \boldsymbol{\alpha}_{\mathbf{X}}^T & \boldsymbol{\alpha}_{\mathbf{D}}^T \end{bmatrix}$ is a 1×12 vector of coefficients, which we separate as $\boldsymbol{\alpha}_{\mathbf{X}}^T$, a 1×10 vector representing the effect of covariates on the outcome, and $\boldsymbol{\alpha}_{\mathbf{D}}^T$, a 1×2 vector corresponding to the treatment effects. In all simulations the conditional standard deviation of the outcome was equal to 4, i.e., $\sigma_Y = 4$.

Three scenarios were constructed to reflect different degrees of overlap of confounding for the two exposures: M1: No Common Confounding, M2: Partially Common Confounding, and M3: Common Confounding. Directed acyclic graphs (DAGs) for each scenario are shown in Figure 4 on page 66.

Tables 2, 3, and 4 on pages 58-60 display the coefficients in the vectors $\boldsymbol{\beta}_1^T$, $\boldsymbol{\beta}_2^T$ and $\boldsymbol{\alpha}^T$ for each scenario. In M1, the two exposures D_1 and D_2 each have five covariates, with none in common; for each exposure, two of the covariates are true confounders (i.e., also associated with Y). The outcome Y is a function of the two exposures as well as the four true confounders, none of which are shared between D_1 and D_2. In M2, D_1 and D_2 again have five covariates each, but they share three in common. Two of the shared covariates are true confounders, and each exposure has a confounder that is not shared with the other exposure. The outcome Y is again a function of the two exposures and the four true confounders, two of which are shared and two of which are not. In M3, D_1 and D_2 share the same five covariates. Four of these are true confounders, and Y is a function of the two exposures and four common confounders. In all scenarios, the treatment effect for each of the exposures was set to 1, i.e., $\boldsymbol{\alpha}_{\mathbf{D}}^T = (1, 1)$.

The three simulation scenarios were run with a sample size of $n = 200$ for a total of $B = 1000$ Monte Carlo repetitions using R Version 4.0 [R Core Team, 2020]. For each repetition, weights were estimated using mvGPS and the three univariate methods with the proper set of confounders specified for each exposure. For example, for Scenario M1,

weights were constructed for D_1 using X_2 and X_4 while D_2 depended on X_6 and X_9 (see Table 2). The three univariate methods, entropy balancing, CBGPS and PS, were implemented using the `WeightIt` package in R [Greifer, 2020]. The parametric version of CBGPS was used. The mvGPS method was implemented using the mvGPS package in R. All methods were compared against an unweighted approach equivalent to applying a weight of 1 to all observations.

Weighted Pearson correlations between the exposures and confounders were used to assess balancing performance [Austin, 2019; Zhu et al., 2015]. We examined maximum absolute correlation, which reflects the most imbalanced confounder after weighting and has been shown to be a key metric to assess balance [Diamond and Sekhon, 2013], and the average absolute correlation, which summarizes how well balance is achieved over all confounders. These correlation values were taken over both sets of exposures.

Effective sample sizes, $(\Sigma_i w_i)^2/\Sigma_i w_i^2$, were calculated to summarize the relative power of each method [Kish, 1965]. The weights were then used to estimate the dose-response model. The performance metrics were absolute total bias, $\Sigma_j|\alpha_{D_j} - \hat{\alpha}_{D_j}|$ for $j = 1, 2$, and root mean squared error, $\sqrt{\frac{1}{n}\Sigma_i(y_i^* - \hat{y}_i)^2}$, where $i = 1, \ldots, 500$ samples, y^*, were drawn from a uniform grid on the convex hull, $\mathcal{H}$, over the observed joint distribution of the two exposures. Each of the metrics was averaged over the 1000 repetitions.

All methods reduce the effective sample size when weights are applied to the sample. Of particular concern for practitioners are extreme weights. When sample sizes are small or moderate, extreme weights can have an outsized influence. They may also result in limited power to detect treatment effects and erratic estimation [Kang and Schafer, 2007]. One remedy is to trim extreme weights [Huber et al., 2013; Lee et al., 2011]. As all simulations were run with moderate sample sizes, i.e., $n = 200$, we wanted to test performance when weight trimming was applied as might be done in practice by analysts when faced with extreme weights. Our simulation analyses were thus repeated using trimmed weights for each method, w_q, where $q \in \{0.99, 0.95\}$. Weights were trimmed at both the upper and lower

bounds of the respective sample percentile such that values above or below the thresholds were replaced with the threshold value.

6.2 Simulation Results

Figure 5 on page 67 plots the absolute maximum correlation between the exposures and confounders for each method along with the original unweighted correlations for comparison. In general, with no common confounding or partially common confounding, the mvGPS method substantially outperformed each univariate method. However, for common founding, mvGPS performs best only when the marginal correlation is low. The performance of univariate methods tended to cluster differently based on the degree of confounding overlap. In models with low overlap, performance was clustered based on exposure, D_1 or D_2, but as the degree of overlap increased, performance became clustered by type of estimation, Entropy, CBGPS, or PS. Applying trimmed weights, we see a slight improvement for $q = 0.99$ while $q = 0.95$ has little to no effect.

Figure 6 on page 68 shows the average absolute exposure-covariate correlation along with a reference line at 0.1, a benchmark suggesting sufficient covariate balance [Zhu et al., 2015]. For all simulation models, the mvGPS is consistently near the 0.1 threshold. For the univariate methods, we see trends in performance similar to those observed for the maximum correlation, but differences between methods are smaller. Entropy methods consistently had the lowest average correlation, especially with high overlap or high marginal correlation. Trimming the weights tended to eliminate the effect of the marginal correlation on the mvGPS, resulting in flatter lines for $q = 0.99$ and $q = 0.95$, particularly for models with at least some common confounding.

Figure 7 on page 69 displays trends in effective sample size for each method across the various simulation scenarios, with a reference line at 100, which is often a minimum desirable quantity for inference in the dose-response model [Vegetabile et al., 2020]. The mvGPS

method tends to have lower effective sample sizes compared to the univariate methods. Importantly, using untrimmed weights, the mvGPS method has effective sample sizes less than 100 in the presence of partial or common confounding, indicating particularly low power. Trimming the weights increases the effective sample size for all methods and the difference between methods decreases as q increases.

Figure 8 on page 70 shows the results of each method with respect to total absolute bias for the treatment effects estimated from weighted regression. Generally, the mvGPS method has the lowest total bias, with the exception of high correlation in the model with partially common confounding or no common confounding. Of particular note, although the effective sample size and balancing diagnostics were lower for mvGPS in the common confounding model, it significantly outperforms all univariate methods with respect to bias even with high marginal correlation. We also observe that certain univariate methods have greater bias than the unweighted estimates, such as those that estimate weights using D_2 for the common confounding model. Trimming the weights tended to reduce bias for mvGPS when there was high correlation of the exposures, particularly under models with either partially overlapping confounding or no common confounding, while slightly increasing the bias for the common confounding models.

Figure 9 on page 71 shows the root mean squared error based on 500 points sampled along a grid from the convex hull, $\mathcal{H}_q$, of the exposures. The precision of predicted values for the mvGPS is often worse than that for univariate methods. As ρ increases, the mvGPS method has worse performance, with decreased power from low effective sample sizes. Trimming the weights helps reduce this trend and reduces the root mean squared error across all methods.

In summary, using a multivariate method for weight estimation is critical to achieve balance as univariate methods in general do not effectively balance on the confounders for both exposures. The multivariate method protects against any single confounder being strongly imbalanced across either exposure at the expense of slightly lower average balance, while the univariate methods have potentially large imbalance on the unused exposure dimension. The

notable exception is when there is high overlap in terms of confounders. In this case univariate methods can sufficiently balance confounders, particularly when the marginal correlation of exposures is high. However, despite achieving balance in these scenarios, the univariate methods still resulted in high total bias of the treatment effect estimates. Although mvGPS weights were advantageous with respect to balance and bias, they tend to produce smaller effective sample sizes, resulting in lower power and higher root mean squared error. Weight trimming, particularly with $q = 0.99$, offers a potential remedy to reduce these effects while also maintaining balance and low bias.

CHAPTER 7

Application

The WIC program is designed to provide nutrition education, 'vouchers' for selected healthy food, and referrals. The rapid increase in childhood obesity rates in the early 2000s led to interventions to increase physical activity, and improve access to healthy food especially in communities where affordable fresh produce is not available. In this motivating example, we estimated the causal effects of interventions that were implemented by individual WIC clinics in attempts to meet the specific needs of the communities they served. As discussed in Section 2, the intervention programs were classified as using macro or micro strategies and we calculated two continuous dose measures for $n = 1079$ census tracts. The outcome was the difference in average obesity prevalence from post, 2012-2016, to pre, 2007-2009, intervention period, calculated as $Y = \bar{p}_{post} - \bar{p}_{pre}$ at the census tract level. Negative values of Y indicate that the prevalence of obesity decreased. We hypothesized that areas with more macro and micro strategies would have the greatest reduction in rates of obesity.

Data on potential census tract-level confounders came from three sources: US Census American Community Survey (ACS) 5-year estimates [United States Census Bureau, 2020], WIC administrative data, and the National Establishment Time-Series (NETS) [Walls & Associates, 2013]. Variables from the ACS captured community level demographic characteristics such as median household income, education level, primary language spoken at home, and ethnicity, which have been shown in previous research to be associated with obesity rates [Nobari et al., 2013, 2018a]. WIC administrative data were used to calculate average pre-treatment overweight and obesity prevalence for each census tract. Overweight

and obesity prevalence were considered potential confounders because agencies may have directed interventions towards clinics in higher prevalence regions. Finally, NETS provided information on neighborhood food environments, specifically on the density per square mile of unhealthy and healthy food outlets [Anderson et al., 2020; Wang et al., 2006]. Previous analysis has shown that higher density of healthy outlets was associated with lower obesity prevalence among low-income preschool-aged children in Los Angeles County [Chaparro et al., 2014]. Both macro and micro propensity dose equations included the same set of potential confounders from these three data sources, but each exposure was assessed separately to determine if higher order polynomial terms for any confounders were needed. These analyses showed that both macro and micro dose had quadratic relationships with education level and density of food outlets, while only macro dose had evidence of a quadratic relationship with median household income.

After defining the appropriate functional form for each exposure and confounder, weights were then estimated using the mvGPS method and the three univariate methods discussed in Section 6.1. To maintain the assumption of positivity, data used for estimating the weights were restricted to the trimmed convex hull $H_{0.95}$ shown in Figure 1 on page 63, where a bivariate normal distribution is plausible. The marginal correlation of exposures was moderate, $r = 0.28$, in this high-density region. As both exposures had nearly identical sets of potential confounders, the data generating mechanism was akin to the common confounding scenario described in the simulations. Therefore, to protect against extreme weights and reduce the variability of the resulting dose-response estimates, weights for each method were trimmed using $q = 0.99$.

Table 5 on page 61 shows the balancing diagnostics, maximum absolute correlation and average absolute correlation, and the effective sample sizes. The confounders were significantly imbalanced prior to weighting with the average absolute correlation above 0.2 and the maximum absolute correlation above 0.4. All methods were able to improve balance, but the mvGPS method had substantially greater reduction in imbalance than the univariate

methods. The average absolute correlation and maximum absolute correlation were reduced to 0.04 and 0.10 respectively after applying mvGPS weights. The effective sample size for mvGPS was reduced from the original sample of $n = 1079$ to 604. However, since the population included over 1000 census tract units, the power was still reasonably high.

Finally, we applied weights from the mvGPS method to estimate the joint effect of macro and micro exposure doses on change in obesity prevalence using weighted least squares regression and compare these to unweighted estimates. Only exposures within the trimmed convex hull, $H_{0.95}$, were used to estimate treatment effects and predict the dose-response surface. The dose-response model for both methods included linear terms for each exposure and an interaction between the two exposures.

Figure 10 on page 72 shows the weighted and unweighted dose-response surfaces along with a reference plane of no change in obesity prevalence. Both the weighted and unweighted surfaces suggest reductions in obesity prevalence from our pre-intervention period, 2007-2009, to the post-intervention period, 2012-2016, for all dose combinations. This is consistent with studies showing a decrease in obesity risk among WIC-participating children in Los Angeles County associated with the 2009 change in the WIC food packages [Chaparro et al., 2019; Nobari et al., 2018b]. The unweighted dose-response surface is a monotonic plane; increases in micro dose and in macro dose are each associated with greater reduction in obesity prevalence, the associations are additive, and the greatest reduction in prevalence corresponds to the highest levels of macro and micro doses. The mvGPS dose-response surface is more complex and shows an interaction effect. At low levels of macro dose, increases in micro dose are associated with a steep reduction in obesity prevalence. However, as macro dose increases, high micro dose becomes gradually less effective. In the quadrant where both macro and micro dose are high, higher micro doses appear to be less beneficial rather than more beneficial. Table 6 on page 62 shows a simplified summary of the dose response surfaces in Figure 10, where the two methods are used to estimate the change in obesity prevalence and corresponding 95% confidence interval at the four quadrants of bivariate exposure. There

are several possible explanations for the observed differences between the methods in the high macro, high micro quadrant. There could be important confounders that were not accounted for in the analysis. There could also have been measurement error in estimating exposures. We noted that the data set included several observations with high macro and high micro doses that were assigned high weights and had either no decrease or a slight increase in child obesity prevalence. Further investigation of these census tracts may yield more information and guide model refinements.

CHAPTER 8

Discussion

In this work, we introduced methodology for generating a multivariate generalized propensity score, mvGPS, to be used in estimating the causal effect of multiple simultaneous continuous exposures in observational or non-randomized studies. We have developed the R package `mvGPS` available at https://cran.r-project.org/package=mvGPS in the CRAN repository to implement the methods.

Through simulations we have shown that, when estimating the causal effects of two simultaneous exposures, mvGPS weights are effective at reducing both the maximum and average absolute correlation between exposures and confounders. Further, the weights can minimize bias in estimating the dose-response function in realistic data generating settings with moderate sample sizes. The simulations identified two key factors that affect performance. When the exposures have highly overlapping sets of confounders or large marginal correlation, the mvGPS method may generate extreme weights, resulting in smaller effective sample sizes and higher average root mean squared error. Trimming the weights improves performance in these situations. We suggest that in settings with a high degree of overlap in the confounders or moderate to large marginal exposure correlation, weights should be trimmed at $q = 0.99$.

While our method could in principle be extended to an arbitrary number of continuous exposures, we have confined attention in our simulations and application to the setting of two exposures. Assessing the joint effect of two interventions is a common scientific question, and we expect that there are many practical applications of the methods. The resulting

dose-response surface for bivariate exposures can be easily visualized and interpreted, which is key in practice. Further work is needed to explore performance for settings with more than two exposures. In particular, positivity and achieving adequate balance may be increasingly difficult as dimensions of exposure increase. For higher order exposures, dimension reduction techniques such as principal component analysis or manifold learning might be applied to transform the problem to a lower order continuous exposure space.

We applied the mvGPS method to evaluate the joint effectiveness of macro and micro intervention strategies used in childhood obesity programs on change in obesity prevalence among low-income preschool aged children. Due to non-random selection of participating clinics, there was significant imbalance on potential confounders as evidenced by large absolute maximum and average correlations prior to weighting. The mvGPS method achieved superior balance compared to univariate alternatives, drastically reducing the maximum absolute correlation and the average absolute correlation. Estimates of the dose-response surface using weights from the mvGPS method differed substantially from the results of the unweighted surface. The results showed that the most effective intervention combination was higher levels of micro strategies and lower levels of macro strategies. However, our results should be interpreted with caution. As with other causal inference methods, all confounders must be adequately captured and modeled to produce unbiased estimates. Thus our estimated treatment effects may be biased due to unknown confounders. In particular, communities that received higher levels of macro and micro doses may have been inherently more difficult to change due to a complex interplay of community and personal factors not captured by our set of potential confounders. In addition, we have assumed that the potential outcomes of the change in obesity given macro and micro exposures are well-defined via SUTVA as discussed in Section 5.2. Specifically, SUTVA stipulates that multiple versions of the exposures do not exist. In our application, however, there are potentially different sets of interventions that can yield the same macro and micro exposure scores. In the presence of multiple versions of the exposures, the resulting potential outcomes may be unidentifiable.

Further, we had to estimate exposure to the interventions at the census tract level using various assumptions, which may have resulted in measurement error. Thus our application, while demonstrating the methods, has important limitations.

Our approach assumes that the exposures have a multivariate normal distribution. The multivariate normal distribution is particularly attractive for working in higher dimensions. In our case, it allows for the full conditionals used to generate weights in Equation 5.3.3 to be tractable univariate normal densities. Further, the asymptotics of the estimates are well behaved due to the central limit theorem. We note that it is common practice when using the generalized propensity score for continuous treatments in the univariate case to assume normality [Fong et al., 2018; Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Robins, 2000; Zhu et al., 2015]. However, this reliance on multivariate normality is an important limitation of the methodology, particularly in assessing its validity [Mecklin and Mundfrom, 2004]. Possible extensions include replacing the multivariate normal distribution with non-parametric or semi-parametric alternatives as has been done recently with univariate methods [Kennedy et al., 2017; Tübbicke, 2020; Vegetabile et al., 2020]. Other possible extensions include allowing for time-varying outcomes and exposures such has been recently proposed for CBGPS [Huffman and van Gameren, 2018]. Additionally, SUTVA might be relaxed to test for potential geographic interference [VanderWeele, 2008; Verbitsky-Savitz and Raudenbush, 2012].

TABLES & FIGURES

Table 1: Intervention Program Strategies

#	**Name**	**Group**	n (%)
1	Government Policies	Macro	1 (3%)
2	Institutional Polices	Macro	4 (12%)
3	Infrastructure Investments	Macro	3 (9%)
4	Business Practices	Macro	4 (12%)
5	Group Education	Micro	21 (66%)
6	Counseling	Micro	14 (44%)
7	Health Communication	Micro	17 (53%)
8	Home Visitation	Micro	8 (25%)
9	Screening and Referral	Micro	14 (44%)

Each intervention program was classified based on the strategies that were implemented. Programs could use multiple strategies, e.g., Group Education and Counseling. Strategies are categorized as either micro and macro based on whether they directly targeted individuals or the population at large. The final column represents how many of the 32 programs used that particular strategy.

Table 2: Coefficients for Simulation Scenario M1: No Common Confounding

	Param.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	D_1	D_2
D_1	$\boldsymbol{\beta}_1^T$	1	0.5	0.25	0.1	0.75	0	0	0	0	0	-	-
D_2	$\boldsymbol{\beta}_2^T$	0	0	0	0	0	1	0.5	0.25	0.1	0.75	-	-
Y	$\boldsymbol{\alpha}^T$	0	0.5	0	1	0	0.2	0	0	1	0	1	1

All values of the covariates enter in each equation linearly with the respective coefficients shown in the table. In this scenario, covariates $X_1 - X_5$ are associated with exposure D_1; however, only X_2 and X_4 are true confounders also associated with Y. Similarly, covariates $X_6 - X_{10}$ are associated with exposure D_2; however, only X_6 and X_9 are true confounders also associated with Y. There are no common confounders in this model. Each exposure has a true treatment effect coefficient of 1.

Table 3: Coefficients for Simulation Scenario M2: Partially Common Confounding

	Param.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	D_1	D_2
D_1	$\boldsymbol{\beta}_1^T$	0	0	1	0.5	0.25	0.1	0.75	0	0	0	-	-
D_2	$\boldsymbol{\beta}_2^T$	0	0	0	0	1	0.5	0.25	0.1	0.75	0	-	-
Y	$\boldsymbol{\alpha}^T$	0	0	0.5	0	0	1	0.2	0	1	0	1	1

All values of the covariates enter in each equation linearly with the respective coefficients shown in the table. In this scenario, covariates $X_3 - X_7$ are associated with exposure D_1; however, only X_3, X_6, and X_7 are true confounders also associated with Y. Similarly, covariates $X_5 - X_9$ are associated with exposure D_2; however, only X_6 X_7, and X_9 are true confounders also associated with Y. Common confounders of exposure D_1 and D_2 are X_6 and X_7. Confounder of exposure D_1 only is X_3, and X_9 is a confounder of D_2 only. Each exposure has a true treatment effect coefficient of 1.

Table 4: Coefficients for Simulation Scenario M3: Common Confounding

	Param.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	D_1	D_2
D_1	$\boldsymbol{\beta}_1^T$	1	0.5	0.25	0.1	0.75	0	0	0	0	0	-	-
D_2	$\boldsymbol{\beta}_2^T$	0.8	0.8	0.05	0.4	0.55	0	0	0	0	0	-	-
Y	$\boldsymbol{\alpha}^T$	0.5	0	1	0.2	1	0	0	0	0	0	1	1

All values of the covariates enter in each equation linearly with the respective coefficients shown in the table. In this scenario, covariates $X_1 - X_5$ are associated with exposure D_1 and D_2; however, only X_1, X_3, X_4 and X_5 are true confounders also associated with Y. Common confounders of exposure D_1 and D_2 are X_1, X_3, X_4, and X_5. There are no confounders of D_1 or D_2 only in this model. Each exposure has a true treatment effect coefficient of 1.

Table 5: Covariate Balance

Max Abs. Corr.	**Avg. Abs. Corr.**	**ESS**	**Method**
0.12	0.04	637	mvGPS
0.18	0.08	580	CBGPS (Macro)
0.22	0.07	659	PS (Macro)
0.25	0.08	541	Entropy (Macro)
0.35	0.12	779	Entropy (Micro)
0.37	0.14	828	PS (Micro)
0.49	0.16	679	CBGPS (Micro)
0.41	0.19	1079	Unweighted

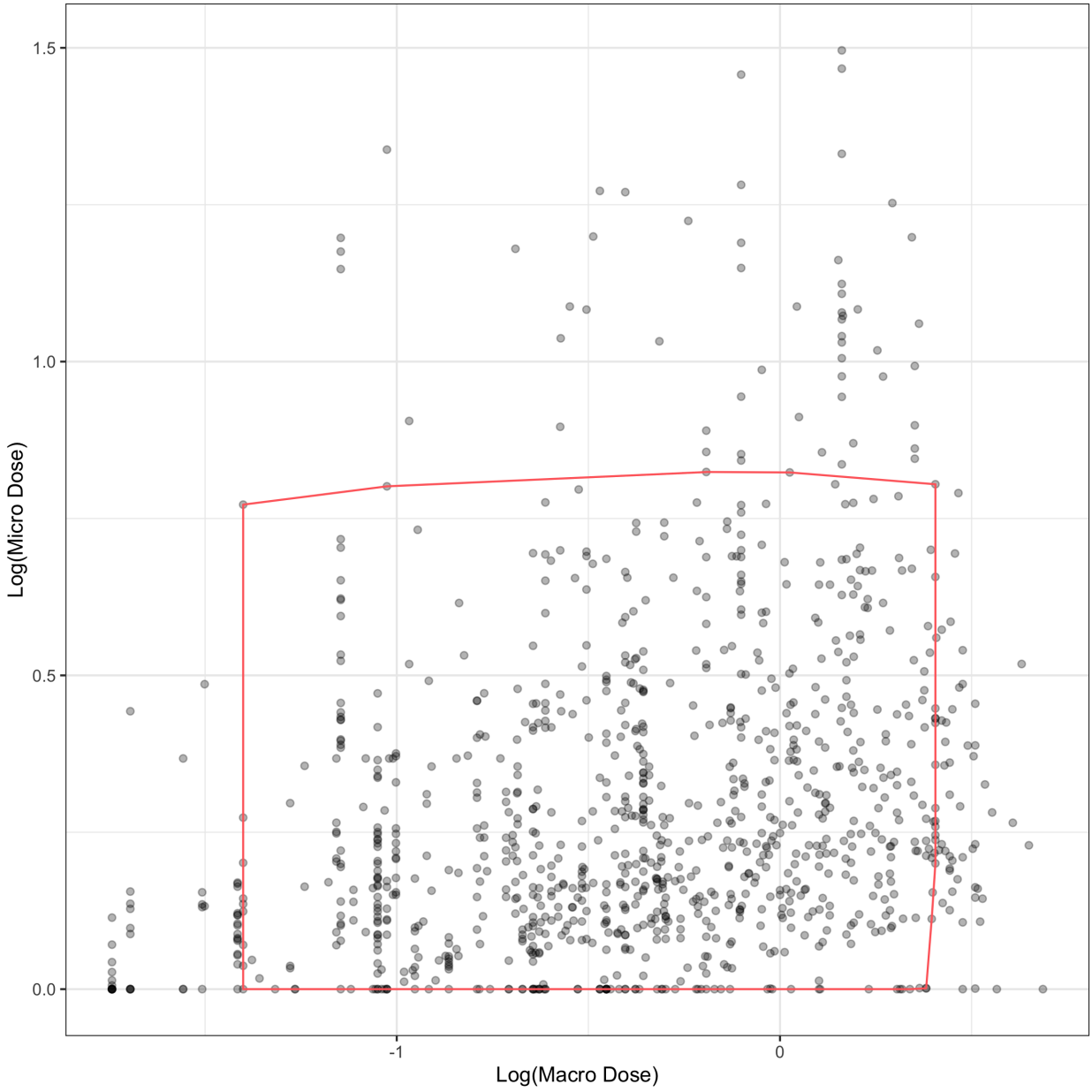
Assessing balance and effective sample size (ESS) using various weighted methods for the motivating example. Data used for estimating weights were restricted to $\mathcal{H}_{0.95}$. The univariate methods were applied separately to the two exposure metrics, macro or micro dose. The correlations are taken over both exposure metrics. The unweighted method represents the original values before applying weights. The weights for each method are trimmed using $q = 0.99$.

Table 6: Dose Response Quadrant Comparison

Macro Dose	**Micro Dose**	**Unweighted**	**mvGPS**
Low	Low	0.22 (-0.58, 1.01)	0.12 (-0.72, 0.97)
	High	-1.86 (-3.82, 0.10)	-3.51 (-5.40, -1.61)
High	Low	-1.80 (-2.51, -1.09)	-1.77 (-2.46, -1.08)
	High	-2.41 (-3.57, -1.25)	-1.22 (-2.27, -0.17)

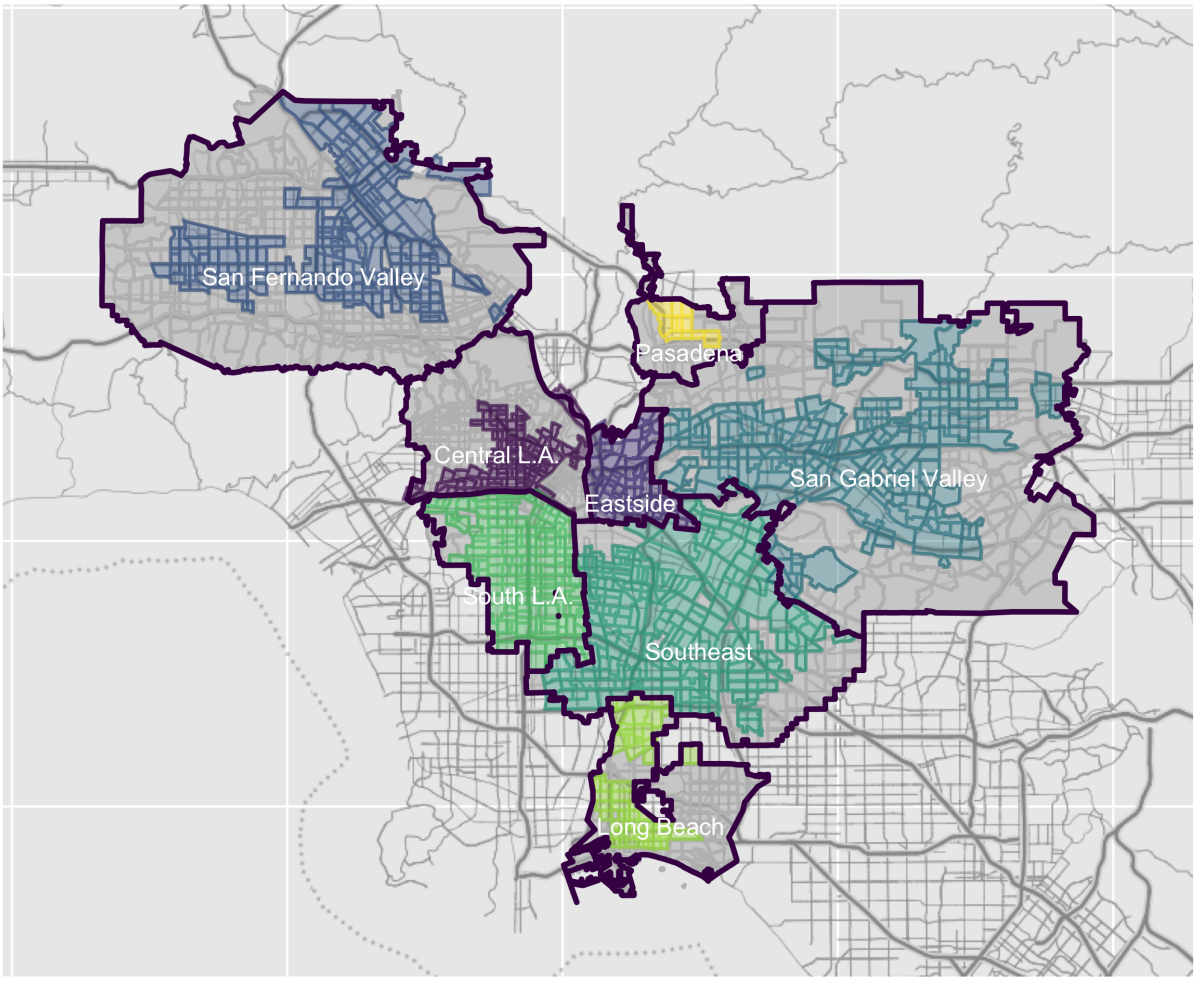
Point estimates and corresponding 95% confidence intervals for change in obesity prevalence at four quadrants of the trimmed convex hull inference region for the unweighted and multivariate generalized propensity score (mvGPS) methods. "High" corresponds to the 90^{th} percentile and "Low" corresponds to the 10^{th} percentile of the convex hull.

Figure 1: Joint Distribution of Macro and Micro Intervention Doses



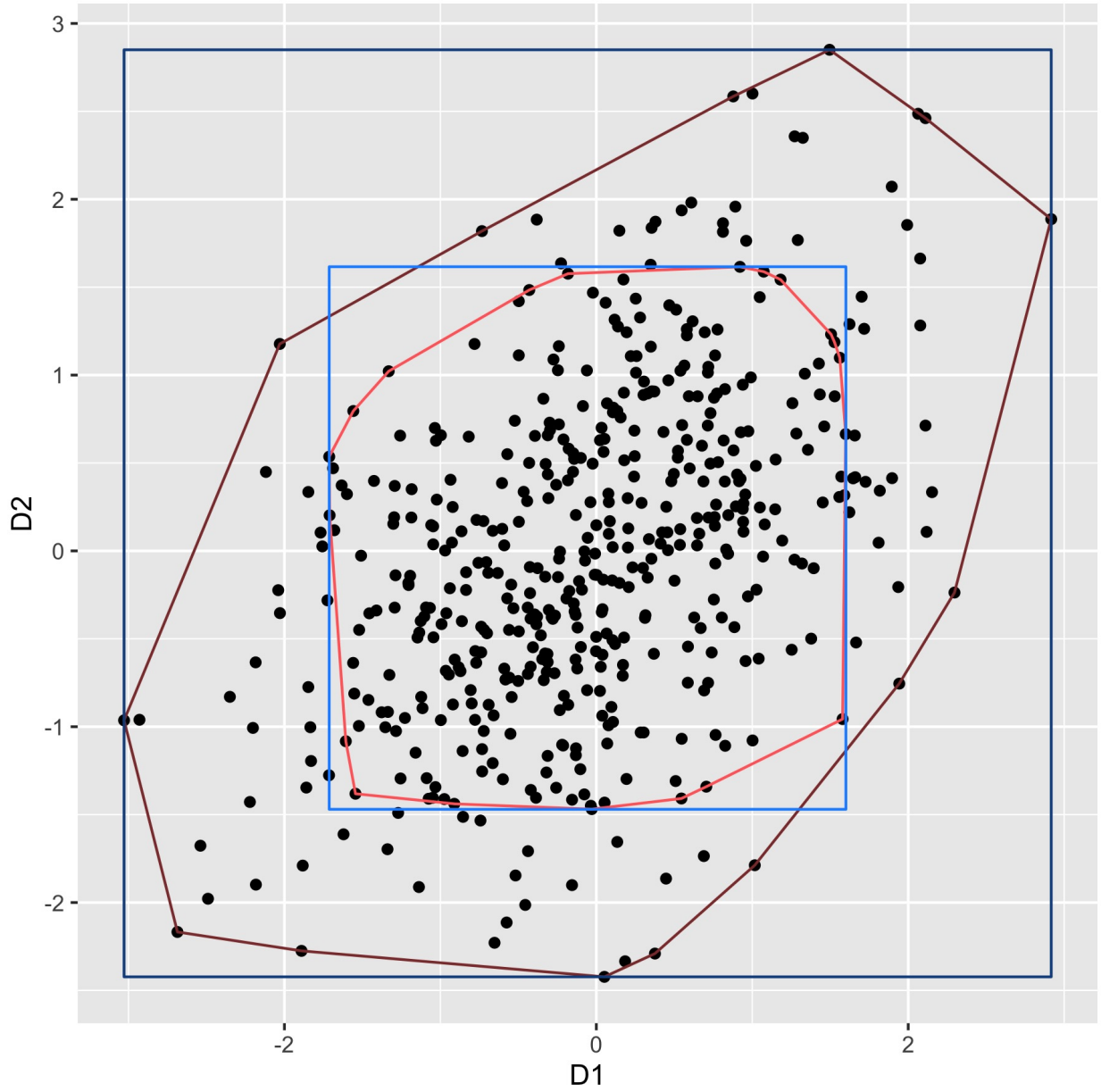
Observed values of log macro and micro exposure doses averaged over the study intervention period 2010-2016. Trimmed convex hull region with $q = 0.95$, i.e., $\mathcal{H}_{0.95}$, is shown in red

Figure 2: Units of Analysis: Census Tracts from 8 Regions



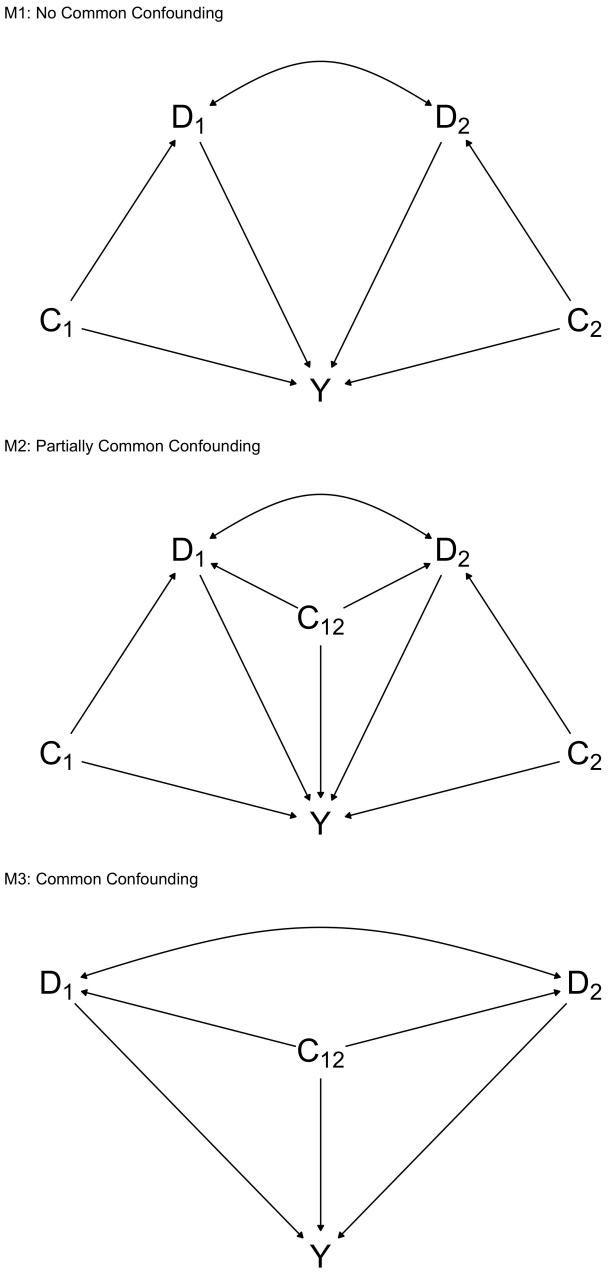
This map highlights the 8 regions in Los Angeles County that were targeted as part of the ECOSyS data collection, and the census tracts in these regions with at least 30 WIC-enrolled children over the period 2007-2016. This resulted in a total of $n = 1079$ census tracts that were the unit of analysis for estimating the effect of childhood obesity intervention programs by WIC.

Figure 3: Defining Estimable Region with Bivariate Exposure



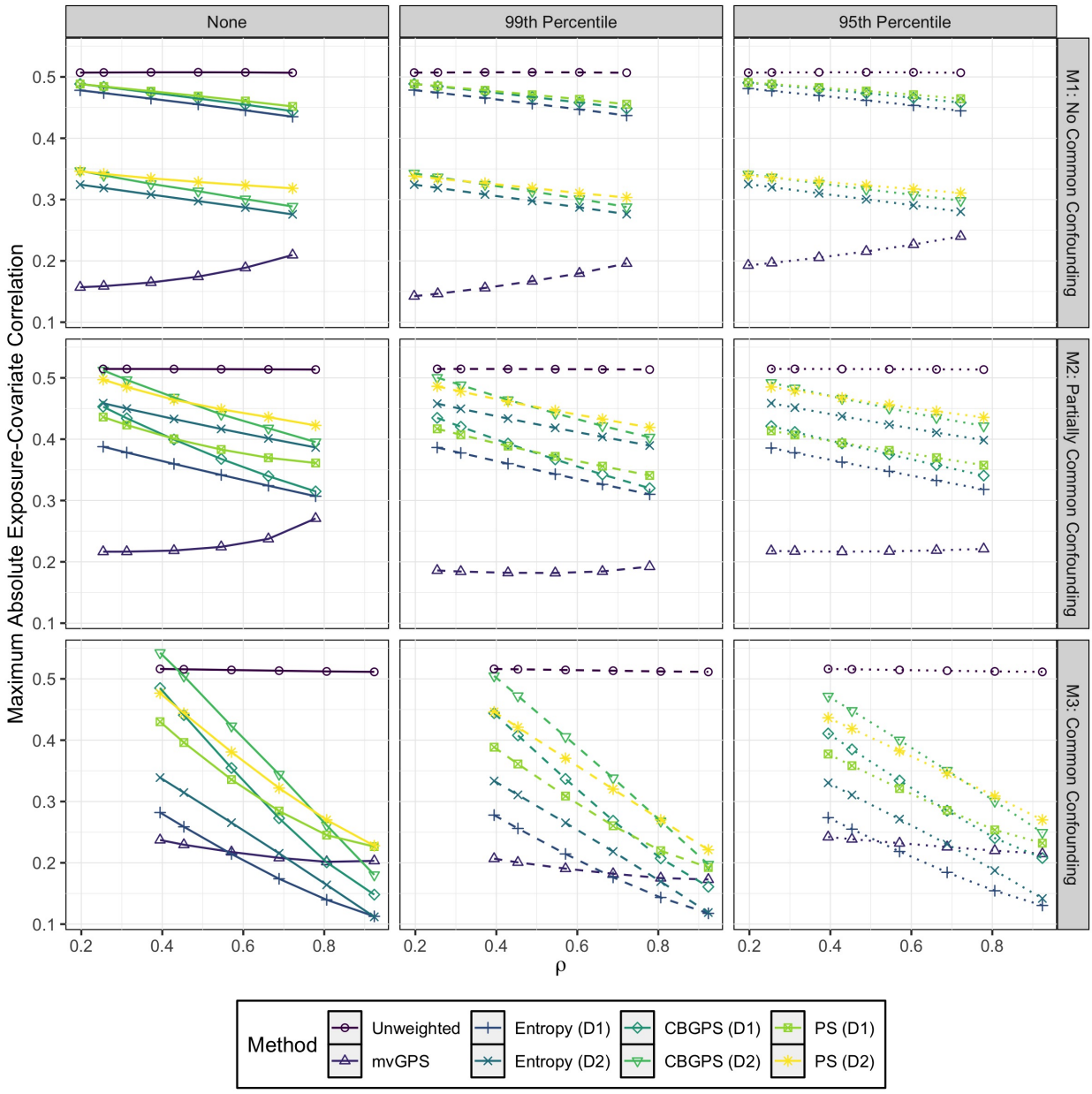
Sample of $n = 500$ units drawn from population where $D_1 \sim N(0,1)$, $D_2 \sim N(0,1)$, $\text{Cov}(D_1, D_2) = 0.5$. Region defined by dark blue box corresponds to $\mathcal{G}$ while region defined in dark red represents $\mathcal{H}$. Trimmed regions are also shown for $q = 0.95$ with the light blue box corresponds to $\mathcal{G}_{0.95}$ while the region defined in light red represents $\mathcal{H}_{0.95}$

Figure 4: Simulation Scenarios



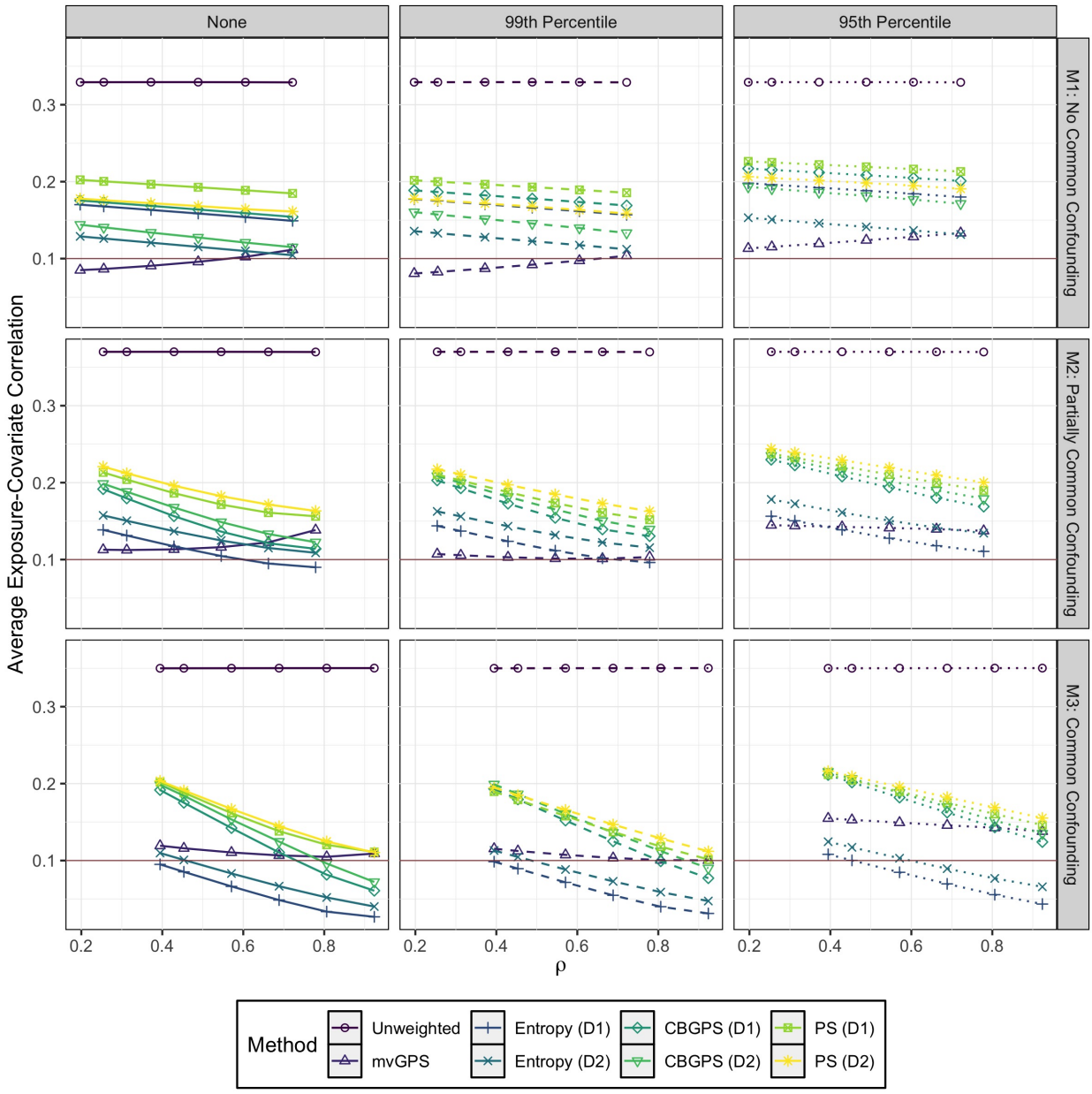
Directed acyclic graphs for each of the three data generating simulation scenarios. D_1 and D_2 are continuous exposure measures and Y is the outcome of interest. C_1 and C_2 represent confounder sets that are specific to exposures D_1 and D_2, while C_{12} represents a confounder set common to both exposures.

Figure 5: Assessing Covariate Balance: Maximum Absolute Exposure-Covariate Correlation



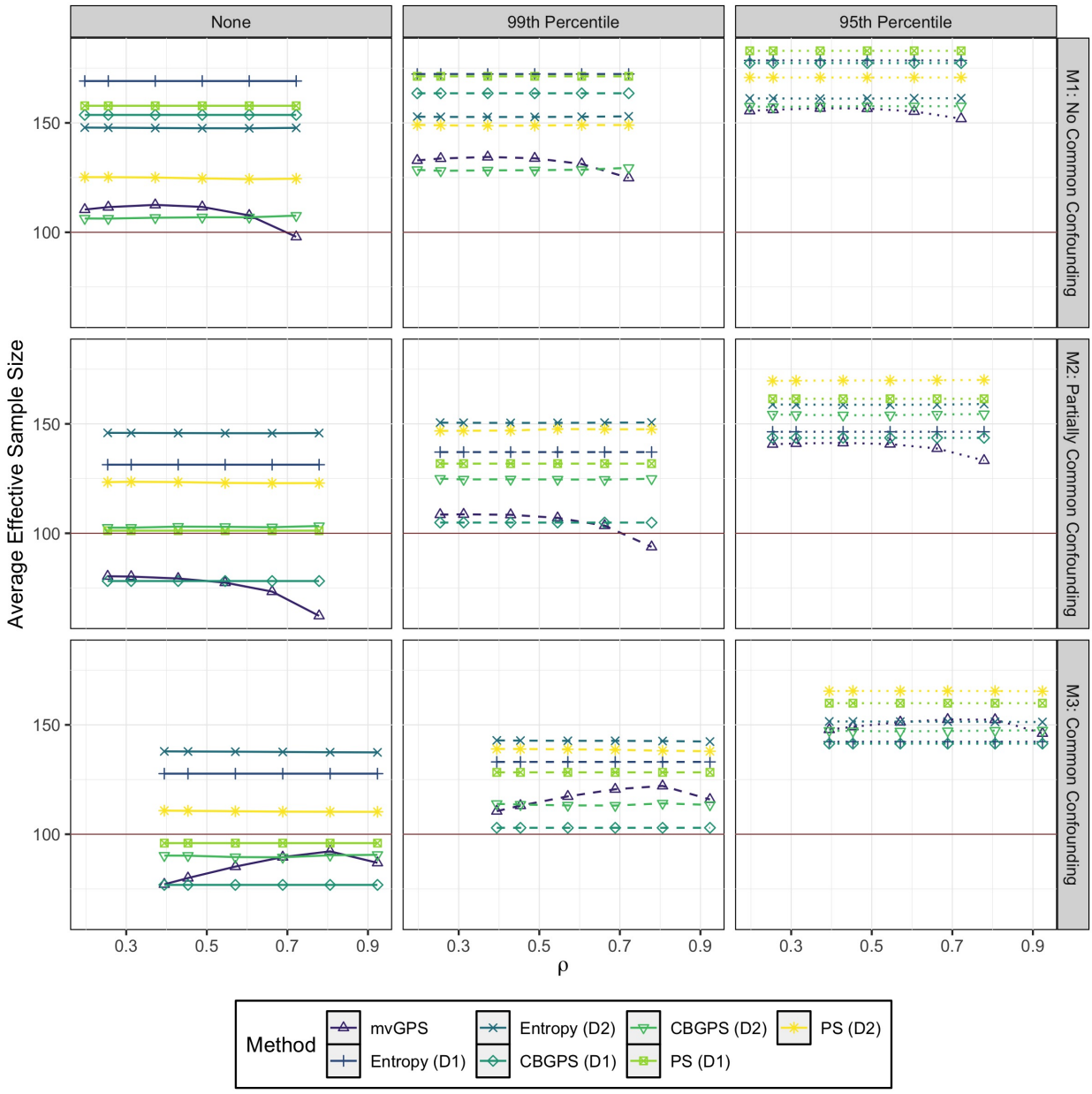
Rows correspond to the three simulation scenarios, M1, M2 and M3, and each column corresponds to quantiles used for weight trimming. The y-axis is the average maximum absolute exposure-covariate correlation for $n = 200$ from $B = 1000$ repetitions. This maximum is taken across both exposure values, D_1 and D_2. The x-axis, ρ, is the marginal correlation of the exposures. For univariate methods, weights were generated twice, once for each exposure variable.

Figure 6: Assessing Covariate Balance: Average Absolute Exposure-Covariate Correlation



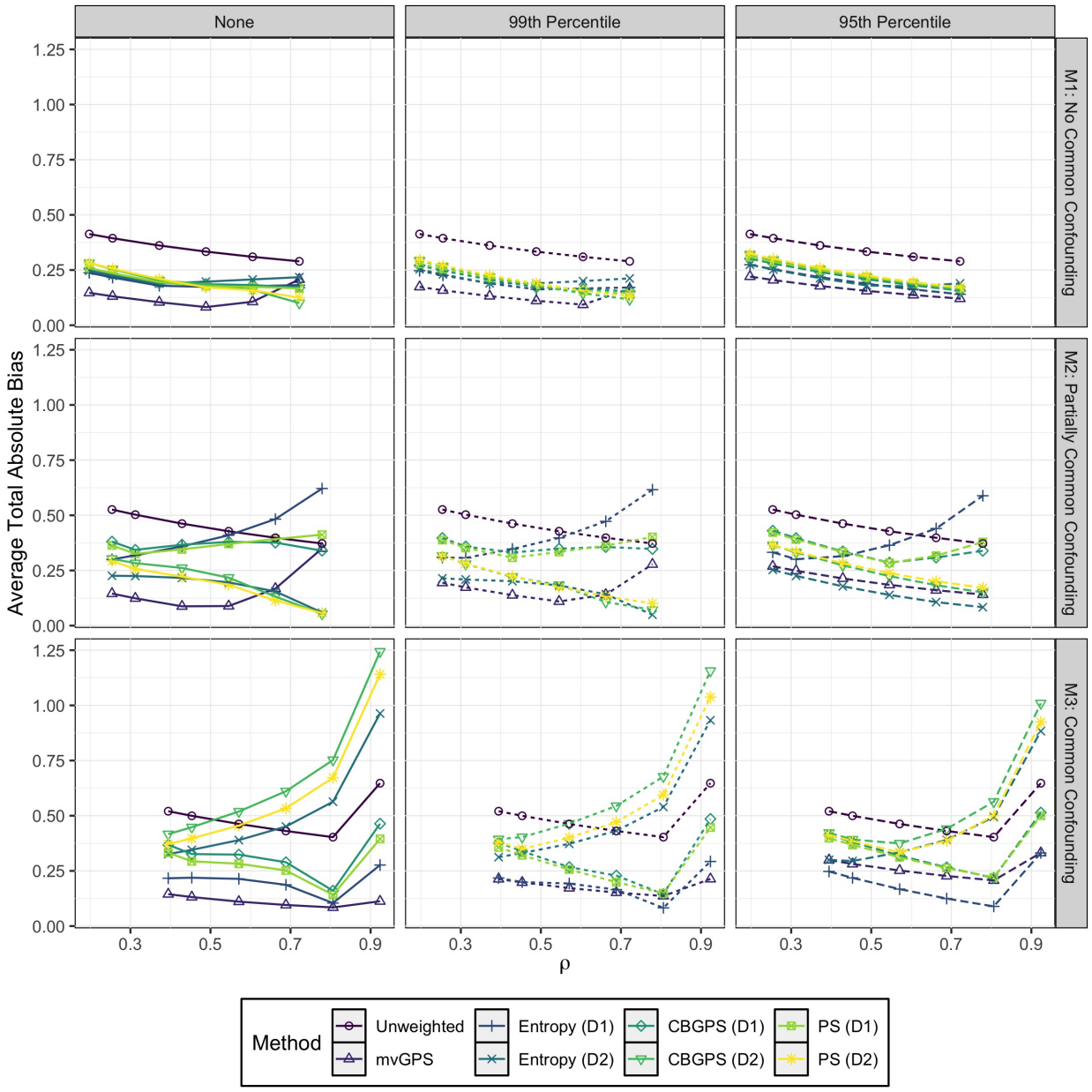
Rows correspond to the three simulation scenarios, M1, M2 and M3, and each column corresponds to quantiles used for weight trimming. The y-axis is the average absolute exposure-covariate correlation for $n = 200$ from $B = 1000$ repetitions. This average is taken across both exposure values, D_1 and D_2. The x-axis, ρ, is the marginal correlation of the exposures. For univariate methods, weights were generated twice, once for each exposure variable. The red line corresponds to an average value of 0.1, which is often used as a benchmark for sufficient balance.

Figure 7: Effective Sample Size



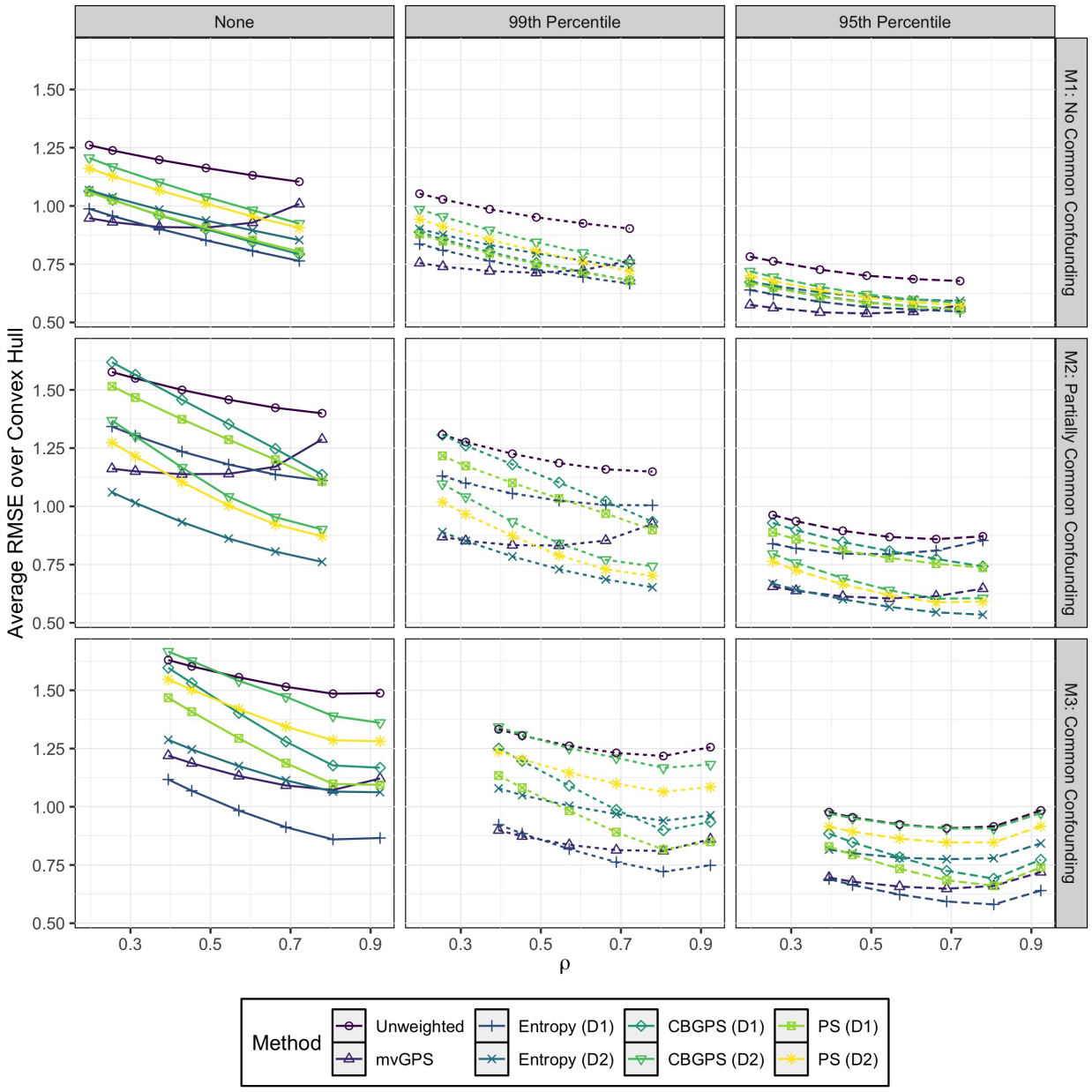
Rows correspond to the three simulation scenarios, M1, M2 and M3, and each column corresponds to quantiles used for weight trimming. The y-axis is the average effective sample size, $(\Sigma_i w_i)^2/\Sigma_i w_i^2$, for $n = 200$ from $B = 1000$ repetitions. The x-axis, ρ, is the marginal correlation of the exposures. The red line corresponds to an effective sample size of 100 which is often a minimum desirable quantity for hypothesis testing and inference of the dose-response model.

Figure 8: Outcome Modeling Performance Metric: Average Total Absolute Bias



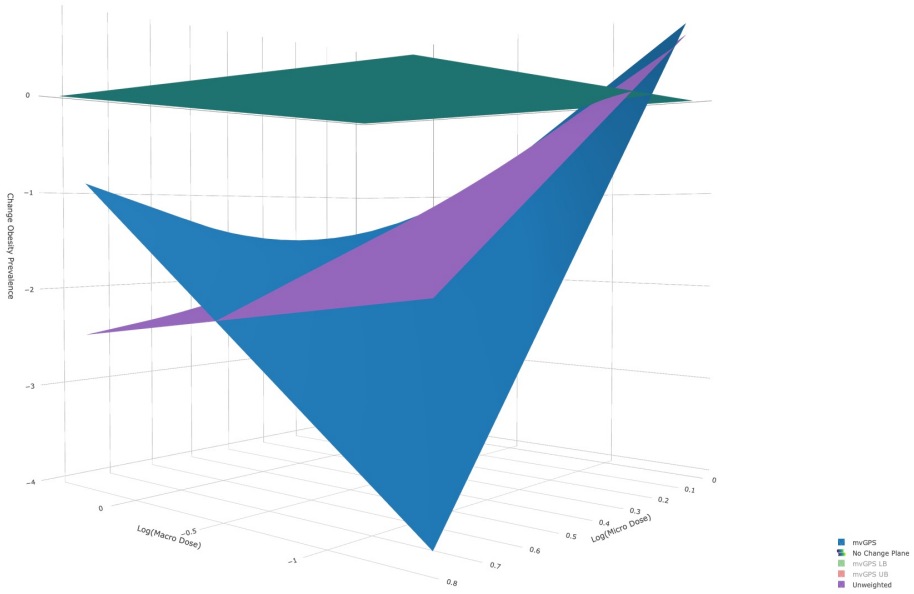
Rows correspond to the three simulation scenarios, M1, M2 and M3, and each column corresponds to quantiles used for weight trimming. The y-axis is the average total absolute bias, $\Sigma_j|\alpha_{D_j} - \hat{\alpha}_{D_j}|$, for $n = 200$ from $B = 1000$ repetitions. The x-axis, ρ, is the marginal correlation of the exposures. For univariate methods, weights were generated twice, once for each exposure variable.

Figure 9: Outcome Modeling Performance Metric: Average RMSE



Rows correspond to the three simulation scenarios, M1, M2 and M3, and each column corresponds to quantiles used for weight trimming. The y-axis is the average root mean squared error (RMSE) for 500 points sampled on a convex hull H_q grid for $n = 200$ from $B = 1000$ repetitions. The x-axis, ρ, is the marginal correlation of the exposures. For univariate methods, weights were generated twice, once for each exposure variable.

Figure 10: Estimated Dose-Response Surface of Change in Obesity Prevalence as a Function of Macro and Micro Intervention Dose



Estimated dose-response surface of change in obesity prevalence as a function of log macro and micro dose, obtained using mvGPS weights and unweighted. The surface is restricted to the convex hull of observed bivariate exposure, $H_{0.95}$, shown in Figure 1 and points are sampled evenly along this grid. A reference plane of no change is included. For a 3D interactive version of the dose-response surface that includes lower and upper bound 95% confidence interval surfaces for the mvGPS method, visit https://williazo.github.io/resources/.

Appendix A

Details of Dose Construction

This appendix provides details as to how we quantified levels of exposure to different types of obesity intervention strategies from WIC intervention programs in order to understand their effect on changes in childhood obesity prevalence, with regard to our motivating application described in Chapter 2. The ultimate goal of dose construction was to estimate the levels of exposure to macro and micro intervention strategies at the census tract level. Figure A.1 provides an overview of this process. Each intervention program was associated with one or more WIC clinics that implemented the program. Each program was parsed as to which of nine possible intervention strategies it used. Some strategies are considered macro strategies while others were considered micro strategies, as shown in Table 1. Then a program specific reach score was applied to generate a continuous bivariate intervention dose index (IDI) for macro and micro strategies associated with that program at the corresponding clinics. These IDIs were then mapped from clinics to census tracts where WIC-participating children lived using catchment areas. Any census tract whose boundary fell within the catchment area of a clinic was assigned the corresponding macro and micro dose associated with that clinic. Refinements to this calculation were made to account for distances between clinics and census tracts with overlapping catchment areas. Once this process was completed for each program by year, the exposure doses were then aggregated across all WIC programs and years, returning one bivariate measure of exposure for each census tract during the intervention period.

A.1 Intervention Dose Index

The first step in the dose construction as shown in Figure A.1 was to generate the macro and micro intervention dose indices for each intervention program. We used in part the "community intervention dose index" concept described in Wang et al. [2018]. The authors define an intervention dose index (IDI) for each of the nine intervention strategies as

$$IDI = SS \times RS \times FS,$$

where $SS \in [1, 9]$ is a strength score based on a Delphi survey from subject matter experts, $RS \in [0, 1]$ is a reach score reflecting the percent of the target population reached by the program, and $FS \in [0, 1]$ is a fidelity score which is the degree to which the program was followed during implementation.

We assumed that all programs had $FS = 1$ because the data needed to assess program fidelity were not available. The reach score, RS, was estimated using interviews with WIC employees who were asked to rate the percentage of WIC clients that they believed the program reached. Finally, we did not use strength scores so that the corresponding dose-response estimates would be based strictly on data alone.

Our adapted version of the IDI returned a continuous dose for each strategy utilized by a program that corresponded to the reach score associated with the program. Thus each program had a vector of nine IDIs, one for each of the nine potential program strategies. As mentioned in Chapter 2, there was particular interest in estimating the joint effect of macro and micro strategies. Therefore, the strategy-specific doses were summed by strategy type (macro versus micro) to yield a continuous bivariate dose for each program.

A.2 Catchment Area

Using this algorithm, we obtained a bivariate measure of exposure for each intervention program. Each program was associated with one or more WIC clinics. However, the outcome

of interest, obesity prevalence, was available at the census tract level. To assign exposures to census tracts, we constructed catchment areas for each clinic. A clinic's catchment area was defined as a circle centered at the clinic's latitude and longitude with a radius such that a certain percent of the clients served by the clinic resided within the catchment area. Any census tract whose boundary intersected the circle was assumed to be in the clinic's catchment area. For macro strategies, the radius was set to encompass 80% of the clients served and for micro strategies, it was set to encompass 50%. This reflects a belief that macro strategies have potential for wider geographic impact than micro strategies.

A.2.1 Refinements

Three refinements were applied when converting exposure levels to census tracts using catchment areas: radius truncation, micro exposure inverse distance weighting, and overlapping catchment area adjustment.

The first refinement was to truncate the clinic radii. Some smaller clinics served families that were highly dispersed geographically, perhaps because the families moved but preferred to travel to their regular clinic to receive services. The unadjusted radii for such clinics was sometimes quite large. To prevent such clinics from having out-sized impacts, the radii for macro and micro catchment areas were truncated at the 90^{th} percentile.

Another refinement was applied to micro strategies only. Because micro intervention strategies target individuals, those who live closer to a clinic offering micro interventions may be more likely to benefit than those who live further away. To reflect this belief, micro exposure for a particular clinic and catchment area was weighted by the relative distance between the clinic and the centroid of the census tract. The weights were defined as $w = d_0/d$ where d_0 represents the minimum distance among all census tracts within the micro catchment area of the clinic and d is the distance between a particular census tract and clinic. Constructing weights in this manner ensured that census tracts close to the clinic have $w \approx 1$, that $0 < w < 1$ for all other census tract units, and $w \to 0$ as the $d \to \infty$.

Weights were constructed separately for each clinic that participated in a program. We then multiplied these weights by micro exposure dose for each clinic participating in the program, yielding an inverse distance weighted micro exposure dose.

The final refinement addressed overlapping catchment areas. Certain programs were implemented at a large number of clinics and the catchment areas for the clinics overlapped. In such regions, children have increased potential to receive the intervention. To reflect this, we multiplied the macro exposure dose of the program by the number of overlapping catchment areas. For micro exposures, we took the average weight of the census tract for all overlapping catchment areas and then multiplied by the number of overlapping catchment areas. Figure A.2 provides an example of how micro weighting and overlapping adjustments are performed.

Figure A.3 shows the final catchment areas by clinic for micro and macro strategies after applying these refinements.

A.3 Aggregation

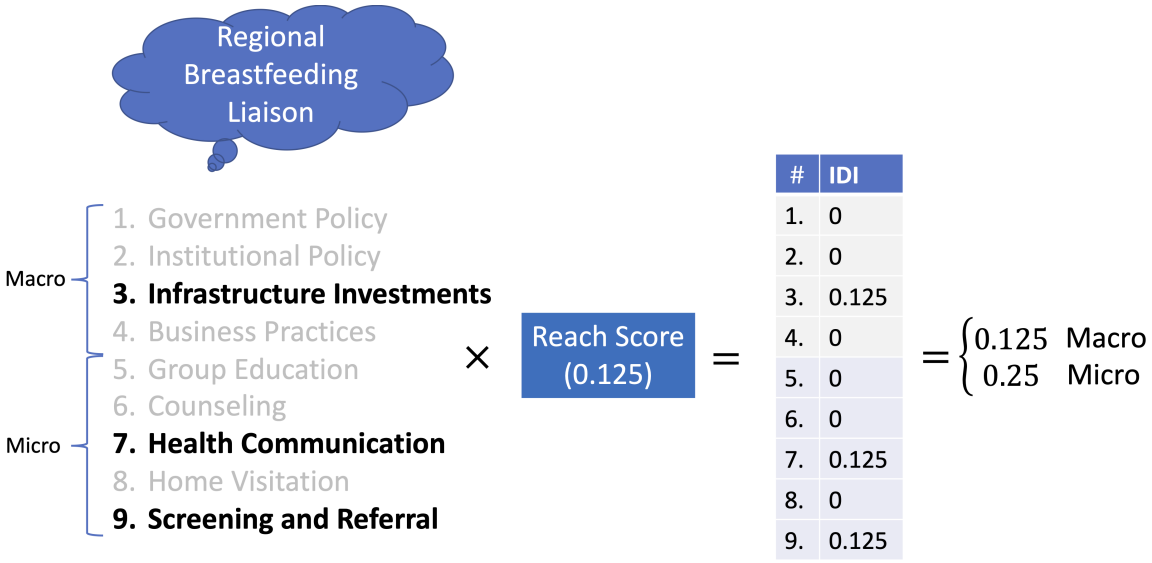
The final step in the dose estimation process was to aggregate exposures at the census tract level. We computed macro and micro exposures for all programs at the corresponding census tracts by year as described above from 2010-2016. For each census tract and year, we had a dose matrix with two columns, for macro and micro exposures, and 32 rows that corresponded to the total number of intervention programs. If a program was not implemented in that year or in that census tract, it would have a macro and micro exposure dose equal to zero. We then summed exposure doses across the number of programs, i.e., the rows of the matrix, to create a bivariate measure of total macro and micro exposure per census tract by year. We then averaged the macro and micro exposures at a census tract across years. As mentioned in Chapter 2, the resultant exposures were log transformed due to skewness. The final average log transformed macro and micro exposure doses for the population of census tracts are

shown in Figure A.4. We can see that micro doses were much more localized with a few high density regions while macro doses were much more evenly distributed throughout the population.

A.4 Limitations

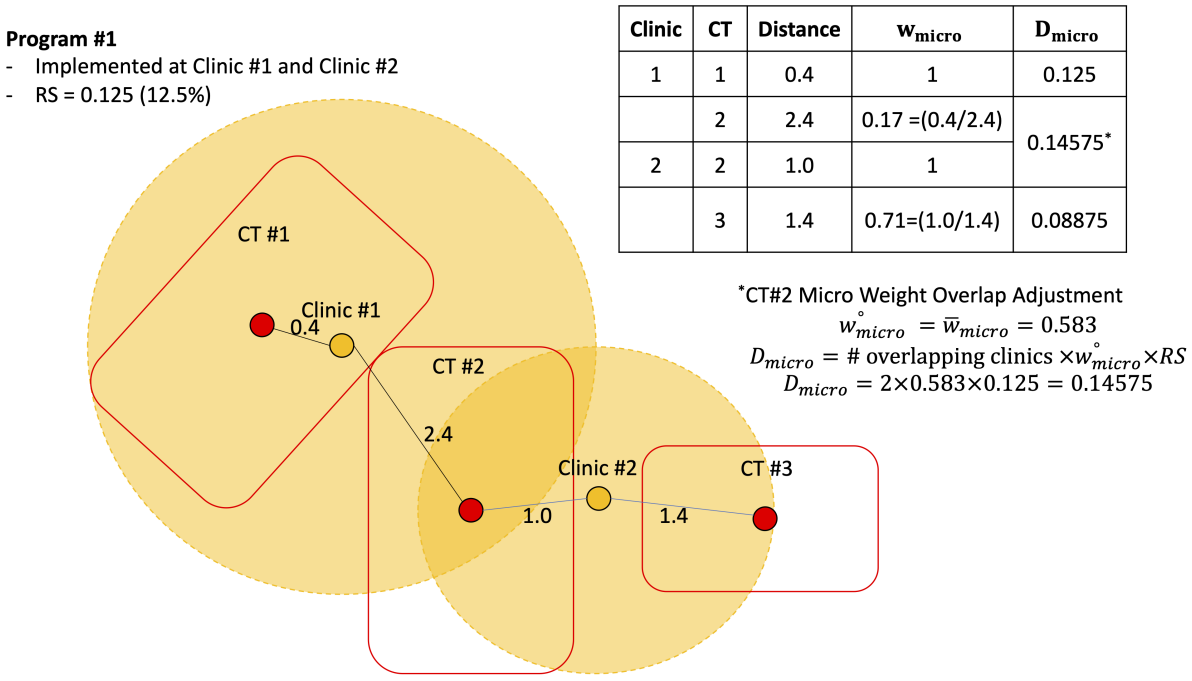
There are several major limitations of the dose estimation process. One limitation is the circular shape of the catchment area, which assumes that the impact and reach of the clinic extends uniformly in all directions when in reality these shapes are likely to be irregular. Another limitation is the lack of fidelity scores for the intervention programs; by ignoring fidelity, we assume that all programs were implemented equally well. Finally, reliance on the catchment area approach to map clinics to census tracts is a major limitation but was necessary because we did not have records of client participation.

Figure A.1: Overview of Dose Construction



The process of converting an intervention program into continuous exposures. In this example the intervention program, "Regional Breastfeeding Liaison", uses one macro and two micro intervention strategies with a program reach score of 12.5%. An indicator for strategy utilization is multiplies by the reach score to produce the corresponding intervention dose index (IDI) for all nine strategies. These IDIs are then aggregated to total macro and micro strategy dose.

Figure A.2: Example of Dose Refinement



Clinic	CT	Distance	w_{micro}	D_{micro}
1	1	0.4	1	0.125
	2	2.4	0.17 =(0.4/2.4)	0.14575*
2	2	1.0	1	
	3	1.4	0.71=(1.0/1.4)	0.08875

*CT#2 Micro Weight Overlap Adjustment

$w^{\circ}_{micro} = \bar{w}_{micro} = 0.583$

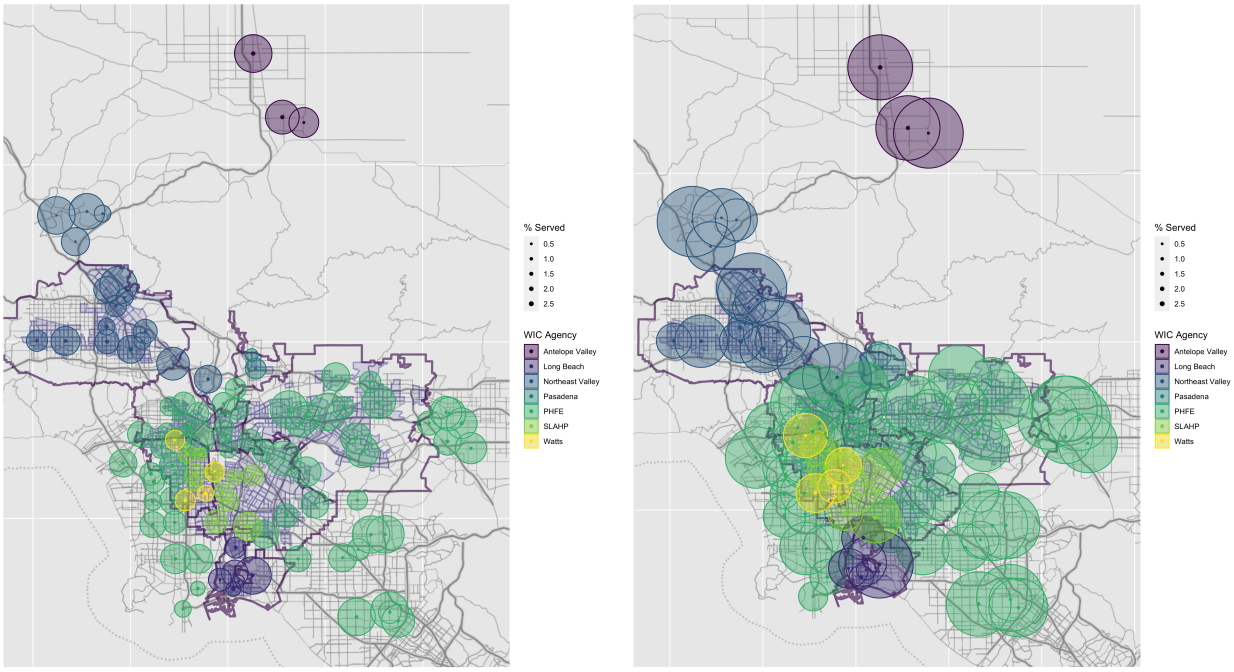
D_{micro} = # overlapping clinics $\times w^{\circ}_{micro} \times RS$

$D_{micro} = 2 \times 0.583 \times 0.125 = 0.14575$

CT=census tract

Example showing the process of refining the exposure dose for an intervention program using micro weights and adjusting for census tracts with overlapping catchment areas. We assume there exists an intervention program (program #1) implemented at clinics #1 and #2 with a program reach score of 0.125. Each clinic has a specific catchment area shown as the light orange shaded circles. The centroid of census tract #2 lies in the catchment area of both clinic #1 and #2. Therefore, the weight for this unit is the average weight for each clinic, 0.583. The micro dose is then constructed by multiplying the number of overlapping catchment areas with the average weight and the estimated reach score. Census tracts #1 and #3 would not require any overlapping adjustment and their micro dose would be the product of their weight and the reach score.

Figure A.3: Catchment Area for WIC Clinics

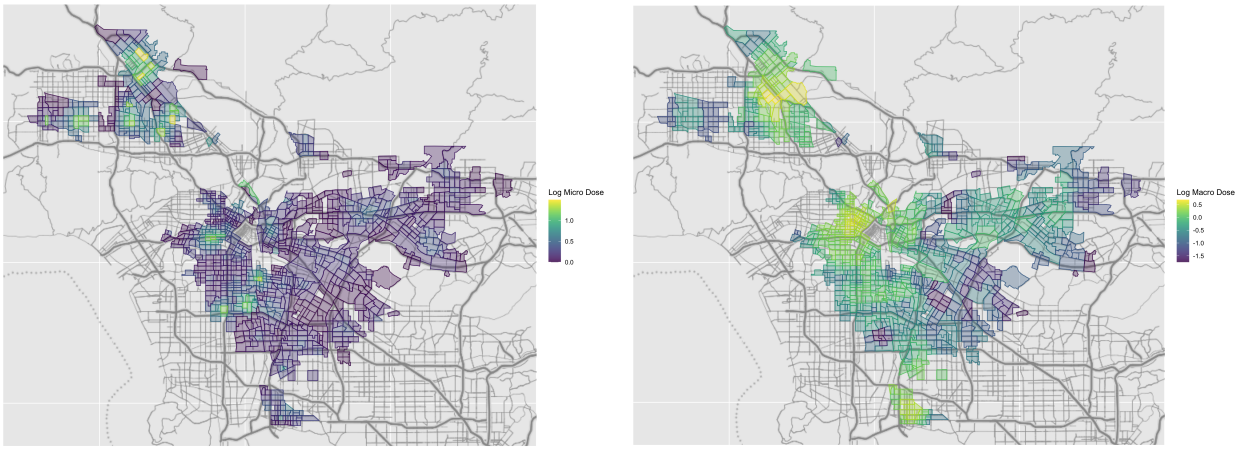


(a) Micro catchment areas

(b) Macro catchment areas

Catchment area for each WIC clinic by strategy group overlaid on the census tracts and regions used in the analysis from Figure 2. Clinics are colored corresponding to their respective WIC agencies, and the size of each point represents the relative number of WIC families that are served at the clinic on average over the period of 2010-2016.

Figure A.4: Geographic Distribution of Macro and Micro Dose



(a) Micro exposure dose

(b) Macro exposure dose

Average micro and macro log exposure dose over the period of 2010-2016 for the $n = 1079$ census tracts that from Figure 2.

Bibliography

Anderson, C. E., C. M. Crespi, M. C. Wang, S. E. Whaley, and M. P. Chaparro (2020): "The neighborhood food environment modifies the effect of the 2009 WIC food package change on childhood obesity in Los Angeles County, California," *BMC Public Health*, 20, 1–11.

Austin, P. C. (2019): "Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures," *Statistical Methods in Medical Research*, 28, 1365–1377.

Chaparro, M. P., C. M. Crespi, C. E. Anderson, M. C. Wang, and S. E. Whaley (2019): "The 2009 Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) food package change and children's growth trajectories and obesity in Los Angeles County," *Am J Clin Nutr*, 109, 1414–1421.

Chaparro, M. P., S. E. Whaley, C. M. Crespi, M. Koleilat, T. Z. Nobari, E. Seto, and M. C. Wang (2014): "Influences of the neighbourhood food environment on adiposity of low-income preschool-aged children in Los Angeles County: A longitudinal study," *J Epidemiol Community Health*, 68, 1027–1033.

Chazelle, B. (1993): "An optimal convex hull algorithm in any fixed dimension," *Discrete & Computational Geometry*, 10, 377–409.

Cochran, W. G. and D. B. Rubin (1973): "Controlling bias in observational studies: a review," *Sankhyā: The Indian Journal of Statistics, Series A*, 35, 417–446.

Colombel, J. F., W. J. Sandborn, W. Reinisch, G. J. Mantzaris, A. Kornbluth, D. Rachmilewitz, S. Lichtiger, G. D'Haens, R. H. Diamond, D. L. Broussard, K. L. Tang, C. J. van der Woude, and P. Rutgeerts (2010): "Infliximab, azathioprine, or combination therapy for Crohn's disease," *New England Journal of Medicine*, 362, 1383–1395.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Dehejia, R. H. and S. Wahba (2002): "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and Statistics*, 84, 151–161.

Diamond, A. and J. S. Sekhon (2013): "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95, 932–945.

Doyle, W. R. (2011): "Effect of increased academic momentum on transfer rates: an application of the generalized propensity score," *Economics of Education Review*, 30, 191–200.

Fan, J. and I. Gijbels (1996): *Local Polynomial Modeling and Its Applications*, London: Champan and Hall.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (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, 153–171.

Fong, C., C. Hazlett, and K. Imai (2018): "Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements," *Annals of Applied Statistics*, 12, 156–177.

Gautret, P., J.-C. Lagier, P. Parola, V. T. Hoang, L. Meddeb, M. Mailhe, B. Doudier, J. Courjon, V. Giordanengo, V. E. Vieira, H. T. Dupont, S. Honoré, P. Colson, E. Chabrière, B. La Scola, J.-M. Rolain, P. Brouqui, and D. Raoult (2020a): "Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial," *International Journal of Antimicrobial Agents*, URL https://doi.org/10.1016/j.ijantimicag.2020.105949.

Gautret, P., J.-C. Lagier, P. Parola, V. T. Hoang, L. Meddeb, J. Sevestre, M. Mailhe, B. Doudier, C. Aubry, S. Amrane, P. Seng, M. Hocquart, C. Eldin, J. Finance, V. E. Vieira, H. T. Tissot-Dupont, S. Honoré, A. Stein, M. Million, P. Colson, B. La Scola, V. Veit, A. Jacquier, J.-C. Deharo, M. Drancourt, P. E. Fournier, J.-M. Rolain, P. Brouqui, and D. Raoult (2020b): "Clinical and microbiological effect of a combination of hydroxychloroquine and azithromycin in 80 COVID-19 patients with at least a six-day follow up: A pilot observational study," *Travel Medicine and Infectious Disease*, 34, 101663, URL https://doi.org/10.1016/j.tmaid.2020.101663.

Gradman, A. H., J. N. Basile, B. L. Carter, and G. L. Bakris (2010): "Combination therapy in hypertension," *Journal of the American Society of Hypertension*, 4, 90–98.

Greifer, N. (2020): *WeightIt: Weighting for Covariate Balance in Observational Studies*, URL https://CRAN.R-project.org/package=WeightIt, R package version 0.9.0.

Greiner, D. J. and D. B. Rubin (2011): "Causal effects of perceived immutable characteristics," *Review of Economics and Statistics*, 93, 775–785.

Guo, S. and M. W. Fraser (2014): *Propensity Score Analysis: Statistical Methods and Applications*, Los Angeles: Sage, 2nd edition.

Hade, E. M. and B. Lu (2014): "Bias associated with using the estimated propensity score as a regression covariate," *Statistics in Medicine*, 33, 74–87.

Hainmueller, J. (2012): "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Hillier, A., J. McLaughlin, C. C. Cannuscio, M. Chilton, S. Krasny, and A. Karpyn (2012): "The impact of WIC food package changes on access to healthful food in 2 low-income urban neighborhoods," *Journal of Nutrition Education and Behavior*, 44, 210–216.

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

Hirano, K. and G. W. Imbens (2004): "The propensity score with continuous treatments," in A. Gelman and X.-L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Hoboken, N.J.: John Wiley & Sons, 73–84.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

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

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

Huber, M., M. Lechner, and C. Wunsch (2013): "The performance of estimators based on the propensity score," *Journal of Econometrics*, 175, 1–21.

Huffman, C. and E. van Gameren (2018): "Covariate balancing inverse probability weights for time-varying continuous interventions," *Journal of Causal Inference*, 6, URL https://doi.org/10.1515/jci-2017-0002.

Imai, K. and D. A. Van Dyk (2004): "Causal inference with general treatment regimes: Generalizing the propensity score," *Journal of the American Statistical Association*, 99, 854–866.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

Jain, R. K. (2001): "Normalizing tumor vasculature with anti-angiogenic therapy: A new paradigm for combination therapy," *Nature Medicine*, 7, 987–989.

Jiang, M. and E. M. Foster (2013): "Duration of breastfeeding and childhood obesity: a generalized propensity score approach," *Health Services Research*, 48, 628–651.

Kang, J. D. and J. L. Schafer (2007): "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22, 523–539.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Non-parametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society: Series B*, 79, 1229–1245.

Khuri, A. I. and S. Mukhopadhyay (2010): "Response surface methodology," *WIREs Computational Statistics*, 2, 128–149.

Kish, L. (1965): *Survey Sampling*, New York: John Wiley & Sons.

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

Kreif, N., R. Grieve, I. Díaz, and D. Harrison (2015): "Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury," *Health Economics*, 24, 1213–1228.

Lechner, M. (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," *Econometric Evaluation of Labour Market Policies*, 43–58.

Lee, B. K., J. Lessler, and E. A. Stuart (2011): "Weight trimming and propensity score weighting," *PLOS ONE*, 6, URL https://doi.org/10.1371/journal.pone.0018174.

Little, R. J. and D. B. Rubin (2014): *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, 2nd edition.

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," in P. Zarembka, ed., *Frontiers in Econometrics*, Cambridge: Academic Press, 105–142.

McFadden, D. (1978): "Modelling the choice of residential locations," in A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds., *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, 72–77.

Mecklin, C. J. and D. J. Mundfrom (2004): "An appraisal and bibliography of tests for multivariate normality," *International Statistical Review*, 72, 123–138.

Neyman, J. (1923): "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Roczniki Nauk Rolniczych Tom X* [in Polish]; translated in *Statistical Science*, 5, 465–480.

Nobari, T. Z., M.-C. Wang, M. P. Chaparro, C. M. Crespi, M. Koleilat, and S. E. Whaley (2013): "Immigrant enclaves and obesity in preschool-aged children in Los Angeles County," *Social Science & Medicine*, 92, 1–8.

Nobari, T. Z., S. E. Whaley, C. M. Crespi, M. L. Prelip, and M. C. Wang (2018a): "Widening socio-economic disparities in early childhood obesity in Los Angeles County after the great recession." *Public Health Nutrition*, 21, 2301–2310.

Nobari, T. Z., S. E. Whaley, M. L. Prelip, C. M. Crespi, and M. C. Wang (2018b): "Trends in socioeconomic disparities in obesity prevalence among low-income children aged 2–4 years in Los Angeles County, 2003–2014," *Childhood Obesity*, 14, 248–258.

Pearl, J. and D. Mackenzie (2018): *The Book of Why: The New Science of Cause and Effect*, New York: Basic Books.

Perelson, A. S., P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, and D. D. Ho (1997): "Decay characteristics of HIV-1-infected compartments during combination therapy," *Nature*, 387, 188–191.

PHFE WIC (2010): "WIC data 2003-2009: A report on low income families with young children in Los Angeles County," Technical report, Special Supplemental Nutrition Assistance Program for Women, Infants and Children.

R Core Team (2020): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Robins, J. M. (2000): "Marginal structural models versus structural nested models as tools for causal inference," in M. Halloran and D. Berry, eds., *Statistical Models in Epidemiology: The Environment and Clinical Trials*, New York, NY: Springer, 95–134.

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

Rosenbaum, P. R. (1987): "Model-based direct adjustment," *Journal of the American Statistical Association*, 82, 387–394.

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

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

Rosenkranz, R. R. and D. A. Dzewaltowski (2008): "Model of the home food environment pertaining to childhood obesity," *Nutrition Reviews*, 66, 123–140.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. B. (1980): "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

Rubin, D. B. (1986): "Comment: which ifs have causal answers?" *Journal of the American Statistical Association*, 81, 961–962.

Sanders, J. M., M. L. Monogue, T. Z. Jodlowski, and J. B. Cutrell (2020): "Pharmacologic treatments for coronavirus disease 2019 (COVID-19): A review," *JAMA*, 323, 1824–1836.

Shannon, C. E. (1948): "A mathematical theory of communication," *Bell System Technical Journal*, 27, 379–423.

Stebbing, J., A. Phelan, I. Griffin, C. Tucker, O. Oechsle, D. Smith, and P. Richardson (2020): "COVID-19: Combining antiviral and anti-inflammatory treatments," *The Lancet Infectious Diseases*, 20, 400–402.

Tübbicke, S. (2020): "Entropy balancing for continuous treatments," *arXiv preprint arXiv:2001.06281*.

United States Census Bureau (2020): "American community survey," URL https://www.census.gov/acs/www/data/data-tables-and-tools/, Last checked on July 30, 2020.

US Department of Agriculture, Food and Nutrition Service (2014): "Final rule: Revisions in the WIC food packages," URL https://www.fns.usda.gov/wic/fr-030414.

VanderWeele, T. J. (2008): "Ignorability and stability assumptions in neighborhood effects research," *Statistics in Medicine*, 27, 1934–1943.

Vegetabile, B. G., B. A. Griffin, D. L. Coffman, M. Cefalu, and D. F. McCaffrey (2020): "Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures," *arXiv preprint arXiv:2003.02938*.

Verbitsky-Savitz, N. and S. W. Raudenbush (2012): "Causal inference under interference in spatial settings: A case study evaluating community policing program in chicago," *Epidemiologic Methods*, 1, 107–130.

Walls & Associates (2013): "National Establishment Time-Series (NETS) database," URL http://youreconomy.org/profile/about.lasso, Last checked on August 22, 2020.

Wang, M. C., C. M. Crespi, L. H. Jiang, T. Nobari, H. Roper-Fingerhut, S. Rauzon, B. Robles, M. Blocklin, M. Davoudi, T. Kuo, K. E. MacLeod, E. Seto, S. Whaley, and M. Prelip (2018): "Developing an index of dose of exposure to early childhood obesity community interventions," *Preventive Medicine*, 111, 135–141.

Wang, M. C., A. A. Gonzalez, L. D. Ritchie, and M. A. Winkleby (2006): "The neighborhood food environment: Sources of historical data on retail food stores," *International Journal of Behavioral Nutrition and Physical Activity*, 3, 15.

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