Finding common support and assessing matching methods for causal inference

by

Sharif Mahmood

M.S., University of Dhaka, 2009

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Abstract

This dissertation presents an approach to assess and validate causal inference tools to estimate the causal effect of a treatment. Finding treatment effects in observational studies is complicated by the need to control for confounders. Common approaches for controlling include using prognostically important covariates to form groups of similar units containing both treatment and control units or modeling responses through interpolation. This dissertation proposes a series of new, computationally efficient methods to improve the analysis of observational studies.

Treatment effects are only reliably estimated for a subpopulation under which a common support assumption holds—one in which treatment and control covariate spaces overlap. Given a distance metric measuring dissimilarity between units, a graph theory is used to find common support. An adjacency graph is constructed where edges are drawn between similar treated and control units to determine regions of common support by finding the largest connected components (LCC) of this graph. The results show that LCC improves on existing methods by efficiently constructing regions that preserve clustering in the data while ensuring interpretability of the region through the distance metric.

This approach is extended to propose a new matching method called largest caliper matching (LCM). LCM is a version of cardinality matching—a type of matching used to maximize the number of units in an observational study under a covariate balance constraint between treatment groups. While traditional cardinality matching is an NP-hard, LCM can be completed in polynomial time. The performance of LCM with other five popular matching methods are shown through a series of Monte Carlo simulations. The performance of the simulations is measured by the bias, empirical standard deviation and the mean square error of the estimates under different treatment prevalence and different distributions of covariates. The formed matched samples improve estimation of the population treatment

effect in a wide range of settings, and suggest cases in which certain matching algorithms perform better than others. Finally, this dissertation presents an application of LCC and matching methods on a study of the effectiveness of right heart catheterization (RHC) and find that clinical outcomes are significantly worse for patients that undergo RHC.

Finding common support and assessing matching methods for causal inference

by

Sharif Mahmood

M.S., University of Dhaka, 2009

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Approved by:

Major Professor
Michael Higgins

Copyright

Abstract

This dissertation presents an approach to assess and validate causal inference tools to estimate the causal effect of a treatment. Finding treatment effects in observational studies is complicated by the need to control for confounders. Common approaches for controlling include using prognostically important covariates to form groups of similar units containing both treatment and control units or modeling responses through interpolation. This dissertation proposes a series of new, computationally efficient methods to improve the analysis of observational studies.

Treatment effects are only reliably estimated for a subpopulation under which a common support assumption holds—one in which treatment and control covariate spaces overlap. Given a distance metric measuring dissimilarity between units, a graph theory is used to find common support. An adjacency graph is constructed where edges are drawn between similar treated and control units to determine regions of common support by finding the largest connected components (LCC) of this graph. The results show that LCC improves on existing methods by efficiently constructing regions that preserve clustering in the data while ensuring interpretability of the region through the distance metric.

This approach is extended to propose a new matching method called largest caliper matching (LCM). LCM is a version of cardinality matching—a type of matching used to maximize the number of units in an observational study under a covariate balance constraint between treatment groups. While traditional cardinality matching is an NP-hard, LCM can be completed in polynomial time. The performance of LCM with other five popular matching methods are shown through a series of Monte Carlo simulations. The performance of the simulations is measured by the bias, empirical standard deviation and the mean square error of the estimates under different treatment prevalence and different distributions of covariates. The formed matched samples improve estimation of the population treatment

effect in a wide range of settings, and suggest cases in which certain matching algorithms perform better than others. Finally, this dissertation presents an application of LCC and matching methods on a study of the effectiveness of right heart catheterization (RHC) and find that clinical outcomes are significantly worse for patients that undergo RHC.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere gratitude and deepest appreciation to my supervisor: Dr. Michael Higgins. His mentorship, valuable suggestions regarding my research, review of my writing and constant inspiration greatly helped to advance my research training and to complete this dissertation. It truly has been an honor and a privilege to work with him.

It is my pleasure to express my admiration to Dr. Gary Gadbury, Dr. Christopher Vahl, Dr. Pietro Poggi-Corradini and Dr. William Hageman for their precious time to serve as my committee members, and their valuable comments and suggestions on improving the quality of my work. Many thanks to the faculty and staff of the department of statistics, staff of the Jardine and staff of the graduate school who made it a warm and enjoyable experience at K-State.

Finally, my special and affectionate thanks to my mother who has encouraged me all my life and my wife for her ceaseless support and devotion during my studies.

Dedication

Dedicated to

Soha and Zayaan

Chapter 1

Introduction

1.1 Introduction

Estimation of treatment effects in observational studies is complicated by the need for an appropriate model—one that adjusts for all important covariates and their interactions. Adjustment can weaken asymptotic precision if an appropriate model is not being considered (Lin, 2013). Furthermore, to make the treated and control groups comparable, we need to find a control group that has same covariate distribution as the treated group. Common approaches for handling these challenges include using prognostically important covariates to form groups of similar units containing both treated and control units (e.g. statistical matching) and/or modeling responses through interpolation. Hence, treatment effects are only reliably estimated for a subpopulation under which a common support assumption holds—one in which treatment and control covariate spaces overlap.

Incomplete overlap and imbalance are issues for causal inference largely because they cause model dependency. In general, when the treatment and control groups are unbalanced, the simple comparison of group averages is not a good estimate of the average treatment effect due to the presence of confounders. Even though researchers try to fit the best model to estimate treatment effects, their results may become biased due to incorrect specification of the relationship between confounders, treatment, and response in the model. To

ensure robustness of estimates, some analysis must be performed to adjust for pre-treatment differences between the groups.

Researchers often use statistical matching based on pre-treatment confounders to improve the precision of causal estimates. Matching helps to remove the covariate imbalance between treated and control groups in the following way: given a dissimilarity measure between pre-treatment covariates, each unit is grouped with a set of units of the opposite treatment status with small dissimilarity. Common choices for dissimilarity include (possibly weighted) Euclidean and Mahalanobis distances. Often researchers use the propensity score—which is the probability of being a treated unit given the observed covariates—or genetic matching to automatically weight covariates when the number of covariates is large (Rosenbaum and Rubin, 1983; Diamond and Sekhon, 2012). In many cases, applying matching methods on large observational study can answers questions that are difficult using randomized trial designs. For example, large administrative datasets are frequently used by pharmaceutical researchers to discern rare but dangerous side-effects of medications that were approved for sale based on trials of a few thousand persons but may eventually be used by millions of people per year (Stuart et al., 2013).

Imbalance can increase when there are a few observed units that are far away from the rest of the units in the sample. In that case, treatment effect estimation is only reliable if we can trim those units. However, trimming those units to obtain an interpretable region of common support is a big challenge (King and Zeng, 2006), and many methods have been proposed to find a common support region. We develop a method called largest connected components (LCC) to eliminate those distanced units and form a subpopulation under which all units have acceptable match. LCC seeks a balance between computational efficiency, estimation efficiency, flexibility of permitted regions, performance in high-dimensional covariate spaces, and interpretability of the common support region.

We analyze a study on right heart catheterization (RHC)—a procedure involving inserting a catheter into the heart to monitor critically ill patients—to compare the performance of LCC against other common support methods. We supplement our findings thorough simulation study based on real observational data. We show that the covariate imbalances

are reduced between treated and control group under LCC. In addition, LCC improves existing methods by efficiently constructing regions that preserve clustering in the data while ensuring interpretability of the region through the distance metric. Our results also show that treatment effect estimation is more robust to choose of model]under common support.

1.2 Overview of Statistical Literature on Causal Inference

One common misconception of many researchers when confronted with causation and correlation is that 'Correlation implies causation' (Beebee et al., 2009). In literature, causation is defined as: "Cause is an event followed by another (effect)," and "Without the first event (cause), the second (effect) would never happen," which forms the foundation for the sufficient and necessary conditions for causality (von Giźycki and Coit, 1891). These intuitive causal definitions were translated into statistical language by Splawa-Neyman et al. (1990) using the 'potential outcomes' paradigm to define 'causal effects' (Neyman's framework).

1.2.1 Potential Outcomes Framework

In Neyman's framework, the causal effect is defined as the comparison of potential outcomes Y_1 under treatment and potential outcomes Y_0 under control, i.e., Y_1 versus Y_0 for a given unit i. This comparison can be measured either in the form of a difference (additive scale) or a ratio (multiplicative scale) or some other generalized contrasts. As defined in Holland (1986), the fundamental problem of causal inference is that once treatment assignment has occurred, each subject is assigned either to treatment or control, so only one of the two potential outcomes is observed. The potential outcomes that we do not observe are known as '*counterfactuals*'. Given a counterfactual condition, e.g., 'What would be response if a treated patient received no treatment', we estimate the potential treatment effect.

Let Y_{i1} and Y_{i0} denote the potential outcomes of unit i under treatment and control

respectively—that is, Y_{it} is the hypothetical potential outcome if that unit receives treatment $T_i \in \{0, 1\}$. We observe T_i and Y_i , where

$$Y_i = Y_i(T_i) \equiv \begin{cases} Y_{i1} & \text{if } T_i = 1, \\ Y_{i0} & \text{if } T_i = 0. \end{cases}$$

In a simple observational study with four subjects, we might observe data like that summarized in Table 1.1. Since we cannot observe the potential outcomes under treated and control units, we aim to compute the sample average of the response of observed treated and control units.

Subject	Potential Y_1	Potential Y_0	T	Observed Y_1	Observed Y_0	Causal Effect
1	Y_{11}	Y_{10}	1	Y_{11}	?	?
2	Y_{21}	Y_{20}	1	Y_{21}	?	?
3	Y_{31}	Y_{30}	0	?	Y_{30}	?
4	Y_{41}	Y_{40}	0	?	Y_{40}	?
Mean				$\frac{Y_{11}+Y_{21}}{2}$	$\frac{Y_{30}+Y_{40}}{2}$	$\frac{Y_{11}+Y_{21}}{2} - \frac{Y_{30}+Y_{40}}{2}$

Table 1.1: Fundamental problem of causal inference in terms of potential outcomes. The row "Mean" gives standard sample estimates.

1.2.2 Review of Causal Inference

In evaluating causation or causal effects, randomized experiments are the "gold standard" (Hall, 2007). Their strength stems from the principle of randomization or lack of bias towards any covariate levels, as noted by Fisher (Hall, 2007; Fisher, 1925; Basu, 1980). However, randomized experiments are not always feasible in medical science and may be expensive compared to observational studies. Providing a causal interpretation of an estimate obtained from observational data requires additional assumptions or conditions under which we could imagine some form of chance mechanism was involved in the process of data collection.

Suppose we have a random sample of size n from a large population in which there are n_1 treated units and n_0 control units. Thus $n = n_1 + n_0$. For each unit i in the sample, let T_i

indicate whether or not the treatment of interest was received, with $T_i = 1$ if unit i belongs to the treatment regime, and $T_i = 0$ if unit i belongs to control regime. We denote the K observed covariates for individual i as the column vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})' \in \mathbb{R}^K$.

1.3 Assumptions

We need the following two assumptions to estimate the average treatment effect.

Assumption 1.3.1 (Unconfoundedness). We assume that $T_i \perp\!\!\!\perp (Y_{i1}, Y_{i0}) \,|\boldsymbol{X}$.

Assumption 1.3.2 (Common Support). For all $x \in X$, $0 < P(T_i = 1|\boldsymbol{X}) < 1$.

The first assumption is known as '*ignorability*' or unconfoundedness assumption (Rosenbaum and Rubin, 1983) and asserts that, conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes. The second assumption ensures common support in the covariate distributions and it is required to have a comparable treatment and control group. Together, unconfoundedness and overlap constitute a property known as"*strong ignorability*." For example, in a study of a risky medical procedure, sick people are most likely to receive treatment and are less likely to survive. We would predict a person that gets treatment that have a lower survival chance under both treatment and control. Hence, the covariates $\boldsymbol{X}$ must include initial health (along with possibly other covariates) in order to ensure the assumption of ignorability or unconfoundedness holds.

Making a causal statement or interpretation requires that the observational study imitates a randomized experiment where all the covariates are equally distributed between the treated and control groups. However, such balance between the treated and untreated groups is not usually seen in observational studies. Under the strong ignorability assumptions, observational studies can be viewed as quasi-randomized experiments, i.e., treatment assignment can be assumed random conditional on measured covariates. It is then possible to make causal inferences with the hope that the untestable assumptions are approximately true. Unfortunately, without subject-area knowledge or use of additional information to justify the assumptions, such inferences cannot be validated.

1.4 Neyman-Rubin Potential Outcomes Model

We assume the Neyman-Rubin potential outcomes model for response (Splawa-Neyman et al., 1990; Holland, 1986; Rubin, 1973), where potential outcomes are non-random. In particular, we assume that randomness in the response entirely depends on the treatment assignment. The observed outcome, Y_i, can be written as:

$$Y_i = Y_{i1}T_i + Y_{i0}(1 - T_i). \tag{1.1}$$

Inherent in 1.1 is the SUTVA, i.e., that treatment status of a unit does not affect the response of any other units (Heckman and Robb, 1985). Since, for each subject depending on the treatment status, only one potential outcome is observed—the treatment effect for unit i, denoted $\tau_i \equiv Y_{i1} - Y_{i0}$, is unobservable.

1.5 Quantity of Interest

The causal estimands of interest depend largely on the design of the study and the research question asked. The average treatment effect

$$\text{ATE} \equiv E[Y_{i1} - Y_{i0}],$$

measures the average treatment effect across all units for the whole population. The average treatment effect on the treated

$$\text{ATT} \equiv E[Y_{i1} - Y_{i0}|T = 1],$$

measures the average effect over the distribution of the treated units. ATE is common quantity of interest for randomized experiments, while ATT is more frequently used for observational studies.

Unbiased estimation of the ATE or ATT may not be possible if there are subsets of

the population that violate the common support assumption. Thus, it may be necessary to identify a subpopulation that satisfies common support and to isolate analysis to that subpopulation. This necessitates a change the quantity of interest to a conditional ATE (CATE) or conditional ATT (CATT)—the ATE or ATT conditional on observations being in the subpopulation. Under Assumption 1.3.1, the average treatment effect (ATE) for the subpopulation with $X = x$ equals

$$\begin{aligned} \text{CATE} &\equiv E[Y_1 - Y_0|X = x] \\ &= E[Y|T = 1, X = x] - E[Y|T = 0, X = x] \end{aligned}$$

almost surely. Then the conditional average treatment effect can be estimate for a given value of X. The average treatment effect on the treated can be written as:

$$\begin{aligned} \text{CATT} &\equiv E[Y_1 - Y_0|T = 1, X = x] \\ &= E[E[Y|T = 1, X = x]|T = 1] - E[E[Y|T = 0, X = x]|T = 1] \\ &= E[Y|T = 1, X = x] - E[E[Y|T = 0, X = x]|T = 1] \end{aligned}$$

Often, inference is restricted to the sample of the units at hand. In this case, the quantity of interest becomes on sample average treatment effect:

$$\text{SATE} \equiv \frac{1}{n}\sum_{i=1}^{n}[Y_{i1} - Y_{i0}]$$

or sample average treatment effect on the treated:

$$\text{SATT} \equiv \frac{1}{n}\sum_{i=1,T_i=1}^{n}[Y_{i1} - Y_{i0}].$$

Let us define the observed response for ith treated unit as y_{i1} and the observed response for the ith control units as y_{i0}. A baseline estimator of SATE is the difference in the sample

means of the observed outcome variable between the treated and control groups:

$$\hat{\tau} = n_1^{-1} \sum_{i=1}^{n_1} y_{i1} - n_0^{-1} \sum_{i=1}^{n_0} y_{i0}.$$

The difference between sample estimate and population estimate is known as estimation error.

We focus on the most basic goal of statistical inference—the deviation of an estimate from the truth. Examples include as unbiasedness, consistency, efficiency, asymptotic distribution, admissibility and mean-square error. These statistical criteria can each be computed from our results (by taking expectations, limits, variances, etc.), but all are secondary to understanding and ultimately trying to reduce estimation error in a real life scenario.

The estimation error (Δ) can be decomposed into two parts (Imai et al., 2008): error due to sample selection ($\delta_s \equiv$ATE-SATE) and error due to treatment imbalance ($\delta_t \equiv$ SATE-$\hat{\tau}$). Moreover, sample selection and treatment imbalance each can be divided into two parts—due to selection on observed (X) and unobserved (U) covariates. Thus, the estimation error can be written as:

$$\begin{aligned} \Delta &\equiv \delta_s + \delta_t \\ &= \delta_{sx} + \delta_{su} + \delta_{tx} + \delta_{tu}, \end{aligned}$$

where $\delta_s = \delta_{sx} + \delta_{su}$ and $\delta_t = \delta_{tx} + \delta_{tu}$. If the sample is representative part of the population then $E(\delta_s) = 0$. On the other hand, blocking in experimental design and matching in observational studies ameliorate treatment imbalance (δ_t). Other biases that could arise in empirical analysis—for example, post-treatment effects, measurement error, simultaneity, lack of compliance with the treatment assignment and missing data—are out of scope of this study.

1.6 Stratification and Randomized Experiments

Randomized experiments are the gold standard for statistical studies because they allow the greatest reliability and validity of estimation of treatment effects. They involve randomly selecting a sample and randomly allocating the subjects in the sample across the treatment groups. For example, if an experiment compares a new drug against a standard drug, then the patients should be randomly allocated to either the new drug or to the standard drug control using randomization. Random selection avoids selection bias by identifying a given population and guaranteeing that the probability of selection from this population is related to the potential outcomes only by random chance. Random allocation of treatment can even guarantee the absence of omitted variable bias (in expectation) without adjustment for confounding variables.

In the subsequent discussion of observational studies, we consider methods for (non-randomized) observational data that can be viewed as analyzing the data as if they arose from hypothetical stratified randomized experiment (Imai et al., 2008). Suppose there are s strata in which the jth stratum contains n_j units. In each stratum, n_{jt} units receives the treatment t $(t = 1, 2, \ldots, r)$. Furthermore, n_{jt} is fixed across randomization, resulting in n_j distinct assignments to treatment for each stratum j. Under the potential outcomes framework with multiple responses, the ith unit in the jth stratum when exposed to treatment regime t has potential outcome $Y_{ij}(t)$. Each unit is associated with observed covariates $\boldsymbol{x}_{ij}$. The treatment effect of treatment t with respect to t' for the ith unit in jth stratum, $\tau_{ij,tt'} = Y_{ijt} - Y_{ijt'}$, is unobservable as each unit is either receives exactly one treatment condition.

The observed response under Neyman-Rubin causal model (NRCM) can be written as:

$$Y_{ij} = \sum_{t=1}^{r} y_{ijt} T_{ijt}$$

where T_{ijt} takes the value 1 if ith unit in jth stratum receives tth treatment and 0 otherwise. We assume that all the units satisfy the assumptions 1.3.1 and 1.3.2. There are $n = \sum_{j=1}^{s} n_j$ units in the study, of whom $n_t = \sum_{j=1}^{s} n_{jt}$ receive the treatment t. Let $\mathcal{F} = \{Y_{ijt}, Y_{ijt'}, \boldsymbol{x}_{ij}, i =$

$i, \ldots, n_j, j = 1, \ldots, s\}$. For a randomized experiment, we can write that $\mathbb{P}(T_{ijt} = 1|\mathcal{F}) = n_{jt}/n_j$.

1.7 Post Stratification and Observational Studies

Using post stratification to obtain a causal interpretation of a treatment effect estimate from observational data requires additional considerations and assumptions. One way to obtain this is through the assumption that strata satisfying treatment symmetry can be formed. Miratrix et al. (2012) define the *assignment symmetric* in the following way:

Definition 1.7.1. A randomization is *assignment symmetric* if the following two properties hold:

- All $\binom{n_j}{n_{jt}}$ ways to treat n_{jt} units in stratum j are equiprobable, given n_{jt}.
- For all strata j, j', with $j \neq j'$, the treatment assignment pattern in stratum j is independent of the treatment assignment pattern in stratum j', given n_{jt} and $n_{j't}$.

We assume data are selected in a manner that does not generate selection bias under assignment symmetry. We also assume that researchers analyzing observational data have sufficient information in their measured pretreatment variables X so that it is possible to obtain the strata satisfying treatment symmetry. However, in the observational study case, n_{jt} and n_j may be random. Assumption 1.3.1 asserts that treatment and the unobserved potential outcomes are independent after conditioning on observed covariates and the observed potential outcomes, so we can ignore all unobserved variables. Under the definition 1.7.1, we assume that each unit in a stratum is equally probable to receive the treatment. Therefore, post-stratification recovers the randomization inference for an observational study.

Theorem 1. An unbiased estimate of the strata-level treatment effect of treatment t with respect to t' is:

$$\hat{\tau}_{j,tt'} \equiv \sum_{i:i\in j} \left[\frac{y_{ijt}T_{ijt}}{n_{jt}} - \frac{y_{ijt'}T_{ijt'}}{n_{jt'}}\right],$$

and its variance is

$$\mathrm{Var}(\hat{\tau}_{j,tt'}) = \Big[\mathbb{E}\Big(\frac{n_j}{n_{jt}}\Big)\frac{\sigma^2_{j,t}}{n_j-1} - \frac{\sigma^2_{j,t}}{n_j-1}\Big] + \Big[\mathbb{E}\Big(\frac{n_j}{n_{jt'}}\Big)\frac{\sigma^2_{j,t'}}{n_j-1} - \frac{\sigma^2_{j,t'}}{n_j-1}\Big] + 2\,\frac{\gamma_{j,tt'}}{n_j-1}.$$

The proof is given in Appendix A.

Theorem 2. Suppose μ_i and σ_i are the finite expected value and variance of τ_i. Let us define $s_n^2 = \sum \sigma_i^2$. Suppose that for $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} E[(\tau_i - \mu_i)^2 . \mathbb{1}_{\{|\tau_i - \mu_i| > \epsilon s_n\}}] = 0.$$

Under Lindeberg CLT condition, $\frac{1}{s_n}\sum_{i=1}^{n}(\tau_i - \mu_i)$ converges towards the standard normal distribution $\mathcal{N}(0,1)$.

This theorem requires a certain type of degeneracy. It requires a finite number of strata, each with nonzero variance and a sufficiently large number of units.

Lemma 1.7.1. Suppose that $\tau_1, \tau_2, \ldots$ is a sequence of random variables and that the distribution of τ_n converges to the distribution of the constant c as $n \to \infty$. Then $\tau_n \to c$ in probability as $n \to \infty$.

Proof. First note that $P(\tau_n \leq \tau) \to 0$ as $n \to \infty$ if $\tau < c$ and $P(\tau_n \leq \tau) \to 1$ as $n \to \infty$ if $\tau > c$. It follows that $P(|\tau_n - c| \leq \epsilon) \to 1$ as $n \to \infty$ for every $\epsilon > 0$. □

1.8 Observed and Unobserved Bias

Often researchers face situations where comparative studies between two or more groups are necessary to make causal inferences for policy implications. The estimation of the treatment effect can have bias due to both observed and unobserved covariates. Matching methods are popular to estimate the unbiased estimate of the treatment effect both in randomized and non-randomized experiments. In randomized experiments, researchers use matching-type methods to block similar subjects and assign treatments (Greevy, 2004; Higgins et al.,

2016). In non-randomized experiment, researchers use pre-treatment covariates to match the treated subjects with control subjects and attempt to replicate a randomized experiment as if the treatments were randomly assigned. When the covariate distributions of the treated and control subjects are different—i.e. in presence of treatment imbalance—analysis without consideration of the confounders may create a substantial bias. An appropriate matching method should reduce bias due to covariates (or treatment imbalance) by reducing the observed and unobserved covariate imbalances between treated and control groups.

Matching methods have five key steps (Stuart, 2010), with the first three representing the design and the last two representing the analysis:

1. Define "closeness": the distance measure based between units on covariates.
2. Implement a matching method given that measure of distance.
3. Iterate Steps (1) and (2) until a well-matched sample is obtained.
4. Estimate the treatment effect and inference based on the matched sample done in Step (3).
5. Conduct a sensitivity analysis on unobserved confounding variables.

Common measures of distance that can be use in Step 1 of matching are:

- Standardized Euclidean distance:

$$D_{ii'} = \left\| \frac{\boldsymbol{x}_i - \mu_{\boldsymbol{x}}}{s_{\boldsymbol{x}}} - \frac{\boldsymbol{x}_{i'} - \mu_{\boldsymbol{x}}}{s_{\boldsymbol{x}}} \right\|_2,$$

- Mahalanobis distance:

$$D_{ii'} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_{i'})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_{i'})},$$

  where $\boldsymbol{S}$ is the $p \times p$-dimensional variance–covariance matrix of $\boldsymbol{x}$.

- Propensity scores:

$$D_{ii'} = |\pi_i - \pi_{i'}|.$$

The propensity score is the probability of receiving a particular treatment (T) given a vector of observed covariates (Rubin, 1973, 2001). The propensity score is the most popular distance metric in matching (Rosenbaum and Rubin, 1983; Austin, 2007, 2008).

$$\pi_i = P(T_i|X_i) \tag{1.2}$$

Often π_i is estimated with a logistic model. There are two key features of propensity scores: (1) At each value of the propensity score, the distribution of observed covariates (that went into the propensity score) is the same for the treated and control groups. (2) the treatment assignment is independent of potential outcomes given the propensity score (no unmeasured confounders). A detail description of matching methods are discussed in Chapter 3.

1.9 Organization of the Dissertation

So far we have portrayed the assumptions, Neyman-Rubin potential outcomes model and the key concepts of the causal inference framework in a very general way. This framework allows us to to consider dependency of response on both treatment and confounders. Chapter 2 discusses in detail the assumption of common support. We introduce a method to find a common support and compare matching methods to estimate the treatment effect. In Chapter 3, a new matching technique is proposed and performance is compared with conventional matching methods and examines different aspects of the matching method—the runtime and the complexity of the algorithm. Finally, we conclude in Chapter 4.

Chapter 2

Finding Common Support Through Largest Connected Components

2.1 Introduction

In this chapter, we discuss different departures from comparability between treated and control groups. First, incomplete overlap, which occurs if there are regions in the space of relevant pre-treatment covariates where there are treated units but no controls, or controls but no treated units. Second, covariate imbalance, which occurs if the distributions of relevant pre-treatment variables differ for the treatment and control groups. Often researchers find a common support for an interpretable study population and then try to minimize the covariate imbalance by matching (Fogarty et al., 2016).

In observational studies, inference on causal quantities of interest is cleanest if the distribution of prognostically important pretreatment covariates is the same between treatment and control groups. When these distributions differ, estimates of treatment effects may be biased without some adjustment for these covariates. The success of methods to overcome this bias—e.g. statistical matching or modeling—depends not only on successfully selecting and observing the confounding covariates, but also depends on an assumption of covariate overlap or *common support*.

To ensure the common support for an interpretable study population we require overlap of the covariate space. By covariate overlap we mean that, for every treated (control) unit there is at least one matched control (treated) unit and all units forms an interpretable cluster in terms of covariate space. Sometimes a lack of overlap arises from a covariate that itself was used to assign units to treatment conditions. Sometimes it is prudent to abandon causal inferences utterly if common support cannot be found for the study.

When common support is not satisfied, methods for adjusting for covariate imbalances rely on extrapolation for estimating *counterfactuals*—the hypothetical outcomes for treated units if they had received control and vice versa. This can lead to substantial bias in treatment effects if counterfactual models do not account for important covariates or covariate interactions, or if treatment heterogeneity is present in the study (Lin, 2013; Rosenbaum, 2005). At minimum, treatment effect estimates will be highly dependent on the model used (King and Zeng, 2006; Ho et al., 2007).

To prevent these problems, it may be preferable to isolate analysis on a subset of data in which common support holds. Creating this subset, however, presents its own problem—interpretability of estimates on this subset. For example, when satisfying common support requires the removal of treated units, matching estimators on the corresponding subset no longer provide an estimate for the average treatment effect on the treated units (ATT). The ability to interpret the new estimand is highly dependent on the method used to subset the data. All current methods for finding regions of common support undergo trade-offs between interpretability of the region, flexibility on the shape of the region, and feasibility and performance when the number of observations and/or the number of prognostically important covariates are large.

The idea of identifying the common support is not only limited to find an interpretable study population but also to reduce imbalances in covariates between-groups. Covariate balance can be achieved by dropping observations whose characteristics are dissimilar to retained units according to a pre-defined metric. When treatment and control groups do not have common support, the data are inherently limited in what they can tell us about treatment effects in the regions of non-overlap. No amount of adjustment can create direct

treatment/control comparisons, and one must either restrict inferences to the region of overlap, or rely on a model to extrapolate outside this region. Thus, lack of common support is a more serious problem than imbalance (Gelman and Hill, 2007). Since both scenario includes same statistical approach, we discuss the problems together here.

2.1.1 Review of Common Support

In the methodological literature, researchers have conducted substantial research on methods to find common support. Dehejia and Wahba (1999) identify the study population removing treated units whose propensity score are larger than the maximal propensity score among the control units, and removing control units whose propensity score are smaller than the minimal propensity score among the treated units. King and Zeng (2006) interpret the covariate overlap by means of the convex hull of the treated and control covariate distributions. The idea is that interpolation is performed if a given treated (control) individual is in the convex hull of the control (treated) covariate distributions, and extrapolation is performed otherwise. Crump et al. (2009) identify the study population by finding the optimal subsamples which minimizes the efficiency bound for the variance of the study population average treatment effect. Rosenbaum (2012) finds an optimal subset of the sample where one chooses the upper bound on the maximum number of treated units that can be removed from the optimal matching. Zubizarreta et al. (2014) derive cardinality matching to create a largest balanced subset. Fogarty et al. (2016) identify the study population through solving the maximal box problem on important covariates. Many of these existing methods face computational challenges in attaining balanced subsamples with large datasets. Other issues include poor performance for high-dimensional data and interpreting the derived region of common support.

We introduce the largest connected components (LCC) method for finding common support. This method aims to strike a balance between computational efficiency, estimation efficiency, flexibility of permitted regions, and interpretability of the common support region. Given a distance metric that measures dissimilarity of units' pretreatment covariates

and a threshold ω, we form a region of common support consisting of large groups of units where each unit has a match in its group within the ω threshold. Our method allows for many different shapes regions of common support, including non-convex regions. Additionally, it is efficient enough to work when the number of units and the number of covariates are large. We also give a suggestion for a distance metric that may help improve interpretability of the region.

2.2 Framework and Definitions

Consider an study with n units. All units are given either treatment or control: there are n_1 treated and n_0 control units. Each unit i has K observable covariates, denoted $\mathbf{x}_i = (x_1, x_2, \ldots, x_K) \in \mathcal{R}^K$. The treatment status for unit i is denoted using a treatment indicator T_i: $T_i = 1$ if unit i is given treatment and $T_i = 0$ otherwise.

We assume the Neyman-Rubin potential outcomes model for response (Splawa-Neyman et al., 1990; Holland, 1986; Rubin, 1973). Let y_{i1} and y_{i0} denote the potential outcomes under treatment and control respectively—that is, y_{it} is the hypothetical outcome of unit i had that unit received treatment $t \in \{0, 1\}$. The treatment effect for unit i, denoted $\tau_i \equiv y_{i1} - y_{i0}$, is unobservable as no unit receives both treatment and control. The observed outcome, Y_i, can be written as:

$$Y_i = y_{i1}T_i + y_{i0}(1 - T_i). \tag{2.1}$$

In particular, we assume that randomness in response entirely depends on the treatment assignment. Inherent in 2.1 is the stable unit treatment value assumption (SUTVA), i.e., that treatment status of a unit does not affect the response of any other unit.

The causal estimands of interest depends largely on the design of the study and the research question asked. The average treatment effect (ATE: $E[Y_{i1} - Y_{i0}]$) measures, on average, the treatment effect across all units for the whole population. The average treatment effect on the treated (ATT : $E[Y_{i1} - Y_{i0}|T = 1]$) is the average effect over the distribution of

the treated units. The ATE is commonly the quantity of interest for experiments, while the ATT is more frequently used for observational studies.

Unbiased estimation of the ATE requires the assumptions 1.3.1 and 1.3.1. The assumptions hold over all potential realizations $(y_{i1}, y_{i0}, T_i, \mathbf{x}_i)$ in the population. The first assumption, also known as selection on observables, asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes (Rosenbaum and Rubin, 1983). Common support, also known as covariate overlap, ensures that the ATE is well-defined. Together, unconfoundedness and overlap constitute a property known as strong ignorability of assignment. When estimating the ATT, Assumption 1.3.1 can be restricted to realizations in which $T_i = 1$ and 1.3.2 can be relaxed to $0 \leq P(T_i = 1|\mathbf{x}_i) < 1$.

Unbiased estimation of the ATE or ATT may not be possible if there are subsets of the population that violate the common support assumption. Thus, it may be necessary to identify a subpopulation that satisfies common support and to isolate analysis to that subpopulation. This necessitates a change the quantity of interest to a conditional ATE (CATE) or conditional ATT (CATT)—the ATE or ATT conditional on observations being in the subpopulation (Hill and Su, 2013).

2.3 Largest Connected Components

We now develop our method of using largest connected components for finding common support. We begin with a brief description of the graph theoretic framework used for our method.

2.3.1 Graph Theoretic Framework

We view our data as a graph $G = (V = (V_1, V_0), E)$ where V_1 and V_0 are sets of vertices and E is a set of edges. Every treated unit corresponds to a vertex in V_1 and every control unit corresponds to a vertex in V_0. For each pair of vertices $i \in V_1$ and $j \in V_0$, there is an edge $ij \in E$, and no other edges exist in E—hence, G is *bipartite* with $|V_1| = n_1$ and $|V_0| = n_0$.

Each edge $ij \in E$ has a corresponding weight $w_{ij} \geq 0$, which is small if units i and j are similar.

Definition 2.3.1. For a subgraph $G' = (V', E') \subset G$, a connected component $G^* = (V^*, E^*)$ of G' is a subgraph of G' induced on V^* such that, for any two vertices $i, j \in V^*$, there exists a path of edges in E^* that connect i and j—that is, there exists $\{v_0, \ldots, v_m\} \subset V^*$ such that $\{iv_0, v_0v_1, \ldots, v_{m-1}v_m, v_mj\} \subset E^*$—and there is no path of edges that connects a vertex $i \in V^*$ to a vertex $k \in V' \setminus V^*$.

Note that a vertex with no edges is a connected component. If G' is connected, it has exactly one connected component consisting of the whole graph.

For a connected component $G^* = (V^*, E^*)$, the *size* of G^*, denoted $size(G^*)$, depends on the initial quantity of interest: For the ATT, $size(G^*) = |V^* \cap V_1|$, the number of treated units in G^*, and for the ATE, $size(G^*) = |V^*|$. In this way, when restricting inference to units within the largest connected components, the difference between the new and initial quantities of interest decreases (heuristically) as the sizes of these connected components increase.

We find it useful to focus our attention to subgraphs in which edges are drawn between vertices i and j if and only if the corresponding units are "similar enough." This can be formalized through the introduction of *bottleneck subgraphs*.

Definition 2.3.2. The bottleneck subgraph of G given threshold $\omega > 0$, denoted $BG_\omega = (V, BE_\omega)$, is a subgraph on V where an edge $ij \in BE_\omega \subset E$ if an only if $w_{ij} \leq \omega$.

$$BG_\omega \equiv \{ij \in E : w_{ij} \leq \omega\}. \tag{2.2}$$

Observe that, since there are $n_1 n_2$ edges in the original bipartite graph G, there can be at most $n_1 n_2$ unique bottleneck subgraphs.

2.3.2 Connection to Common Support

As discussed before, a region of common support is a subset of the covariate space $\mathcal{X} \subset \mathcal{R}^K$ where there is overlap of the treatment and control covariate distributions: for all $\mathbf{x} \in \mathcal{X}$, $0 < P(T_i = 1|\mathbf{x}) < 1$. While it is impossible to deduce the exact distribution of treatment and control covariates from a sample of data, a good heuristic may be to conclude that the covariate vector $\mathbf{x}_i$ is in a region of common support if unit i has an *acceptable match*—that is, there is a unit j of opposite treatment status whose covariates $\mathbf{x}_j$ are "close enough" to $\mathbf{x}_i$ with respect to a researcher-specified measure of dissimilarity (Ramsey et al., 2010). For example, if i is a treated unit, w_{ij} is a dissimilarity measure computed between i and each control unit j, and ω is the tolerable dissimilarity for an acceptable match, then $\mathbf{x}_i$ is in a region of common support if and only if there is a control unit j with $w_{ij} \leq \omega$.

It follows immediately that regions of common support can be visualized through bottleneck subgraphs. Consider the original complete bipartite graph $G = (V, E)$ where the weight w_{ij} of each edge $ij \in E$ is the covariate dissimilarity between i and j. If unit i is connected to an edge in the bottleneck subgraph $BG_\omega = (V, BE_\omega)$—or equivalently, if i belongs to a non-trivial connected component in BG_ω—then the covariate vector $\mathbf{x}_i$ is in a region of common support.

In addition, while removal of units to form a region of common support may lead to a change in the quantity of interest, it may be necessary in order to obtain accurate treatment effect estimates and reduce model dependence. For example, in the presence of heterogeneous treatment effects, it may be difficult to obtain accurate and reliable estimates for small, isolated clusters of data (Hastings et al., 2006). Hence, while a region of common support may be obtained by considering all non-trivial connected components in $BG_\omega = (V, BE_\omega)$, we recommend forming a region of common support using only the *largest* connected components. This also leads to a natural interpretation of the largest connected components method for finding common support (LCC)—LCC selects the largest subset of units under which common support can be reasonably inferred and interpolation to obtain counterfactual estimates is possible.

A further discussion about the relationship between common support regions and model dependence and the accuracy of treatment effect estimates is found in Section 4.2. Recommended tools to choose the number of connected components and the tolerable dissimilarity threshold is found in Section 2.6.

2.3.3 Choice of Edge Weights

Commonly used measures of dissimilarity—and hence, common choices for edge weights—include Euclidean and Mahalanobis distances between covariate vectors and absolute differences in propensity scores. While our method works for any choice of edge weights, it may help interpretability to choose weights that satisfy the triangle inequality: for any three units i, j, and k:

$$w_{ij} + w_{jk} \leq w_{ik}. \tag{2.3}$$

Moreover, a researcher may have an *a priori* preference for the amount of dissimilarity allowed on each covariate for a pair of units to be considered an acceptable match (e.g.: units i and j are similar enough if their heights are within 3 inches, their weights are within 20 pounds, etc). In this case, a reasonable choice of edge weight may be:

$$w_{ij}^{\infty} \equiv \max_{p^*} \frac{|x_{1p^*} - x_{0p^*}|}{c_{p^*}}, \tag{2.4}$$

where c_{p^*} is the dissimilarity allowed for covariate p^*. Hence, control unit i is an acceptable match for treated unit j if and only if $w_{ij}^{\infty} \leq 1$.

Remark 1. While w_{ij}^{∞} may be more intuitive for continuous-valued covariates, it may still be useful for categorical variables. For example, exact matching of a covariate p with d categories may be accommodated by constructing $d-1$ dummy variables $x_{p1}, x_{p2}, \ldots, x_{p(d-1)}$ and setting $c_p < 1$. This ensures that $w_{ij}^{\infty} > 1$ if i and j differ on p. Relaxations of exact matching may also be accommodated through, for example, using a weighted difference of several categorical variables (e.g. creating an index variable).

2.4 The Largest Connected Components Algorithm

The LCC algorithm requires a dissimilarity measure and a dissimilarity tolerance ω. The number of connected components comprising the common support region can be determined through diagnostic checks discussed in Section 2.6. The steps of the algorithm is given below:

1. Consider all the units as vertices and choose a dissimilarity measure.

2. Find the acceptable matches: find all pairs of units with opposite treatment statuses with dissimilarity less than ω.

3. Form a bottleneck subgraphs where edges join acceptable matches.

4. Find the set of connected components in the bottleneck subgraph.

5. Identify the largest connected components. Ensure that connected components have sufficiently many treated and control units to reliably estimate counterfactuals.

6. Select all the units that are in largest connected components to obtain a region of common support.

2.5 Graphical Presentation of LCC

2.5.1 LCC for a simple example

To demonstrate our algorithm, we construct a synthetic dataset and describe the six steps of LCC algorithm to get the largest connected components (LCC) of the graph when our estimand is the ATT. We generate data with two covariates to aid in the graphical presentation of LCC. First, in Figure 2.1a, we consider all treated and control units as vertices in a graph. Second, in Figure 2.1b we consider a value of ω to get acceptable matches. Since our estimand is ATT, acceptable matches are control units located within the highlighted circle of a treated unit. Third, in Figure 2.1c, we construct all edges between treated units and

their acceptable control matches. Fourth, we obtain the connected components as described in Section 2.3. There are two connected components in 2.1d. Figure 2.1e presents the largest connected components in the graph—pruning all units that are not in the LCC. There are 300 units (150 treated and 150 control units) in the original sample. The LCC contains 203 units among them 99 treated and 104 control units. Finally, all units in the LCC describe the study population under common support, presented in Figure 2.1f.

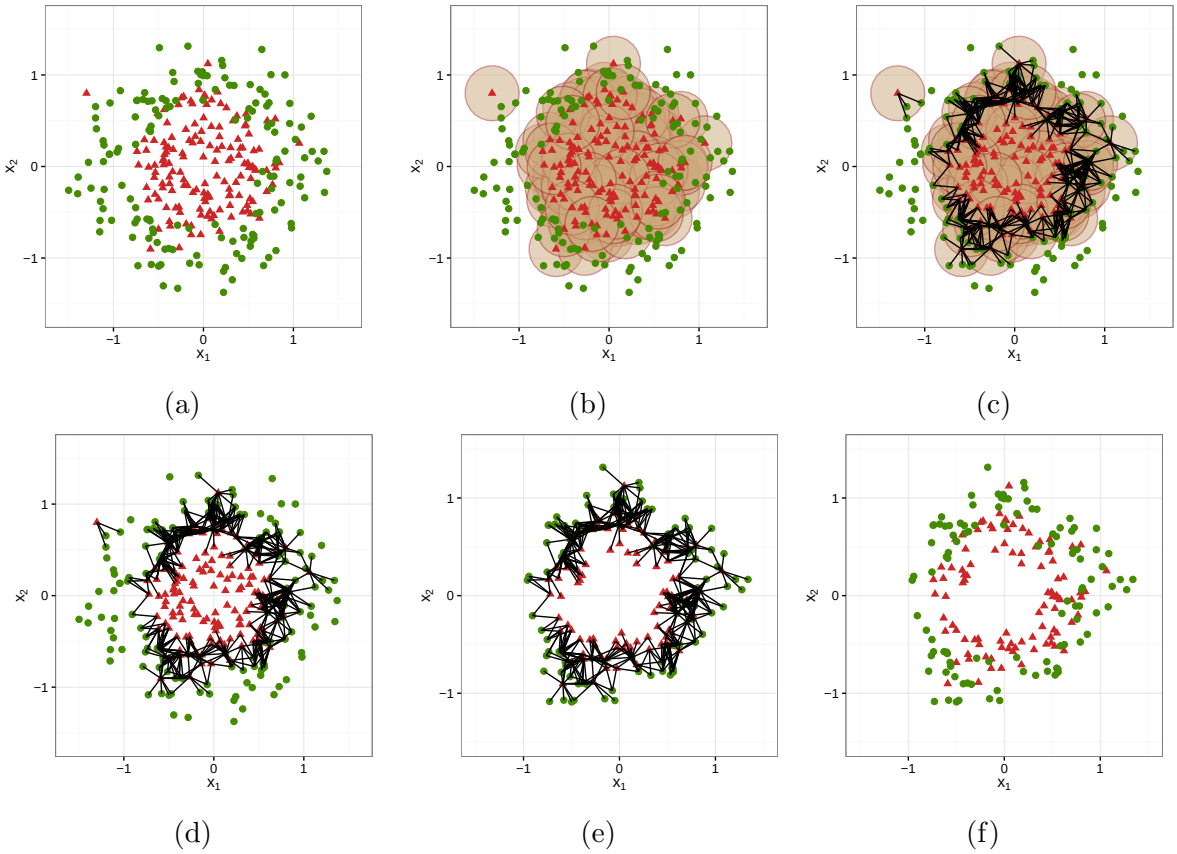


Figure 2.1: Figure 2.1a: plot all the units as vertices. Figure 2.1b: find an acceptable match for all treated units for a given dissimilarity measure. Figure 2.1c: connect all the units that have acceptable matches. Figure 2.1d: find the largest connected components. Figure 2.1e: prune all the units that are not in the largest connected components. Figure 2.1f: form the study population under common support.

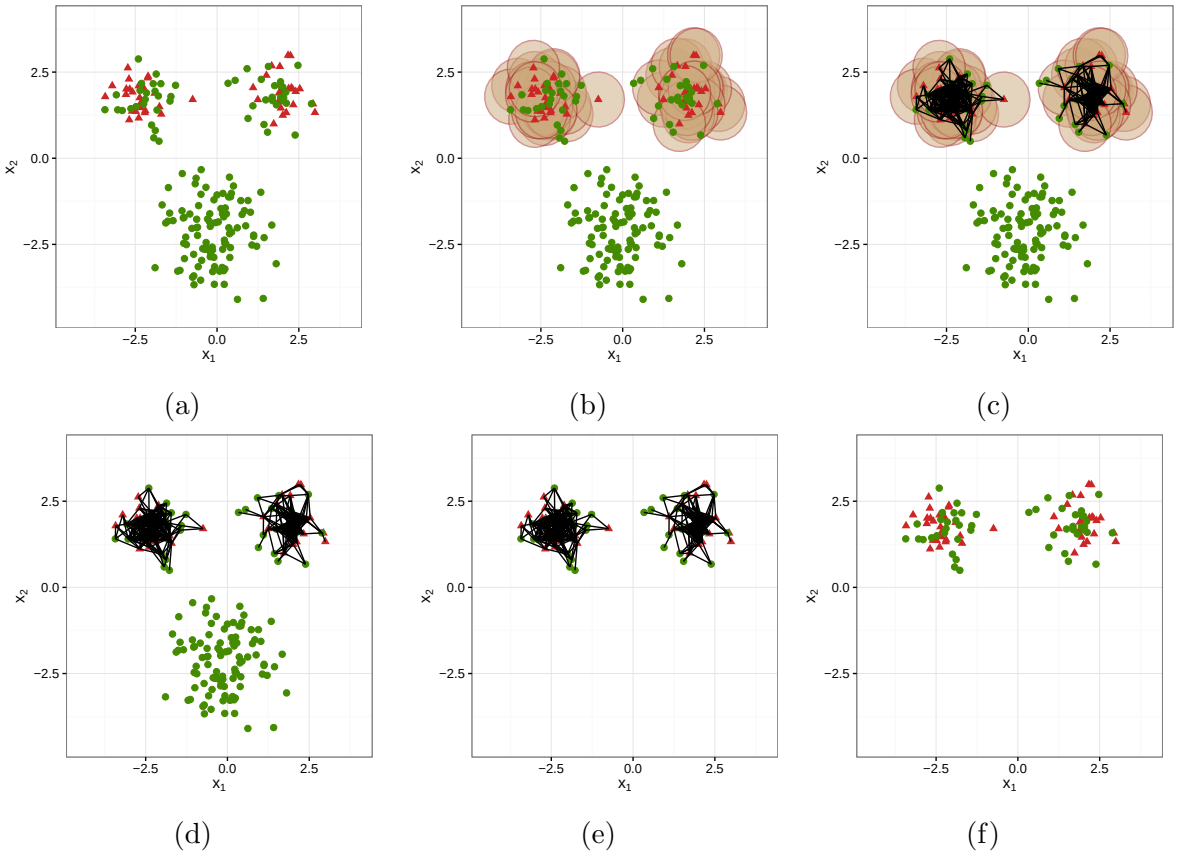


Figure 2.2: Graph 2.2a: plots all the units as vertices. Graph 2.2b: finds an acceptable match for all treated units for a given dissimilar measure. Graph 2.2c: connect all the units that have acceptable matches. Graph 2.2d: finds the largest connected components. Graph 2.2e: prune all the units that are not in the largest connected components. Graph 2.2f: forms the interpretable study population under common support.

2.5.2 LCC for Clustered Data

Often, the treated and control units are clustered and researchers are interested particularly the treatment effect for a particular cluster and/or the combined treatment effect under several clusters rather than just the largest connected components. Figure 2.2 represents a pictorial description of such an example, where we choose two connected components for the interpretable study population. Figure 2.2a–2.2f presents the LCC algorithm to find the common support. There are 200 units (50 treated units and 150 control units). The two connected components contain 25 treated units and 25 control units each.

2.6 Choice of caliper and threshold

Some covariates are important for the analysis and researchers want to reduce the imbalance on some covariates more than others. In that case, the choice of c_p for the important covariates needs to smaller (relatively) compared to the unimportant covariates. Important variables may be considered as those that are most correlated with both treatment assignment and outcome. The imbalance on important variables could cause substantial bias of treatment effect estimates.

Ideally, the threshold ω should be determined through researcher knowledge. However, it may be desirable to choose an optimal ω that minimizes covariate imbalance or the ensures a subpopulation that is sufficiently large.

First, we consider the imbalance over a covariate space $\mathcal{X}$ for the common support. Let us define Θ an L_1 pseudometric on $\mathcal{X}$ which measures the covariate imbalance. Let g and f be the empirical multivariate densities of treated and control units respectively. The objective function to minimize can be written as:

$$\Theta = \frac{1}{2} \int \cdots \int_{\mathcal{X}} |g(x_1, \ldots, x_k) - f(x_1, \ldots, x_k)| \mathrm{d}x_1 \cdots \mathrm{d}x_k, \tag{2.5}$$

where $\Theta = 1$ if and only if two densities are equal and $\Theta = 0$ if and only if two densities do not overlap each other. Equation (2.5) measures the global imbalance under the common support. One of the limitations of equation (2.5) includes no guarantee of an optimal solution. Note that, Θ is invariant to monotonic transformations of X_i and $X_{i'}$ i.e. for strictly increasing monotonic function Θ computes the minimum of two densities by standard application of the law of transformation. This is a big advantage of this measure compared with L_2 or other type metrics, as this is the only member of the L_p class that possesses this property (Anderson et al., 2012).

Second, the representation of Θ can be an expectation:

$$\Theta = E[\min\{1, l_{g_i, g_{i'}}(X)\}] = E[\min\{1, l_{g_i, g_{i'}}(X)\}],$$

where $l_{g_i,g_{i'}} = g_i(x)/g_{i'}(x)$. The choice of the parameter ω can be chosen to minimize the Θ.

Third, the representation of imbalance parameter can be expressed to minimize the imbalance is to Kolmogorov-Smirnov (KS) statistic between the empirical distributions of the treated units and controls:

$$\Theta = \sup_{x \in \mathcal{X}} |\mathcal{F}_p(x) - \mathcal{G}_p(x)|,$$

where $\mathcal{F}_p(\cdot)$ and $\mathcal{G}_p(\cdot)$ are the empirical cumulative distribution functions of the treated and control units for covariate p. This approach ensures that every unit has a matched unit based on the empirical distributions.

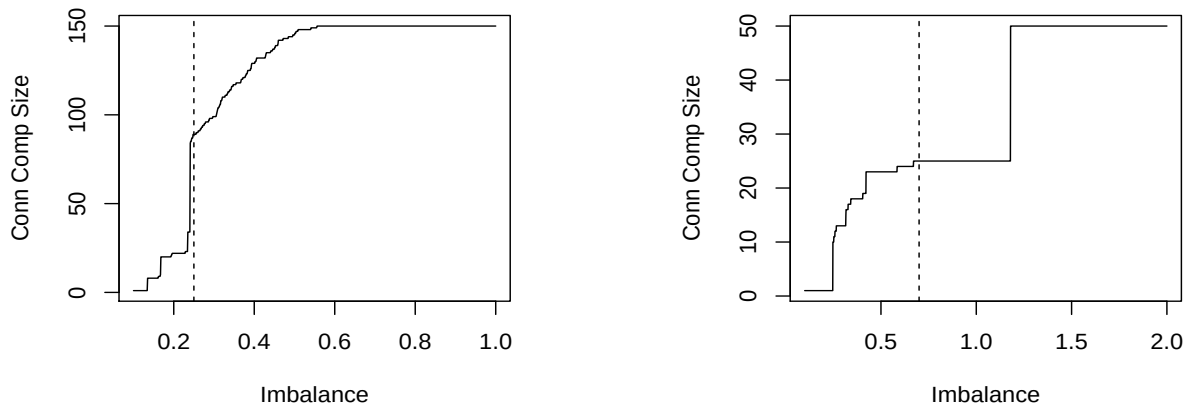
As $\omega \to \infty$, all the treated units are matched with all the control units (a complete graph), and consequently the bias of the treatment effect will be high and the variance of the estimate will be low. Again as $\omega \to 0$, all units are treated (control) units are matched with the control units that have same covariate (an exact match), and consequently the bias of the treatment effect will be low but the variance of the estimate will be high.

Proposition 1. When $n \to \infty$ and $w \to 0$, then $g_i(\boldsymbol{x}) = g_{i'}(\boldsymbol{x})$, $a.s.$ where g_i and $g_{i'}$ are the density function of treated and control units respectively.

Proof. When $w \to 0$ then, $\exists$ at least one match of treated and control in the covariate space. Again, when $n \to \infty$, $Pr(\boldsymbol{x}_i = \boldsymbol{x}_{i'}) = 1$. Hence $g_i(\boldsymbol{x}) = g_{i'}(\boldsymbol{x})$, $a.s.$. □

The choice of the parameter ω can also be selected based on the size of thee connected components that a researcher might want to consider based on equation (3.1). For example, a line graph of the connected component size against the ω parameter can be drawn. Flattening of this line suggests a greater compromise between the researcher's classification of an acceptable match and the size of a connected component. We recommend choosing a value of ω just before portions where this graph flattens.

Figure 2.3 shows that as the ω increases, the connected component size increases. Figure 2.3a and 2.3b represent the choice of the parameter ω for Figures 2.1 and 2.2, respectively. In Figure 2.3a, we choose the value of ω when the connected component size is 99. In Figure 2.3b, we choose the value of ω when the connected component size is 25.



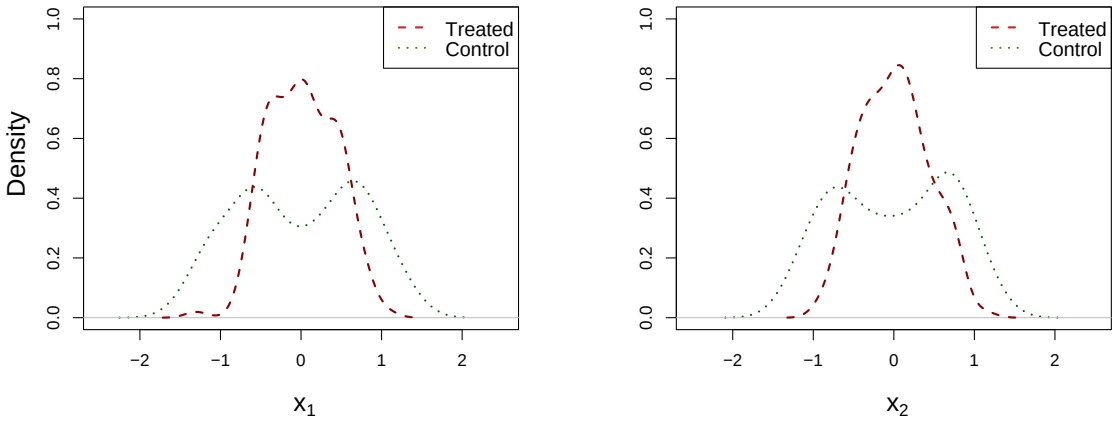
(a) (b)

Figure 2.3: Graph 2.3a: the largest connected components with $w = 0.25$. Graph 2.3b: the largest connected components with $w = 0.8$.

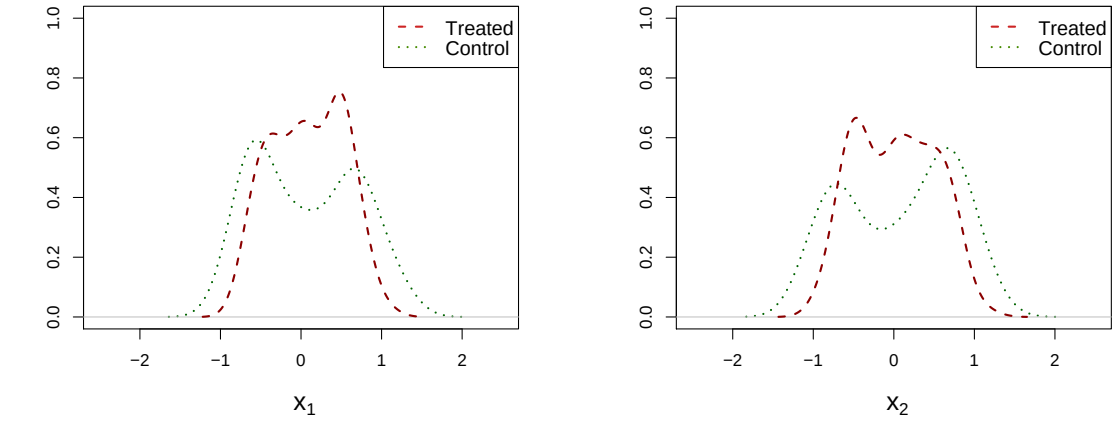
2.7 Covariate Imbalance Reduction Under LCC

It is helpful when making causal inferences to make the treated and control units comparable. To ensure the treated units are control units are comparable we plot the densities of the covariate between treated and control units. If the densities are similar between two groups then we say the two groups are comparable. Additionally, researchers can strive for equality between the univariate covariate means/proportions to compare two groups. If the standardized mean/proportion differences are significant we say the two groups are not comparable and there exits covariate imbalance in the data.

The LCC reduces the covariate imbalance and helps to make the treated and control group comparable. We plot the densities for two covariates for treated and control group for the original sample and sample under LCC. A visual inspection shows that covariate imbalance reduces drastically under common support.

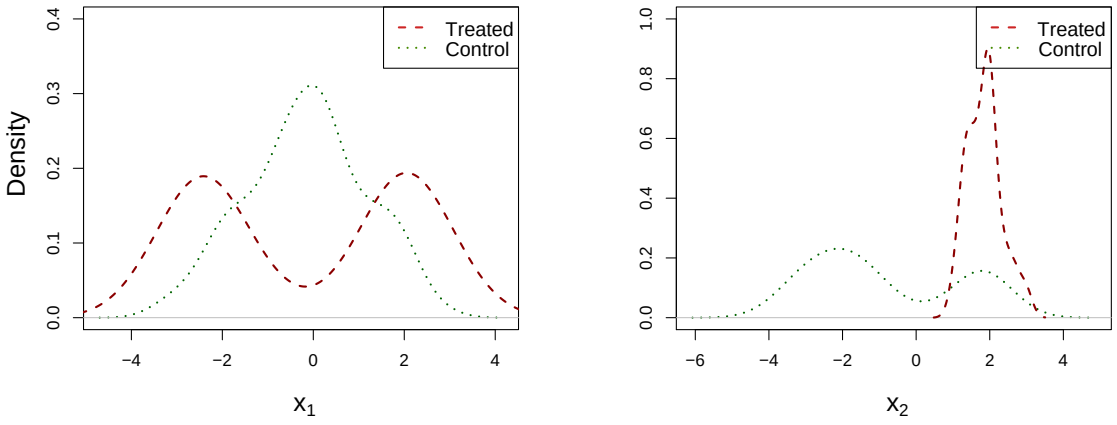


(a) (b)



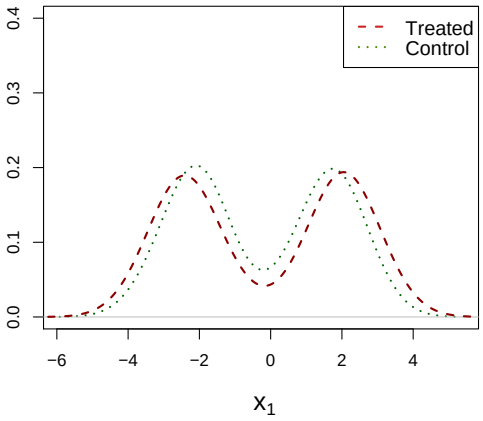
(c) (d)

Figure 2.4: Graphs 2.4a and 2.4b give the densities of the covariates under the original sample and 2.4c and 2.4d give the densities under LCC for the figure 2.1.

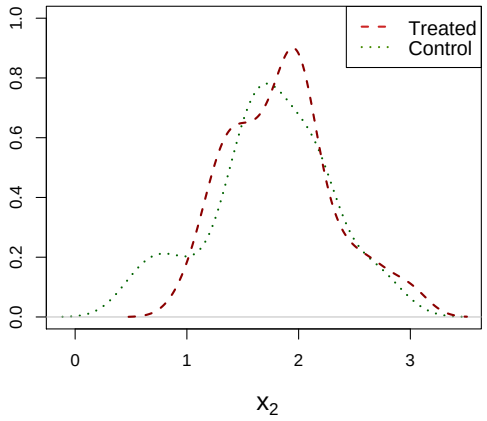


(a)

(b)



(c)



(d)

Figure 2.5: Graphs 2.5a and 2.5b give the densities of the covariates under the original sample and 2.5c and 2.5d give the densities under LCC for the figure 2.2.

2.8 Simulation

2.8.1 The Setup

We focus on a simple setting where the units' covariates are distributed on a plane.

$$x_1, x_2 \sim U(-2, 2).$$

There are two treatment conditions $t \in \{0, 1\}$, which is assigned at random based on the following propensity score:

$$P(T_i = 1|X) = 1 - \frac{x_1^2 + x_2^2}{8}.$$

The treatment assignment is generated from Bernoulli distribution with $B(n, p)$. We construct the treatment assignment in such a way so that the it will guarantee that individual probabilities of the occurrence of the treatment assignment will lie within the unit interval. So, in middle on the plane, i.e. $x_1 = 0$ and $x_2 = 0$, units have a high chance of receiving treatment and units in the edge side has high chance of receiving control units. The probability function is chosen in such a way to have a lack of overlap between treated and control units in the data.

We perform a series of Monte Carlo simulations to compare the performances of several estimation methods coupled with different methods of fining common support. The methods are: The correctly specified model; $y = x_1^2 + x_2^2 + T + \epsilon$, a misspecified model; $y = x_1 + x_2 + T + \epsilon$, nearest neighbor matching with misspecified propensity score (NN), nearest neighbor matching with correctly specified propensity score (NNCorr), full matching (FULL) with misspecified propensity score, full matching with correctly specified propensity score (FullCorr) and genetic matching.

We considered the simulation under four difference common support: without common support (Without CS), largest connected components (LCC), maximum box with correctly specified propensity score (Maxbox)and Convex hull. The simulations were run under two settings for the sample size, $n = 500$ and $n = 5000$. We considerd a Monte-Carlo simulation

to 1000 and 250 simulated data to generate the 500 units and 5000 units, respectively. We evaluated the performance of the common support methods using the following three criteria: (i) bias in estimating treatment effects; (ii) standard deviation of the estimated treatment effect and (iii) the mean squared error of estimated treatment effects.

2.8.2 Homogeneous Treatment Effect

The response model for homogeneous treatment effect was defined:

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + T + \epsilon,$$

where ϵ was standard normal. The expected outcome reached its minimum at (0,0). The true effect is 1.

Method	**Without CS**			**LCC**			**Maxbox**			**Convex Hull**		
	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE
LinCorr	1.001	0.106	0.011	1.005	0.105	0.011	1.039	0.149	0.023	0.983	0.123	0.015
LinearMis	-0.605	0.251	2.640	-0.218	0.151	1.505	-0.363	0.134	0.423	-0.168	0.246	1.423
NN	-0.587	0.072	2.522	-0.332	0.196	1.813	-	-	-	0.035	0.233	0.985
NNCorr	0.003	0.207	1.038	0.160	0.414	0.860	0.013	0.265	1.043	0.096	0.295	0.903
Full	0.044	0.370	1.049	0.120	0.220	0.818	-	-	-	-0.016	0.245	1.091
FullCorr	0.069	0.295	0.945	0.484	0.062	0.270	0.047	0.363	1.039	0.019	0.312	1.058
Genetic	0.859	0.194	0.057	0.898	0.366	0.143	0.865	0.083	0.025	0.799	0.222	0.089

Table 2.1: Estimated treatment effect, standard deviation and mean squared error under homogeneous treatment effect when n=500. On average, LCC selects 485 units, Maxbox selects 345 units and Convex Hull selects 470 units.

Table 2.1 presents the result of estimated treatment effect when sample size is small. The methods are performing well under these situations. We see that LCC performs much better than any other methods. Genetic matching perform better than any other matching methods under this simulation.

Table 2.2 presents the result when the sample size is large. In the result we see that LCC performs well among common support methods and genetic matching performs well among matching methods.

Method	**Without CS**			**LCC**			**Maxbox**			**Convex Hull**		
	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE
LinCorr	1.000	0.034	0.001	0.999	0.033	0.001	1.003	0.062	0.004	0.995	0.033	0.001
LinearMis	-0.598	0.079	2.561	-0.163	0.273	1.427	-0.137	0.256	1.358	-0.161	0.155	1.372
NN	-0.370	0.225	1.927	0.189	0.781	1.208	-	-	-	0.381	0.858	1.046
NNCorr	0.001	0.058	1.002	0.516	0.192	0.271	0.005	0.092	0.998	0.849	0.103	0.016
Full	0.030	0.263	1.01	0.161	0.786	1.259	-	-	-	0.012	0.161	1.003
FullCorr	0.087	0.428	0.999	0.628	1.155	1.307	0.187	1.057	1.767	0.381	0.858	1.046
Genetic	0.773	0.109	0.064	0.988	0.113	0.013	0.891	0.095	0.021	0.877	0.100	0.025

Table 2.2: Estimated treatment effect, standard deviation and mean squared error under homogeneous treatment effect when n=5000. On average, LCC selects 4900 units, Maxbox selects 3600 units and Convex hull selects 4960 units.

2.8.3 Heterogeneity

The response model is given by

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + T + T x_1 x_2 + \epsilon,$$

The true treatment effect is 0.977 and 1.18 for $n = 500$ and $n = 5000$ respectively.

Method	**Without CS**			**LCC**			**Maxbox**			**Convex Hull**		
	ATT = 0.892			ATT=.963			ATT=.976			0.974		
	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE
LinCorr	0.996	0.162	0.026	0.996	0.165	0.027	0.962	0.284	0.079	1.018	0.138	0.019
LinearMis	-0.585	0.276	2.587	-0.385	0.26	1.987	0.273	0.409	0.689	-0.099	0.289	1.289
NN	-0.555	0.246	2.477	-0.423	0.238	2.08	-	-	-	0.086	0.353	0.957
NNCorr	0.011	0.222	1.027	0.511	0.222	0.288	-0.004	1.21	2.423	0.909	0.182	0.041
Full	0.030	0.346	1.06	-0.003	0.285	1.086	-	-	-	0.057	0.343	1.005
FullCorr	0.054	0.342	1.011	-0.009	0.31	1.115	0.117	0.945	1.643	0.011	0.442	1.17
Genetic	0.799	0.236	0.096	0.844	0.208	0.155	0.857	0.316	0.099	0.853	0.249	0.083

Table 2.3: Estimated treatment effect, standard deviation and mean squared error under heterogeneous treatment effect when n=500. On average, LCC selects 490 units, Maxbox selects 351 units and Convex Hull selects 460 units.

Table 2.3 presents the result of estimated treatment effect when sample size is small under heterogeneous model. The methods are performing well under these situations. We see that LCC performs much better than any other methods. Genetic matching perform better than

any other matching methods under this simulation.

Method	**Without CS**			**LCC**			**Maxbox**			**Convex Hull**		
	ATT = 0.973			ATT=1.032			ATT=1.053			ATT=1.137		
	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE	$\hat{\tau}$	SD	MSE
LinCorr	1.007	0.124	0.015	1.000	0.115	0.013	0.091	0.008	1.005	0.044	0.002	
LinearMis	-0.623	0.175	2.663	-0.388	0.178	1.957	0.400	0.140	0.379	-0.436	0.078	2.067
NN	-0.543	0.221	2.431	-0.421	0.216	2.065	-	-	-	-0.388	0.09	1.933
NNCorr	-0.007	0.05	1.016	0.527	0.191	0.260	-0.004	0.124	1.022	0.689	0.064	0.101
Full	-0.005	0.144	1.030	0.034	0.146	1.090	-	-	-	0.012	0.108	0.987
FullCorr	-0.002	0.199	1.043	0.038	0.165	1.105	-0.041	0.384	1.225	0.638	0.384	0.263
Genetic	0.854	0.094	0.030	0.986	0.107	0.011	0.891	0.095	0.021	0.955	0.079	0.008

Table 2.4: Estimated treatment effect, standard deviation and mean squared error under heterogeneous treatment effect when n=5000. On average, LCC selects 4850 units, Maxbox selects 3680 units and Convex Hull 4900.

Table 2.4 presents the result when the sample size is large. In the result we see that LCC performs well among common support methods and genetic matching performs well among matching methods. Note, in these simulation the Maxbox is discarding more units compare to LCC and Convex Hull. Again, Convex Hull is very computationally expensive. In large data, it is hard to compute the Convex Hull. Compare to the performance with Convex Hull and Maxbox, LCC is performing better in this simulation setup.

2.9 SUPPORT Data

We analyses Right Heart Catheterization (RHC) data (Connors et al., 1996) to find the largest connected components. RHC is a diagnostic procedure used for critically ill patients. To measure effectiveness of RHC in an observational setting, using data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The SUPPORT study collected data on hospitalized adult patients at 5 medical centers in the U.S.. It also includes a rich set of variables relating to the decision to perform the RHC and the outcome. Connors et al. (1996) found that after adjusting for ignorable treatment assignment conditional on a range of covariates, RHC appeared to lead to lower survival than not performing RHC. This conclusion contradicted popular perception among practitioners that RHC was beneficial.

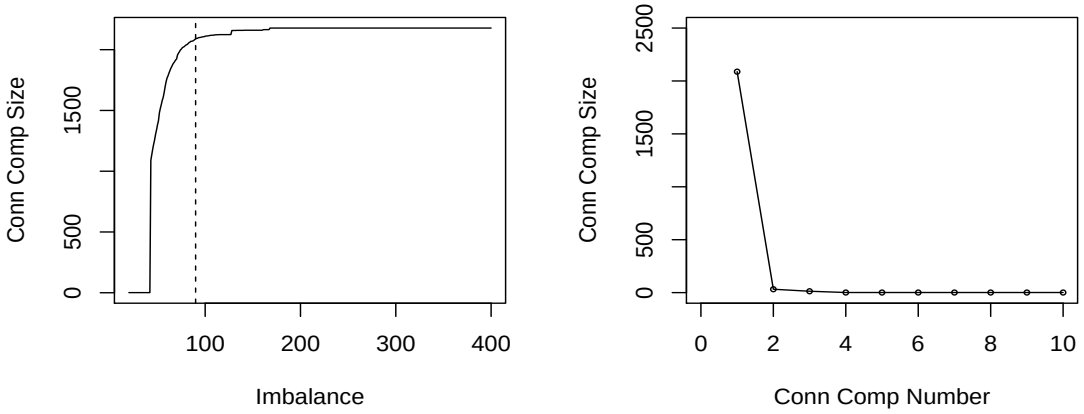
We have data on 5735 individuals of whom 2184 were treated and 3551 were controls. For each individual, we observe the treatment status, which equals to 1 if RHC was applied within 24 hours of admission, and 0 otherwise. Clinical outcome is an indicator for survival at 30 days. Support data shows that there are 68% of the RHC patients have clinical outcome compare to 63% of the No RHC patients. There are 50 covariates for covariate adjustment based on scientific knowledge. Table B.1 shows the covariate imbalance in the SUPPORT data.

Out of 50 covariates there are 32 covariates that have absolute standard differences are more that 0.1. This presents a serious covariate imbalance between treated and control groups.

	SUPPORT Data			Under LCC		
	No	Yes	Total	No	Yes	Sum
No RHC	1315.00	2236.00	3551.00	1271.00	2100.00	3371.00
RHC	698.00	1486.00	2184.00	669.00	1419.00	2088.00
Total	2013.00	3722.00	5735.00	1940.00	3519.00	5459.00

Table 2.5: The number of units under RHC and NO RHC corresponding to the clinical outcomes for orignial data and LCC data.

Table 2.5 shows we trim 276 units from the data that are far away from the rest of the



(a) (b)

Figure 2.6: Right Figure show the choice of the caliper and left Figure show that for given caliper, CC size decreases drastically after the first connected component, suggesting one CC in the common support

data. Thus, in the original data we have 5735 observation units whereas under common support we have 5459 observation units.

2.10 Discussion

Our method for finding common support successfully reduces model dependency, reduces imbalances between treatment and control groups and improves estimates of treatment effects. Our algorithm successful forms an annulus, a shape that is not possible under many common support methods. Our method yields an interpretable region. The LCC is the largest cluster of data that have comparable matches. Interpretability can be aided through the dissimilarity measure $\omega_{ii'}^{\infty}$. Creating the graph in Step 3 and finding connected components in Step 4 can be done very efficiently (Higgins et al., 2016), leading to a low runtime of the algorithm.

Chapter 3

The Performance of Largest Caliper Matching: A Monte Carlo Simulation Approach

3.1 Introduction

In this chapter, we consider six popular matching algorithms and their performance based on the simulation result. We check the bias, empirical standard deviation and the mean square error of the estimates in the simulation under different treatment prevalence and different distributions of covariates. A Monte Carlo simulation study and a real data example are employed to examine the performance of these methods. It is shown that matched samples improve estimation of the population treatment effect in a wide range of settings. Also, findings about the relative performance of the different matching methods are provided to help practitioners determine which method should be used under certain situations.

Matching methods are popular to estimate the unbiased estimate of the treatment effect both in randomized and non-randomized experiments. In randomized experiment, researchers use matching methods to form pair/block similar subjects and assign treatments. In non-randomized experiment, researchers use pretreatment covariates to match the treated

subjects with control subjects and attempt to replicate a randomized experiment as if the treatments were randomly assigned. When the covariate distributions of the treated and control subjects are different—crude analysis could make a substantial bias. An appropriate matching method should reduce bias due to covariates by reducing the observed and unobserved covariate imbalances between treated and control groups.

There are plenty of matching methods that have been developed in literature that improve the covariate balance iteratively by estimating a distance between treated units and potential controls, finding the matches, and checking balance until a satisfactory level is achieved. When there are large number of covariates—it is impossible to reduce the imbalance of all covariates altogether. The goal can be achieved by propensity score matching of treated and control groups that reduce bias due to the covariates (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999). Recently propensity score matching has been criticized as a matching method that can increase imbalance if the propensity score model is misspecified (Diamond and Sekhon, 2012; King and Nielsen, Working Paper). Another common approach that can reduce the imbalance between treated and control groups is Euclidean/Mahalanobis distance matching. One limitation of such distance metric is that if there is an extreme outlier in one covariate for a unit—the estimated variance for that covariate will be high, and Euclidean/Mahalanobis distance ignore the differences in that covariate. Gu and Rosenbaum (1993) reported that if a binary covariate that takes values 1 and 0 with probabilities p and $1-p$; whenever $p \to 0$, Mahalanobis distance tries to match a rare treated unit with this covariate equal to 1. Once the matched sample is selected through distance metric, very simple methods can be used to analyze the outcomes, and typical analysis of matched samples do not require the parametric assumptions of most regression methods (Rosenbaum and Rubin, 1985).

The quantity of interest for the outcome analysis depends on the researcher's objectives—for continuous response the most common estimand is average treatment effect (ATE) or average treatment effect on the treated (ATT) and odds ratio for the binary outcomes. Note that, if a matching method that discards both treated and control units to find a fine balance—do not result ATE or ATT. In this article, we focus on ATT to compare

the performance of the estimation of largest caliper matching compare with other matching methods.

Section 3.3 describes matching methods that have been considered in this article. Section 3.4 describes a series of Monte Carlo simulations to examine the performance of these methods in estimating treatment effects. Particularly, we report on bias, standard deviation and mean square error (MSE) of the estimates. Section 3.5 presents analysis of the right heart catheterization data. Finally, in Section 3.6, we summarize our findings.

3.2 Motivation

A common quantity of interest in observational studies is the average treatment effect on the treated (ATT)—the average difference for each treated unit between their observed response and their hypothetical response given the control condition. Estimation of the ATT is complicated by confounding—the presence of covariates that jointly affect treatment status and response.

Matching estimators are one proposed solution to deal with confounding (Abadie and Imbens, 2002). To estimate the ATT through matching, units are first assigned to groups with similar values on confounding covariates (or functions of covariates, e.g., propensity scores) so that each group with at least one treated unit also has at least one control unit (Rosenbaum and Rubin, 1983; Rubin, 2001; Austin, 2007). Each treated unit is assigned to exactly one group, but control units may belong to several groups, or no groups at all—control units that belong to a group are called *matches*. This grouping should, intuitively, ensure approximately identical multivariate distributions on confounders between treated and matched control groups. Then, an estimate of ATT is obtained by computing a weighted average of within-group differences. For example, one-to-one matching (a frequently used matching technique) finds, for each treated unit, an acceptable control match and obtains an estimate of the ATT by taking the average of the differences between the treatment and matched control responses. To check that matching "worked" for a sample of units, it is common to verify that univariate imbalance on confounders is small between treated and

matched control groups (Stuart, 2010). In fact, univariate balance is often in the objective of matching methods (Diamond and Sekhon, 2012; Zubizarreta, 2012).

Asymptotically unbiased estimation of the ATT requires two assumptions: selection on observables and common support. Jointly, these assumptions are referred to as *strong ignorability*. Selection on observables requires that all confounding variables are observable and included in the matching procedure. Common support states that, for all treated units with confounding covariates $\mathbf{x}$, there must be a non-zero probability of observing a control unit with covariates $\mathbf{x}$. Finite samples have an additional layer of complexity: treated units may have observed covariate values that lie in the theoretic common support but may not have exact matches with a control unit. Hence, to be successful in practice, matching algorithms must also must incorporate some method for determining dissimilarity between two vectors of confounders.

Most matching methods act on the problems of selection on observables and dissimilarity between confounding vectors, leaving finding a region of common support to another method. However, recent methods have been developed to help satisfy both assumptions of strong ignorability and measuring covariate dissimilarity simultaneously. These methods often involve the formation of large subsets of data (with respect to the quantity of interest) for which univariate balance is possible on each covariate (Imai and Ratkovic, 2014). Of particular note is *cardinality matching*—which constructs a subset of data with as many treated units as possible while still ensuring univariate balance between treated and control units across the entire subset (Zubizarreta et al., 2014).

Cardinality matching can be an effective method for improving ATT estimation. However, the method comes with a substantial computational cost: exact cardinality matching is an NP–hard problem. Hence, when dealing with large datasets, it may be necessary to focus on efficient heuristic or approximate solutions to cardinality matching. We present one such heuristic solution: largest caliper matching (LCM).

3.3 Methods

Several studies have been conducted to compare the matching methods. Elze et al. (2017) compared four propensity score matching methods to covariate adjustment on four cardiovascular observational studies. Austin (2014) compared 12 matching methods for 1:1 matching on the propensity score. Ming and Rosenbaum (2000) observed that substantially greater bias reduction is possible if the number of controls in match to each treated unit is not fixed. Gu and Rosenbaum (1993) compared optimal matching with nearest neighbor matching based on Mahalanobis distance. In this article we consider six different matching methods: nearest neighbor matching with replacement (NNWR), nearest neighbor matching without replacement (NNWOR), optimal matching (OPT), full matching (FL), genetic matching (GM) and largest caliper matching (LC). The choice of selecting a matched sample differs in the methods and each serves to achieve a specific objective.

3.3.1 Nearest Neighbor Matching With Replacement

NNWR matching matches all treated subjects to their nearest control subjects based on a distance metric. In this method, each treatment subject can be matched to the closest control subject, even if that control subject is matched more than once. Because this approach can provide closer matches on the distance than nearest-available matching without replacement, it can be beneficial for reducing bias in the analysis. In our analysis, we used Mahalanobis distance metric to find the nearest control for the treated subjects. An illustration of the method is shown in Figure 3.1a.

3.3.2 Nearest Neighbor Matching Without Replacement

NNWOR requires that each match contains exactly one treated subject and exactly one control subject also known as 1:1 match. Once a control subject is matched with a nearest treated subject that control subject is no longer eligible for consideration as a match for other treated subjects. That is why, NNWOR is also known as "greedy" matching. Matching

without replacement can be beneficial when there are enough good matches. Mahalanobis distance metric is used in the analysis to find nearest control for the treated subjects. The method is illustrated in Figure 3.1b.

3.3.3 Optimal Matching without Replacement

The optimal matching method seek to match subjects to minimize a global discrepancy measure, like the sum of distances within matched sets (Rosenbaum, 1989). Greevy (2004) develops the idea to improve matching methods with the goal of optimizing the overall similarity of matched subjects. Most often, optimal matching refers to matching without replacement, as optimal matching with replacement is equivalent to NNWR matching. In our analysis, exactly one treated subject is matched with exactly one control subject so that we minimize the sum of the Mahalanobis distances. Figure 3.1c illustrates the method.

3.3.4 Full Matching

Full matching considers that there exist at least one matched control (treated) subject for every treated (control) subject. Again, the treated (control) subjects are not matched with the matched control (treated) subjects. One can choose $1 : k$ or $k : 1$ matching in full matching. The flexibility of this matching method can result in using more of the data at hand and yield more effective comparisons (in terms of effective sample size) and closest-possible matches on any given distance (Hansen, 2004). In our analysis, we considered 1:3 matching in full matching with calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score. The choice of the ratio was based initial performance before we conduct the whole simulation. Figure 3.1d illustrates how the subjects would be matched using the method.

3.3.5 Genetic Matching

Diamond and Sekhon (2012) proposed genetic matching that automates the iterative process of checking and improving overall covariate balance to determine the given covariates' weight

and ensures convergence to the optimal matched sample. They proposed a distance metric for the method that minimize the overall imbalance by minimizing the largest individual discrepancy based on p-values from paired t-tests. Figure 3.1e present the sample that would be matched using the method.

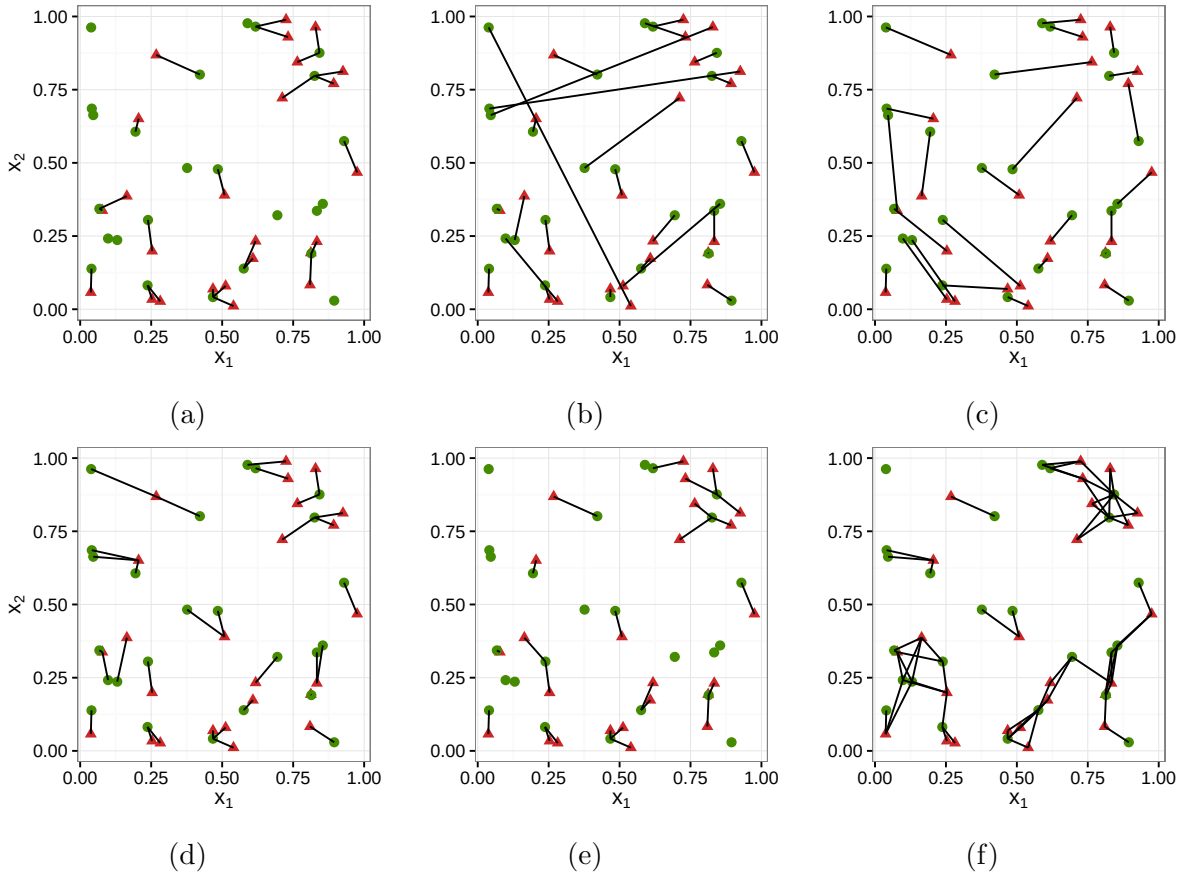


Figure 3.1: Illustration of different matching methods. The sample consists of 50 subjects, both treated and control groups have 25 subjects each. We observe two covariates x_1 and x_2, for each subject. The red triangles indicate treated subjects and green circles indicate control subjects. Edges (based on Mahalanobis distance) indicate matched groups. A good matching method should avoid long edges, as they corresponds to increase covariate imbalance.

3.3.6 Largest Caliper Matching

We introduce a method that provides a data-driven approach to select the maximum amount of imbalance that researchers want to accept for a match given a covariate, namely largest

caliper matching. For largest caliper matching we consider the following distance metric:

$$w_{ij}^{\infty} \equiv \max_{p^*} \frac{|x_{1p^*} - x_{0p^*}|}{c_{p^*}}, \tag{3.1}$$

where c_{p^*} is the dissimilarity allowed for covariate p^*. Hence, control unit i is an acceptable match for treated unit j if and only if $w_{ij}^{\infty} \leq 1$.

We note several importance of largest caliper matching: First, largest caliper matching match based on the amount of imbalance that researchers want to accept for a covariate. For example, $w_{ij}^{\infty} = 0$ means exact match based on pth covariate that researchers want to use for matching. Again, $w_{ij}^{\infty} = \infty$ means match on pth covariate is negligible. Often it is not possible to reduce the imbalance for every covariates altogether, equation (3.1) might not be optimal by random choice of the c_{p^*}. We recommend to choose the c_{p^*} based on the important covariates that are related to the treatment assignment and study outcome. For large data set one can consider c_{p^*} as the caliper for the propensity score (Lunt, 2014). Austin (2011) observed optimal calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score when estimating differences in means and differences in proportions in observational studies. Second, largest caliper matching is a heuristic matching method—for a given c_{p^*}—the average run time of the method is faster than optimal matching. Third, the choice could be based on the quantity of interest. For example, if the quantity of interest is average treatment effect for the treated (ATT) (average treatment effect for the control (ATC)), then we chose the c_{p^*} in such a way so that every treated (control) subject has at least one matched control (treated) subject. Finally, largest caliper matching ensures to discard the extreme subjects that can increase the substantial bias in the analysis (King and Zeng, 2006).

3.4 Monte Carlo Simulations

The simulations performed in the current paper are simplistic matching simulations proposed in the literature (Austin, 2014; Pirracchio et al., 2015). We conduct a number of Monte Carlo

simulations to compare the performance of six matching methods on binary outcome. In each simulated sample we compute an estimate $\hat{\tau}$ of the true parameter τ. We assessed the performance of each method using the following three criteria:

- Bias in estimating treatment effects: $\bar{\tau} - \tau$ where $\bar{\tau} = \sum_{l=1}^{N} \hat{\tau}/N$.
- Standard deviation of the estimated treatment effect: $\sqrt{\sum_{l=1}^{N}(\hat{\tau} - \bar{\tau})^2/(N-1)}$.
- Mean square error of estimated treatment effects: $\sqrt{\sum_{l=1}^{N}(\hat{\tau} - \tau)^2/N}$.

3.4.1 The Setup

We considered $\boldsymbol{X}$ be a vector of 5 covariates that had effect both on the treatment assignment and the outcome. The treatment assignment model was generated from a linear combination of the covariates:

$$\text{logit}(\pi_t) = \beta_{0,t} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (\log(1.25), \log(1.5), \log(1.75), \log(2), \log(2))$. Thus, there were one covariate that had a weak effect on each of treatment effect and outcomes, one covariate had a moderate effect on each treatment assignment and outcomes, one covariate that had a strong effect on each of treatment assignment and outcomes, and two covariates that had a very strong effect on both treatment assignment and outcomes. The intercept of the treatment assignment model ($\beta_{0,t}$) was generated so that the proportion of subjects in the simulated sample that were treated was fixed at a desired proportion. We assigned treatment status (denoted by z) of subjects from a Bernoulli distribution with parameter π_t. The dichotomous outcome was generated using the following logistic model:

$$\text{logit}(\pi_{\text{o}}) = \beta_{0,\text{o}} + \tau z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

We then generated a binary outcome for each subject from a Bernoulli distribution with parameter π_{o}. We selected the intercept, $\beta_{0,\text{o}}$, in the logistic outcome model so that the

incidence of the outcome would be approximately 0.10 if all subjects in the population were control. In a given simulated data set, we simulated a binary outcome for each subject, under the assumption that all subjects were not treated ($z = 0$). We then calculated the incidence of the outcome in the simulated data set. A bisection approach is used to determine that value of $\beta_{0,\mathrm{o}}$ that would result in an incidence of 0.10.

We selected the conditional log odds ratio τ so that average odds in treated subjects due to treatment would be approximately 0.5. The same value of τ was used to generate a cohort of $n = 5000$ in a given scenario. Because we were simulating data with a desired ATT, the value of τ would depend on the proportion of subjects that were treated. This approach allows for variation in subject-specific treatment effects. The logistic model is used to simulate data with an underlying average treatment effect in the treated because such an approach will guarantee that individual probability of the occurrence of the outcome will lie within [0,1].

In Monte Carlo simulations, we consider a complete factorial design in which the following two factors were allowed to vary: (1) the distribution of the 5 pretreatment covariates; (2) the proportion of subjects that received the treatment. We considered four different distributions for the 5 pretreatment covariates: (i) the 5 covariates had independent standard normal distributions; (ii) the 5 covariates were from a multivariate normal distribution. Each variable had mean zero and unit variance, and the pair-wise correlation between variables was 0.25; (iii) the first two variables were independent Bernoulli random variables each with parameter 0.5, whereas the other three variables were independent standard normal random variables; (iv) the 5 random variables were independent Bernoulli random variables, each with parameter 0.5. For the second factor, we considered six different levels for the proportion of subjects that were treated: 0.1, 0.15, 0.2, 0.25, 0.3 and 0.35. Hence, there are 24 different scenarios of the study: four different distributions for the pretreatment covariates times six levels of the proportion of subjects that were treated.

In each of the 24 scenarios, we simulated $N = 1000$ datasets, each consisting of $n = 5000$ subjects. There were two reasons to use simulated datasets of size 5000. First, matching methods can be computationally intensive for large data. We considered a moderate size

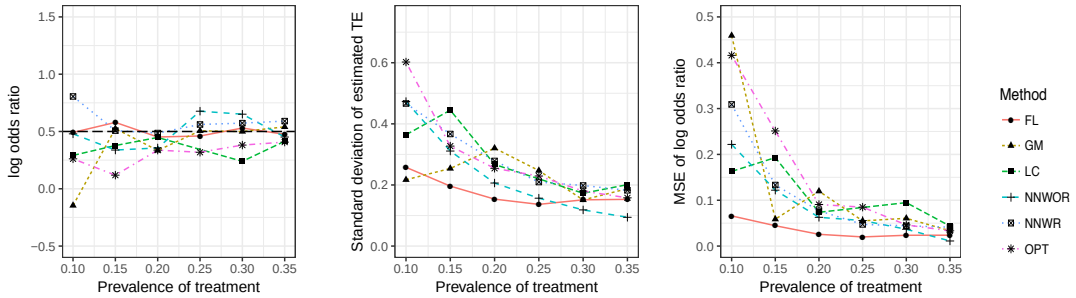
of the data that are available in real life, e.g. SUPPORT data. Second, researchers in different field usually have different size of the data—we observed in most cases these methods have been used in datasets of size around 5000. From the setup, we know the important covariates (i.e. x_4 and x_5) when matching and a good matching method should have more weight on these covariates. Though in real life it is unknown that which variables are important for treatment and outcome but in practice—researchers use the existing literature or subject-matter knowledge and expertise to identify important variables that affect the treatment assignment or outcome. In each matched sample, we estimated the log odds ratio as the treatment effect. As the matched sample removes the effect of confounding due to pretreatment covariates—it was expected the estimates were unbiased.

In SUPPORT data, we check the covariate imbalance by standardized difference. For continuous variables, the standardized difference is defined as $d = (\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_c^2)/2}$, where $\bar{x}_t$ and $\bar{x}_c$ denote the sample mean of the covariate in treated and control subjects, respectively, whereas s_t^2 and s_c^2 denote the sample variance of the covariate in treated and control subjects, respectively. For dichotomous variables, the standardized differences are defined as $d = (\hat{p}_t - \hat{p}_c)/\sqrt{(\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_t c))/2}$, where $\hat{p}_t$ and $\hat{p}_c$ denote the prevalence or mean of the dichotomous variable in treated and control subjects, respectively.

3.4.2 Results

In Figure 3.2 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when the pretreatment covariates were independently normally distributed. Figure 3.2a shows the bias of the methods under different treatment prevalence. A horizontal line has been added to each panel denoting the magnitude of the true log odds ratio 0.5. Figure 3.2b and 3.2c show the standard deviation and mean square error of the estimated log odds ratio, respectively. In general, as the prevalence of treatment increased the precision of the estimates increased for all matching methods. Optimal matching and nearest neighbor matching with/without replacement tended to have similar performance under independently normally distributed covariates. Amongst all methods, 1:3 full matching with caliper show

the less standard deviation and mean square error of the estimated log odds. Largest caliper matching is the second choice in this scenario. Note that when the treatment prevalence is small, e.g. 10%, 1:1 nearest neighbor with/without replacement or optimal matching discards at least 80% of the subjects from the data.

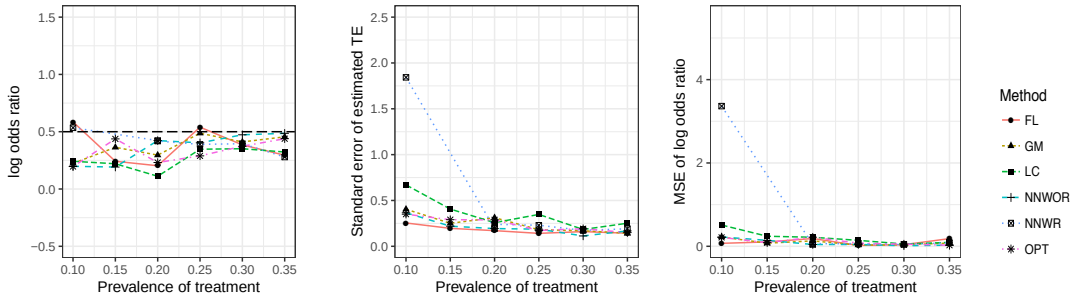


(a) (b) (c)

Figure 3.2: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under independent normally distributed covariates.

Figure 3.3 presents log odds ratio, standard deviation and mean square error of log odds ratio when the pretreatment covariates were multivariate normally distributed. The estimated treatment effect is reported in Figure 3.3a. We see that nearest neighbor matching with replacement performs better than nearest neighbor matching without replacement. Largest caliper matching performed well through different treatment prevalence. The standard deviation and mean square error of the estimated log odds ratio are reported in Figure 3.3b and 3.3c, respectively. Optimal matching and full matching showed less standard deviation and less mean square error in this case. The standard deviation was high for nearest neighbor matching with replacement when the treatment prevalence is low. Genetic matching performed better than any other methods when covariates were multivariate normally distributed.

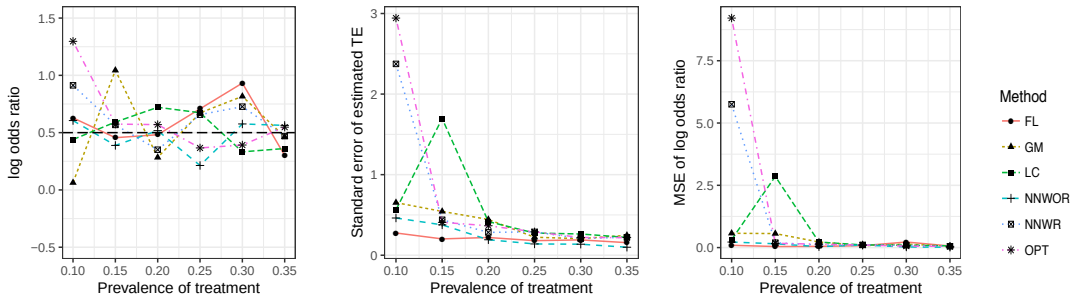
In Figure 3.4 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when the pretreatment covariates were both normally and binary distributed. Figure 3.4a shows the bias of the methods under different treatment prevalence. Largest caliper matching was performing consistent over different prevalence of treatment. Figure 3.4b and 3.4c show the standard deviation and mean square error of the log odds ratio,



(a) (b) (c)

Figure 3.3: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under multivariate normally distributed covariates.

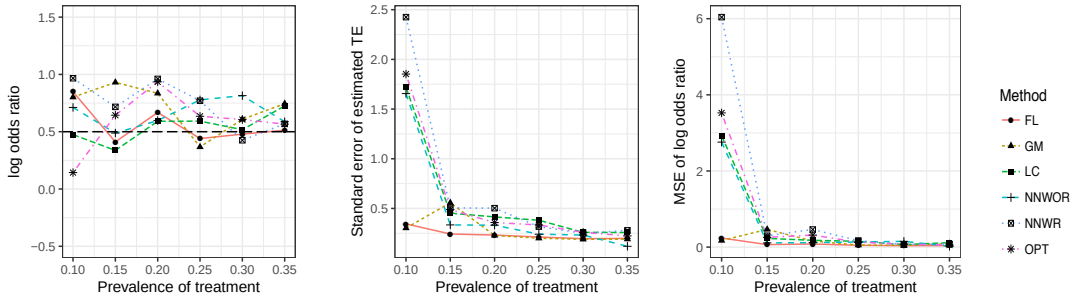
respectively. Optimal matching and nearest neighbor matching with replacement had low precision in presence of low treatment prevalence. Both 1:3 full matching with calipers and largest caliper matching performed better than other matching methods.



(a) (b) (c)

Figure 3.4: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under both normally distributed and binary distributed covariates.

In Figure 3.5 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when pretreatment covariates were independently binary distributed. Figure 3.5a shows the bias of the methods under different treatment prevalence. Both genetic matching and 1:3 full matching performed better than other methods in presence of low treatment prevalence. Figure 3.5b and 3.5c show the standard deviation and mean square error of the estimated log odds ratio, respectively.



(a) (b) (c)

Figure 3.5: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under binary distributed covariates.

3.5 Case Study

We analyzed Right Heart Catheterization (RHC) study to investigate whether RHC led to increase odds of severe clinical outcomes, previously analyzed by several authors (Connors et al., 1996; Imbens, 2001). We applied six matching methods on the effectiveness of RHC using data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The RHC study collected on hospitalized adult patients at 5 medical centers in the U.S. Based on information from a panel of experts a rich set of variables relating to the decision to perform the RHC and outcome. Connors et al. (1996) found that after adjusting for ignorable treatment assignment conditional on a range of covariates, RHC appeared to lead to increase clinical death. This conclusion contradicted popular perception that RHC patients had less risk of clinical outcome. A detailed description of the study can be found in Connors et al. (1996) and Imbens (2001).

We had data on 5735 subjects, 2184 treated patients and 3551 control patients. For each subject we observed treatment status, equal to 1 if RHC was applied within 24 hours of admission, and 0 otherwise. Clinical outcome was an indicator for death within 30 days. There were 68% of the RHC patients that had clinical outcome compared to 63% of the No RHC patients. We considered 50 covariates for covariate matching based on the covariates that are associated with the both RHC and clinical outcome.

In unmatched data, out of 50 covariates there were 32 covariates that had absolute

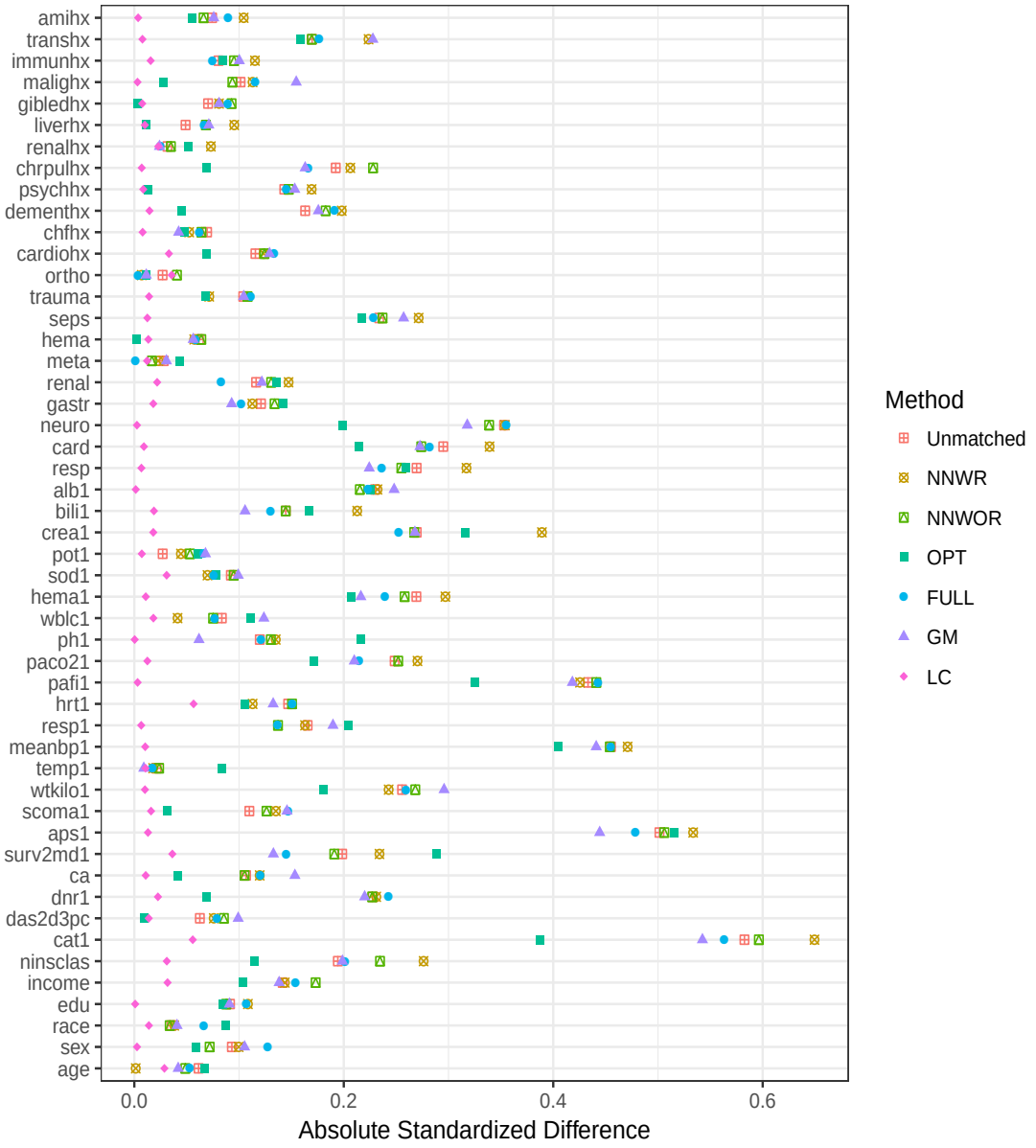


Figure 3.6: Covariate imbalance between treated/control subjects. The dotplot (a Love plot) shows the absolute standardized differences for unmatched and six matched samples.

standardized differences were more than 0.1. NNWR and NNWOR had 34 and 31 covariates that had absolute standardized differences more than 0.1. OPT performed better than nearest neighbor matching in terms of reducing covariate imbalance. LC successfully reduced all the covariate imbalances in the data and the result were consistent with other matching methods. Figure 3.6 reports the standardized difference for each of the 50 covariates in the matched and unmatched data.

We analyzed the unmatched data and the matched samples obtained from six matching methods. Table 3.1 shows the outcome analysis of the SUPPORT data. The second column presents the odds ratios of the analyses. We report that RHC was significant at 5% level of significance under all matching methods.

Method	OR	2.5%	97.5%
Unmatched	1.252	1.119	1.402
NNWR	1.267	1.074	1.492
NNWOR	1.215	1.068	1.383
OPT	1.544	1.364	1.747
FULL	1.167	1.023	1.333
GM	1.243	1.097	1.409
LC	1.276	1.121	1.452

Table 3.1: Odds ratio of RHC group compare to No RHC group with 95% confidence interval.

3.6 Conclusion

The article discusses a new matching technique and compare the relative performance of the method with current existing methods under different Monte Carlo simulations setup. In this section, we briefly discuss our simulation results.

In general, we observed several important facts that researchers need to consider in employing these matching methods. First, as the prevalence of the treated subjects increased from 10% to 35% in data, all methods tend to estimate unbiased estimate in the data and both standard deviation and mean square error of the estimates started to decrease. Second, full matching (in our case 1:3 with caliper) imposed more subjects than other methods—tended to result more precise estimates compared with the other matching methods. Note that full matching would perform better to reduce the covariate bias in the outcome analysis but could worsen covariate imbalance. Third, the choice between nearest neighbor with replacement and nearest neighbor matching without replacement reflected a bias-variance trade-off. In general, the nearest neighbor with replacement had lowest bias but higher variance compares to nearest neighbor without replacement. Some authors demonstrated this fact—matching with replacement produces matches of higher quality than matching

without replacement by increasing the set of possible matches but have greater variability (Abadie and Imbens, 2006). Fourth, when covariates have multivariate normally distributed covariates—Genetic matching tended to have a performance that was at least as good as any of the competing methods. Fifth Sixth, we used Mahalanobis distance metric for nearest neighbor with replacement, nearest neighbor without replacement, optimal matching and full matching. In simulation, we observed that for small number of covariates (in our case we considered five covariates) Mahalanobis distance metric performs much better than propensity score matching. Finally, our conclusions might be restricted to our simulation scenarios and might not apply to situations not represented by our simulated data.

The quantity of interest always depends on the researchers objectives—that need to setup before analysis. If the number of control subjects are insufficient then nearest neighbor without replacement can result in exclusion of some treated subjects from the matched sample. Rosenbaum and Rubin (1985) used the term 'bias due to incomplete matching' to describe the bias that arises when treated subjects are excluded from the matched sample. In many real application, it is could be beneficial to discard some treated subjects without good match to obtain a good covariate balance. If a matching method discards treated subjects—the quantity of interest is no longer ATT. Since, in our simulation we considered the treatment prevalence maximum of 35%, our quantity of interest for all matching methods was ATT.

Our findings show that largest caliper matching performd fair under difference setup. In presence of large number covariates we recommend to use all the covariates that are important for both treatment assignment and outcome. Unnecessary inclusion of covariates in the matching methods could reduce the performance of the methods (Stuart, 2010). Besides employing caliper on covariates—adding calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score for largest caliper matching in large data could make better performance. In this article, the analyses was conducted as a post-stratified sample—all the formed clusters were given weight to estimate ATT. In methodological literature, researchers have conducted substantial research on methods to estimate treatment effects. Besides, computationally they are very convenient—there are several `R` packages

available for matching methods, e.g. `Matching`, `MatchIt` and `optmatch`.

We like to note certain attentions for the users of largest caliper matching. First, in largest caliper matching the analysis is sensitive to the choice of the caliper that could make substantial difference in matched sample. One choice of the caliper could be, consider only the important covariates that have higher standardized difference than a tolerance level. Second, a tighter caliper leads to reduce bias and make good matches but could discard those treated subjects that do not have good matches. Third, largest caliper matching ensures that there is at least one match for all treated subjects when the quantity of interest is ATT. Fourth, largest caliper matching is fast for a given amount of imbalance that researchers want to accept for a covariate. For SUPPORT data set, our largest caliper matching took 2.7 seconds to to run on a desktop computer with 2.7 GHz Intel Core i7 processor and 16.0 GB RAM. Finally, largest caliper matching forms a good match sample that forms a cluster of homogeneous subjects. It successfully discards the control subjects that could increase the imbalance in the data. Combining these characteristics, largest caliper matching is a very flexible and convenient matching method.

Chapter 4

Conclusion

4.1 Introduction

The main goal of this dissertation was to assess and validate causal inference tools to estimate the causal effect of a treatment. Finding the treatment effect in observational studies is complicated by the need to control for confounders. Common practice for controlling include using prognostically important covariates to form clusters of similar units containing both treatment and control units. This formation of similar units tries to reduce the imbalance due to treatment assignment. Under specific assumptions described in this dissertation the causal effect of a treatment can be estimated efficiently. The dissertation proposes a series of new, computationally efficient methods to improve the analysis of observational studies.

Treatment effects are only reliably estimated for a subpopulation under which a common support assumption holds—one in which treatment and control covariate spaces overlap. Given a distance metric measuring dissimilarity between units, graph theory is used to find common support. An adjacency graph is constructed where edges are drawn between similar treated and control units. Regions of common support are determined finding the largest connected components (LCC) of this graph. LCC improves on existing methods by efficiently constructing regions that preserve clustering in the data while ensuring interpretability of the region through the distance metric.

This approach is extended to propose a new matching method called largest caliper matching (LCM). LCM is a version of cardinality matching—a type of matching used to maximize the number of units in an observational study under a covariate balance constraint between treatment groups. While traditional cardinality matching is NP–hard, LCM can be completed in polynomial time. The performance of LCM with other five popular matching methods are shown through a series of Monte Carlo simulations. The performance is measured by the bias, the empirical standard deviation, and the mean square error of the estimates in the under different treatment prevalence and different distributions of covariates. The formed matched samples improve estimation of the population treatment effect in a wide range of settings, and suggest cases in which certain matching algorithms perform better than others.

4.2 Implications

Finding treatment effects in observational studies is complicated by the need for appropriate model–one that adjusts for all the important covariates and their interactions. Adjustment can lead to worsened asymptotic precision if appropriate model is not considered (Lin, 2013). The model dependency is reduce under common support, hence helps to choose researchers flexible and interpretable model. We present the sensitivity of model dependency through a simulation in Figure 4.1. This data set was designed to illustrate the problem of model sensitivity and increased robustness of treatment effects under common support (Ho et al., 2007). The figure shows that common support discards the outliers in the data to reduce model dependency. Some units in data may be influential in terms of covariate but may not be an outlier in regression model. A good data analysis should include the units that would not change the parameter of interest under different choice of the regression model.

In Figure 4.1, each data point plotted as a "▲" for treated units and "●" for control units. We fit two regression models to this data. The first is a linear regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and the second is a quadratic model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \epsilon$. The fitted values of the linear regression and quadratic regression line are given by the dashed

line and solid line respectively. The positive vertical distance between the two lines is this parametric model's treatment effect estimate.

In the raw data, some of the control units are far outside the range of the treated units, and these outlying control units are influential in the parametric models. For the data under common support, treated units are matched with control units that are close in weight (yellow units are discarded), and as a resul,t treatment effect estimates are similar regardless of model specification. The two linear and two quadratic lines also appear on the right graph (on top of one another), truncated to the location of the matched data.

A key problem that generates this model dependence is the presence of control units far outside the range of the treated units. The model estimation thus extrapolates over a range of data that do not include treated and control units and so is particularly sensitive to the set of control units that do not look similar to the treated units. These extrapolations make treatment effect estimates extremely sensitive to a small change in the statistical model (King and Zeng, 2007).

The immediate goal of finding common support ensure overlap, allowing matching methods to reduce imbalances between treatment and control pretreatment covariate distributions, without losing too many observations in the process. The result of this process, when done appropriately, is considerable reduction in model dependence, reduced potential for bias, smaller variance, and as a result, lower mean squared error.

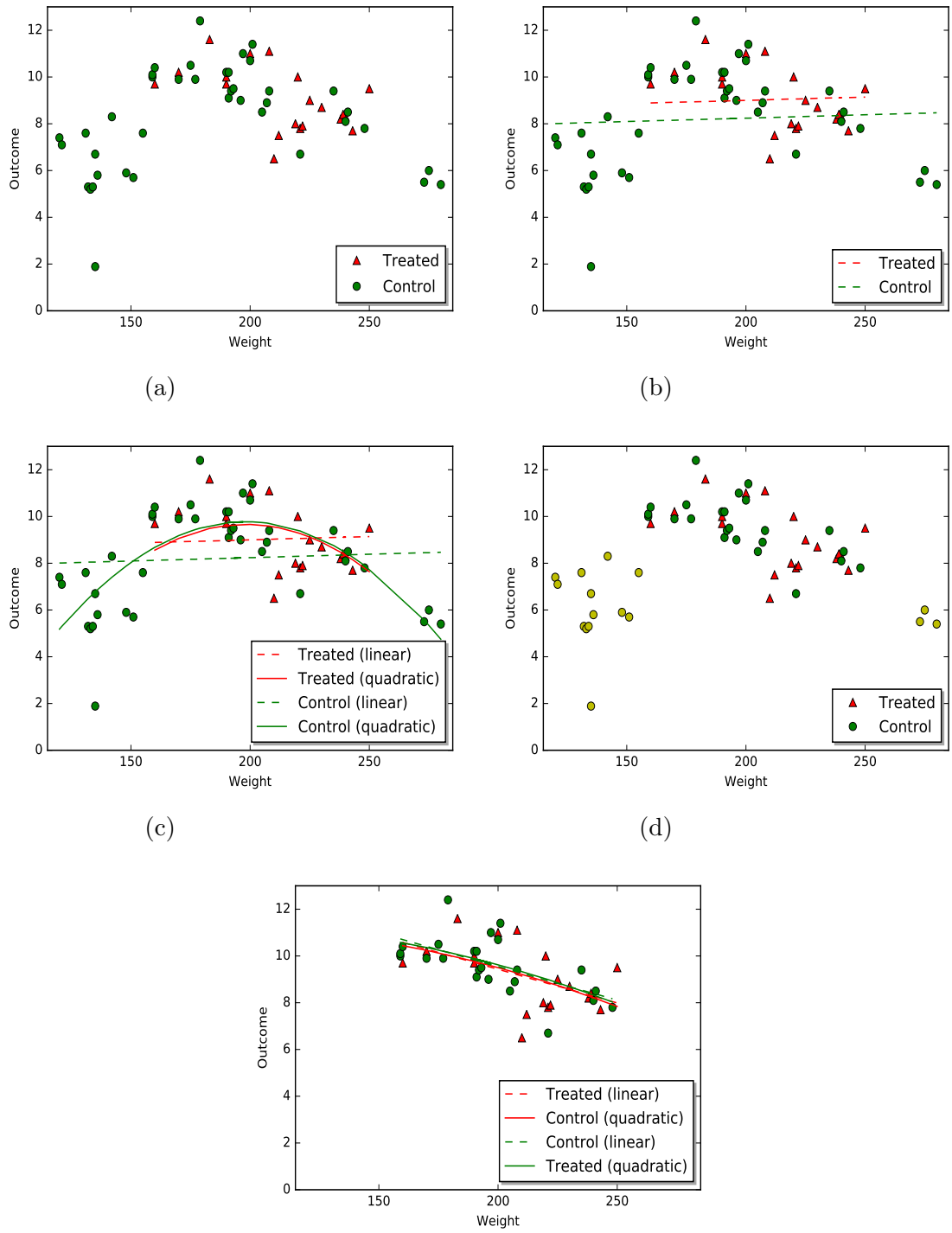


Figure 4.1: The Figures depict estimates of the treatment effect for a linear and quadratic specification, represented by the difference between parallel lines and parabolas, respectively. Red lines are fitted to the treated points and green to the controls. The solid lines are for the quadratic equation and dashed lines are for linear equation.

4.3 Future Research

In this dissertation, we focus on estimation of treatment with two categories. My future research will address more specific questions regarding the treatment heterogeneity and multiple treatment effects.

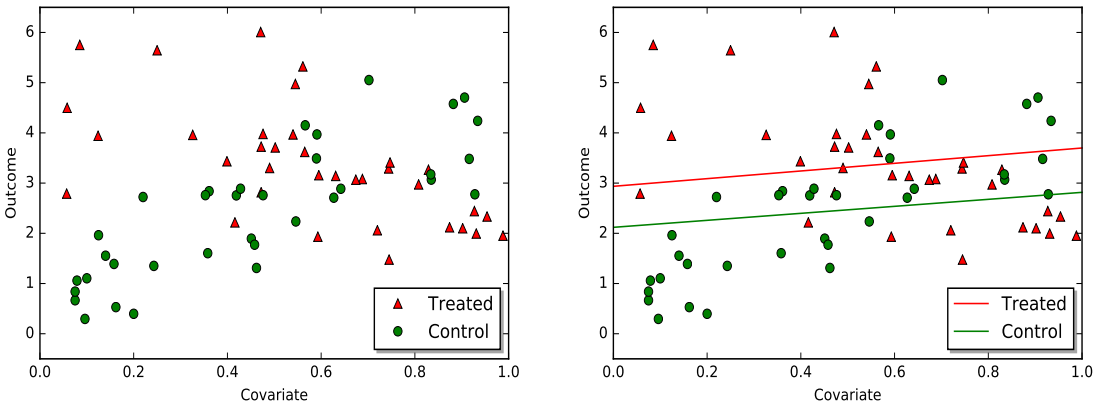
4.3.1 Treatment Heterogeneity

Often, identifying heterogeneous treatments is necessary; because of the effort and cost involved in some studies, investigators frequently use analyses of subgroups of study participants to extract as much information as possible. Heterogeneity of treatment effects in subgroups of patients may provide useful information for the care of patients and for future research (Wang et al., 2007; Gabler et al., 2009). Finding the heterogeneous treatment effects could be tricky as the effects are different under different covariate spaces. A pictorial description of treatment heterogeneity is presented in Figure 4.2. A simple linear model is parsimonious but not efficient in finding a heterogeneous treatment effect. A subgroup analysis may be appropriate in presence of heterogeneity.

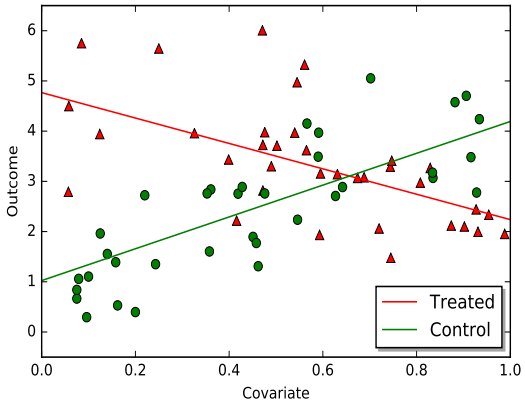
In methodological research, several approaches have been proposed to analysis heterogeneity in the data. One common method is random forests—a learning algorithm that operate by constructing a multitude of decision trees at training time and outputting the class that is the treatment effect of the individual trees (Breiman, 2001; Athey and Imbens, 2016). I would like to investigate the use of case specific random forests (CSRF) for causal inference (Xu et al., 2016). CSRF takes weighted bootstrap resamples to create individual trees for each unit to be predicted—units close to the training subject are given more weight when predicting the response of the test subject.

4.3.2 Multiple Treatments

A treatment that has more than two categories or is continuous or mixed can be complicated in terms of estimating treatment effect (Imai and van Dyk, 2004; Lechner, 1999). Those with



(a) (b)



(c)

Figure 4.2: Heterogeneity is presented in the Figures where a subgroup has a positive treatment effect whereas another subgroup has negative treatment effect. A simple linear model is inefficient to determine the treatment effect. The figure presents an interaction between treatment and the covariate.

more than treatment of interest can follow all the advice herein for one variable at a time, which would involve matching separately for each and working hard to avoid posttreatment bias in the process (Yu et al., 2013). In this dissertation, we stick to a single binary treatment since it greatly simplifies the exposition and improves intuition even for those who will ultimately use more sophisticated treatments.

Bibliography

W. Lin, Ann. Appl. Stat. **7**, 295 (2013), URL http://dx.doi.org/10.1214/12-AOAS583.

P. Rosenbaum and D. Rubin, Biometrika **70**, 41 (1983).

A. Diamond and J. S. Sekhon, Review of Economics and Statistics **95**, 932 (2012).

E. Stuart, E. DuGoff, M. Abrams, D. Salkever, and D. Steinwachs, EGEMS. **3**, 1038 (2013), URL http://doi:10.13063/2327-9214.1038.

G. King and L. Zeng, Political Analysis **14**, 131–159 (2006).

H. Beebee, C. Hitchcock, and P. Menzies, *The Oxford Handbook of Causation*, Oxford Handbooks in Philosophy (OUP Oxford, 2009), ISBN 9780199279739, URL https://books.google.com/books?id=xGnZtUtG-nIC.

G. von Giżycki and S. Coit, *An Introduction to the Study of Ethics*, Introductory science text-books (Sonnenschien, 1891), URL https://books.google.com/books?id=_mJAAQAAMAAJ.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, Statist. Sci. **5**, 465 (1990), URL http://dx.doi.org/10.1214/ss/1177012031.

P. W. Holland, Journal of the American Statistical Association **81**, 945 (1986), ISSN 01621459, URL http://dx.doi.org/10.2307/2289064.

N. S. Hall, Journal of the History of Biology **40**, 295 (2007), ISSN 1573-0387, URL http://dx.doi.org/10.1007/s10739-006-9119-z.

R. Fisher, *Statistical Methods For Research Workers*, Cosmo study guides (Cosmo Publications, 1925), ISBN 9788130701332, URL https://books.google.com/books?id=4bTttAJR5kEC.

D. Basu, Journal of the American Statistical Association **75**, 575 (1980), ISSN 01621459, URL http://www.jstor.org/stable/2287648.

D. B. Rubin, Biometrics **29** (1973), ISSN 0006341X, URL http://dx.doi.org/10.2307/2529684.

J. Heckman and R. Robb, Journal of Econometrics **30**, 239 (1985), URL http://EconPapers.repec.org/RePEc:eee:econom:v:30:y:1985:i:1-2:p:239-267.

K. Imai, G. King, and E. Stuart, Journal of the Royal Statistical Society. Series A: Statistics in Society **171**, 481 (2008), ISSN 0964-1998.

L. W. Miratrix, J. S. Sekhon, and B. Yu, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2012), URL http://dx.doi.org/10.1111/j.1467-9868.2012.01048.x.

R. Greevy, Biostatistics pp. 263–275 (2004), URL http://www.ingentaconnect.com/content/oup/biosts/2004/00000005/00000002/art00263.

M. J. Higgins, F. Sävje, and J. S. Sekhon, Proceedings of the National Academy of Sciences **113**, 7369 (2016), http://www.pnas.org/content/113/27/7369.full.pdf.

E. A. Stuart, Statistical science : a review journal of the Institute of Mathematical Statistics **25**, 1 (2010), URL http://dx.doi.org/10.1214/09-sts313.

D. B. Rubin, Health Services and Outcomes Research Methodology **2**, 169 (2001), ISSN 1572-9400, URL http://dx.doi.org/10.1023/A:1020363010465.

P. C. Austin, The Journal of Thoracic and Cardiovascular Surgery **134**, 1128 (2007), ISSN 0022-5223, URL http://www.sciencedirect.com/science/article/pii/S0022522307012433.

P. C. Austin, Statistics in Medicine **27**, 2037 (2008), ISSN 1097-0258, URL http://dx.doi.org/10.1002/sim.3150.

C. B. Fogarty, M. E. Mikkelsen, D. F. Gaieski, and D. S. Small, Journal of the American Statistical Association **111**, 447 (2016), http://dx.doi.org/10.1080/01621459.2015.1112802, URL http://dx.doi.org/10.1080/01621459.2015.1112802.

P. R. Rosenbaum, The American Statistician **59**, 147 (2005), ISSN 0003-1305 (print), 1537-2731 (electronic).

D. Ho, K. Imai, G. King, and E. Stuart, Political Analysis **15**, 199–236 (2007).

A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*, vol. Analytical methods for social research (Cambridge University Press, 2007).

R. Dehejia and S. Wahba, Journal of the American Statistical Association **94**, 1053 (1999).

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, Biometrika **96**, 187 (2009), http://biomet.oxfordjournals.org/content/96/1/187.full.pdf+html.

P. R. Rosenbaum, Journal of Computational and Graphical Statistics **21**, 57 (2012), http://dx.doi.org/10.1198/jcgs.2011.09219, URL http://dx.doi.org/10.1198/jcgs.2011.09219.

J. R. Zubizarreta, R. D. Paredes, and P. R. Rosenbaum, ArXiv e-prints (2014), 1404.3584.

J. Hill and Y.-S. Su, The Annals of Applied Statistics **7**, 1386 (2013), ISSN 19326157, URL http://www.jstor.org/stable/23566478.

J. Ramsey, S. Hanson, C. Hanson, Y. Halchenko, R. Poldrack, and C. Glymour, NeuroImage **49**, 1545 (2010), ISSN 1053-8119, URL http://www.sciencedirect.com/science/article/pii/S105381190900977X.

J. S. Hastings, T. J. Kane, and D. O. Staiger, Working Paper 12145, National Bureau of Economic Research (2006), URL http://www.nber.org/papers/w12145.

G. Anderson, O. Linton, and Y.-J. Whang, Journal of Econometrics **171**, 1 (2012), ISSN 0304-4076, URL http://www.sciencedirect.com/science/article/pii/S0304407612001108.

A. F. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, and R. M. Califf, JAMA : the journal of the American Medical Association **276**, 889 (1996), http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?

G. King and R. Nielsen, *Why propensity scores should not be used for matching* (Working Paper), URL https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching.

X. S. Gu and P. R. Rosenbaum, Journal of Computational and Graphical Statistics **2**, 405 (1993), http://www.tandfonline.com/doi/pdf/10.1080/10618600.1993.10474623, URL http://www.tandfonline.com/doi/abs/10.1080/10618600.1993.10474623.

P. R. Rosenbaum and D. B. Rubin, The American Statistician **39**, 33 (1985), http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.1985.10479383, URL http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1985.10479383.

A. Abadie and G. W. Imbens, Working Paper 283, National Bureau of Economic Research (2002), URL http://www.nber.org/papers/t0283.

J. Zubizarreta, Journal of the American Statistical Association **107**, 1360 (2012), http://dx.doi.org/10.1080/01621459.2012.703874, URL http://dx.doi.org/10.1080/01621459.2012.703874.

K. Imai and M. Ratkovic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76**, 243 (2014), ISSN 1467-9868, URL http://dx.doi.org/10.1111/rssb.12027.

M. C. Elze, J. Gregson, U. Baber, E. Williamson, S. Sartori, R. Mehran, M. Nichols, G. W. Stone, and S. J. Pocock, Journal of the American College of Cardiology **69**, 345 (2017), ISSN 0735-1097, http://www.onlinejacc.org/content/69/3/345.full.pdf, URL http://www.onlinejacc.org/content/69/3/345.

P. C. Austin, Statistics in Medicine **33**, 1057 (2014), ISSN 1097-0258, URL http://dx.doi.org/10.1002/sim.6004.

K. Ming and P. R. Rosenbaum, Biometrics **56**, 118 (2000), URL http://EconPapers.repec.org/RePEc:bla:biomet:v:56:y:2000:i:1:p:118-124.

P. R. Rosenbaum, Journal of the American Statistical Association **84**, 1024 (1989), http://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.10478868, URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478868.

B. B. Hansen, Journal of the American Statistical Association **99**, 609 (2004), http://dx.doi.org/10.1198/016214504000000647, URL http://dx.doi.org/10.1198/016214504000000647.

M. Lunt, American Journal of Epidemiology **179**, 226 (2014), /oup/backfile/content_public/journal/aje/179/2/10.1093_aje_kwt212/1/kwt212.pdf, URL +http://dx.doi.org/10.1093/aje/kwt212.

P. C. Austin, Pharmaceutical Statistics **10**, 150 (2011), ISSN 1539-1612, URL http://dx.doi.org/10.1002/pst.433.

R. Pirracchio, M. L. Petersen, and M. van der Laan, American Journal of Epidemiology **181**, 108 (2015), /oup/backfile/content_public/journal/aje/181/2/10.1093_aje_kwu253/2/kwu253.pdf, URL +http://dx.doi.org/10.1093/aje/kwu253.

G. W. Imbens, in *on Right Heart Catheterization,? Health Services and Outcomes Research Methodology* (2001), pp. 259–278.

A. Abadie and G. W. Imbens, Econometrica **74**, 235 (2006), ISSN 1468-0262, URL http://dx.doi.org/10.1111/j.1468-0262.2006.00655.x.

G. King and L. Zeng, International Studies Quarterly pp. 183–210 (2007).

R. Wang, S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen, New England Journal of Medicine **357**, 2189 (2007), pMID: 18032770, http://dx.doi.org/10.1056/NEJMsr077003, URL http://dx.doi.org/10.1056/NEJMsr077003.

N. B. Gabler, N. Duan, D. Liao, J. G. Elmore, T. G. Ganiats, and R. L. Kravitz, Trials **10**, 43 (2009), ISSN 1745-6215, URL http://dx.doi.org/10.1186/1745-6215-10-43.

L. Breiman, Machine Learning **45**, 5 (2001), ISSN 1573-0565, URL http://dx.doi.org/10.1023/A:1010933404324.

S. Athey and G. Imbens, Proceedings of the National Academy of Sciences **113**, 7353 (2016).

R. Xu, D. Nettleton, and D. J. Nordman, Journal of Computational and Graphical Statistics **25**, 49 (2016), http://dx.doi.org/10.1080/10618600.2014.983641, URL http://dx.doi.org/10.1080/10618600.2014.983641.

K. Imai and D. A. van Dyk, Journal of the American Statistical Association **99**, 854 (2004), http://dx.doi.org/10.1198/016214504000001187, URL http://dx.doi.org/10.1198/016214504000001187.

M. Lechner, IZA Discussion Papers 91, Institute for the Study of Labor (IZA) (1999), URL http://EconPapers.repec.org/RePEc:iza:izadps:dp91.

C. Yu, J. Legg, and B. Liu, Electron. J. Statist. **7**, 2737 (2013), URL http://dx.doi.org/10.1214/13-EJS856.

P. M. Aronow, D. P. Green, and D. K. K. Lee, Ann. Statist. **42**, 850 (2014), URL http://dx.doi.org/10.1214/13-AOS1200.

Appendix A

Supplement

A.1 The Estimators and Their Variances

A.1.1 The Post-Stratified Estimator

The strata-level mean and variance of potential outcomes for treatment k are:

$$\mu_{j,k} \equiv \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ijk}, \tag{A.1}$$

$$\begin{aligned} \sigma^2_{j,k} &\equiv \frac{1}{n_j} \sum_{i:i\in j}^{n_j} (y_{ijk} - \mu_{j,k})^2 \\ &= \sum_{i=1}^{n_j} \frac{y^2_{ijk}}{n_j} - \Big(\sum_{i=1}^{n_j} \frac{y_{ijk}}{n_j} \Big)^2. \end{aligned} \tag{A.2}$$

The strata-level covariance between treatment k and treatment k' is given by:

$$\gamma_{j,kk'} \equiv \sum_{i=1}^{n_j} \frac{y_{ijk} y_{ijk'}}{n_j} - \Big(\sum_{i=1}^{n_j} \frac{y_{ijk}}{n_j} \Big) \Big(\sum_{i=1}^{n_j} \frac{y_{ijk'}}{n_j} \Big), \tag{A.3}$$

and strata level correlation $\rho_{j,kk'}$ is:

$$\rho_{j,kk'} \equiv \frac{\gamma_{j,kk'}}{\sigma^2_{j,k} \sigma^2_{j,k'}}.$$

The range of $\rho_{j,kk'}$ lies between -1 to 1 i.e $-1 \leq \rho \leq 1$. The estimated correlation coefficient can be written as:

$$\hat{\rho}_{j,kk'} = \frac{\hat{\gamma}_{j,kk'}}{\hat{\sigma}^2_{j,k}\hat{\sigma}^2_{j,k'}}. \tag{A.4}$$

The strata defined by our algorithm have the stratum-specific sample average treatment effect for two treatment k and k' ($\text{SATE}_{j,kk'}$):

$$\tau_{j,kk'} \equiv \sum_{i:i\in j} \Big[\frac{y_{ijk}}{n_j} - \frac{y_{ijk'}}{n_j}\Big].$$

A.1.2 The Overall Estimator

The domain-level mean and variance of potential outcome for treatment k are:

$$\mu_k \equiv \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j} y_{ijk}, \tag{A.5}$$

$$\begin{aligned}\sigma^2_k &\equiv \sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{(y_{ijk} - \mu_k)}{n} \\ &= \sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{y^2_{ijk}}{n} - \Big(\sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{y_{ijk}}{n}\Big)^2.\end{aligned} \tag{A.6}$$

The domain level covariance between for treatment k and treatment k' is:

$$\begin{aligned}\gamma_{kk'} &\equiv \sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{(y_{ijk} - \mu_k)(y_{ijk'} - \mu_{k'})}{n} \\ &= \sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{y_{ijk}y_{ijk'}}{n} - \Big(\sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{y_{ijk}}{n}\Big)\Big(\sum_{j=1}^{s}\sum_{i=1}^{n_j} \frac{y_{ijk'}}{n_j}\Big).\end{aligned} \tag{A.7}$$

We also refer to the correlation of potential outcomes ρ, where

$$\rho \equiv \frac{\gamma_{kk'}}{\sigma^2_k\sigma^2_{k'}}.$$

A post-stratification adjusted estimate is is the sum of weighted treatment effect of the strata, written as:

$$\tau_{kk'} \equiv \sum_{j=1}^{s} \frac{n_j}{n} \tau_{j,kk'}.$$

A.1.3 The Estimate

The estimate of the mean and variance for treatment k in jth stratum is:

$$\hat{\mu}_{j,k} = \sum_{i:i\in j} \frac{y_{ijk} T_{ijk}}{n_{jk}}, \tag{A.8}$$

$$\hat{\sigma}^2_{j,kk'} = \frac{n_j - 1}{n_j} \sum_{i:i\in j} \frac{T_{ijk}\big(y_{ijk} - \sum_{i:i\in j} \frac{y_{ijk} T_{ijk}}{n_{jk}}\big)^2}{n_{jk} - 1}$$

The simple-difference estimator of stratum-specific sample average treatment effect for two treatment k and k' ($\text{SATE}_{j,kk'}$):

$$\begin{aligned}\hat{\tau}_{j,kk'} &\equiv \sum_{i:i\in j} \Big[\frac{y_{ijk} T_{ijk}}{n_{jk}} - \frac{y_{ijk'} T_{ijk'}}{n_{jk'}}\Big] \\ &= \bar{y}_{\cdot jk} - \bar{y}_{\cdot jk'},\end{aligned}$$

where n_{jk} and $n_{jk'}$ denotes the number of units in jth stratum under kth treatment and k'th treatment respectively. The estimator is undefined when $n_{jk} = 0$ or $n_{jk'} = 0$, $\forall j$, i.e. each stratum consists of at least one unit assigned to each treatment. The post-stratification estimate can be obtained as a weighted average of the estimates of $\text{SATE}_{j,kk'}$s:

$$\hat{\tau}_{kk'} = \sum_{j=1}^{s} w_j \hat{\tau}_{j,kk'},$$

where $w_j = n_j/n$, the proportion of units assigned to treatment k and k' in stratum j. These estimators are shown to be unbiased.

A.1.4 Unbiasedness

First we find

$$\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}}\Big) = \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}}|n_{jk}\Big)\Big] = \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}}{n_{jk}}\Big) = \mathbb{E}\Big(\frac{1}{n_j}\Big) = \frac{1}{n_j}.$$

The strata-level estimators are unbiased:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{j,kk'}) &= \mathbb{E}\Big[\sum_{i:i\in j}\Big(\frac{y_{ijk}T_{ijk}}{n_{jk}} - \frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\Big)\Big] \\
&= \sum_{i:i\in j}\mathbb{E}\Big(\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big) - \sum_{i:i\in j}\mathbb{E}\Big(\frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\Big) \\
&= \sum_{i:i\in j} y_{ijk}\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}}\Big) - \sum_{i:i\in j} y_{ijk'}\mathbb{E}\Big(\frac{T_{ijk'}}{n_{jk'}}\Big) \\
&= \sum_{i:i\in j} y_{ijk}\frac{1}{n_j} - \sum_{i:i\in j} y_{ijk'}\frac{1}{n_j} \\
&= \tau_{j,kk'}.
\end{aligned} \tag{A.9}$$

The post-stratification estimator $\hat{\tau}_{kk'}$ is unbiased:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{kk'}) &= \mathbb{E}\Big(\sum_j \frac{n_j}{n}\hat{\tau}_{j,kk'}\Big) \\
&= \sum_j \frac{n_j}{n}\mathbb{E}(\hat{\tau}_{j,kk'}) \\
&= \sum_j \frac{n_j}{n}\tau_{j,kk'} \\
&= \tau_{kk'}.
\end{aligned} \tag{A.10}$$

A.1.5 Variance

First we find the following expectations:

$$\mathbb{E}\Big(\frac{T_{ijk}^2}{n_{jk}^2}\Big) = \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}^2}|n_{jk}\Big)\Big] = \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}}{n_{jk}^2}\Big) = \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}}\Big), \tag{A.11}$$

$$
\begin{aligned}
\mathbb{E}\Big(\frac{T_{ijk}T_{ijk'}}{n_{jk}^2}\Big) &= \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}T_{ijk'}}{n_{jk}^2}|n_{jk}\Big)\Big] \\
&= \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}\frac{n_{jk}-1}{n_j-1}}{n_{jk}^2}\Big) \\
&= \frac{1}{n_j(n_j-1)}\mathbb{E}\Big(\frac{n_{jk}-1}{n_{jk}}\Big) \\
&= \frac{1}{n_j(n_j-1)}\Big(1-\mathbb{E}\frac{1}{n_{jk}}\Big) \\
&= \frac{1}{n_j(n_j-1)} - \frac{1}{n_j(n_j-1)}\Big(\mathbb{E}\frac{1}{n_{jk}}\Big),
\end{aligned}
\tag{A.12}
$$

$$
\begin{aligned}
\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk'}}{n_{jk}n_{jk'}}\Big) &= \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk'}}{n_{jk}n_{jk'}}|n_{jk},n_{jk'}\Big)\Big] \\
&= \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}\frac{n_{jk'}}{n_j-1}}{n_{jk}n_{jk'}}\Big) \\
&= \frac{1}{n_j(n_j-1)}.
\end{aligned}
\tag{A.13}
$$

The variance of $\hat{\tau}_{j,kk'}$ can be written as:

$$
\begin{aligned}
\mathrm{Var}(\hat{\tau}_{j,kk'}) &= \mathbb{E}(\hat{\tau}_{j,kk'}^2) - [\mathbb{E}(\tau_{j,kk'})]^2 \\
&= \mathbb{E}(\hat{\tau}_{j,kk'}^2) - \tau_{j,kk'}^2.
\end{aligned}
\tag{A.14}
$$

The first part of (A.14) can be written as:

$$
\mathbb{E}(\hat{\tau}_{j,kk'}^2) = \mathbb{E}\big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\big)^2 + \mathbb{E}\big(\sum_{i:i\in j}\frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\big)^2 - 2\,\mathbb{E}\Big[\big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\big)\big(\sum_{i:i\in j}\frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\big)\Big].
\tag{A.15}
$$

Using (A.11) and (A.12), the first part in (A.15) can be written as:

$$
\begin{aligned}
\mathbb{E}\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2 &= \mathbb{E}\Big(\sum_{i:i\in j}\frac{y_{ijk}^2T_{ijk}^2}{n_{jk}^2}+\sum_{i:i\in j}\frac{y_{ijk}y_{i'jk}T_{ijk}T_{i'jk}}{n_{jk}^2}\Big)\\
&= \sum_{i:i\in j}y_{ijk}^2\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}^2}\Big)+\sum_{i:i\in j}y_{ijk}y_{i'jk}\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk}}{n_{jk}^2}\Big)\\
&= \sum_{i:i\in j}\frac{y_{ijk}^2}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)+\sum_{i:i\in j}y_{ijk}y_{i'jk}\Big[\frac{1}{n_j(n_j-1)}-\frac{1}{n_j(n_j-1)}\Big(\mathbb{E}\frac{1}{n_{jk}}\Big)\Big]\\
&= \sum_{i:i\in j}\frac{y_{ijk}^2}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)+\big[(\sum_{i:i\in j}y_{ijk})^2-\sum_{i:i\in j}y_{ijk}^2\big]\Big[\frac{1}{n_j(n_j-1)}-\frac{1}{n_j(n_j-1)}\Big(\mathbb{E}\frac{1}{n_{jk}}\Big)\Big]\\
&= \frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\Big[\Big(\sum_{i:i\in j}\frac{y_{ijk}^2}{n_j}\Big)+\frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j(n_j-1)}\Big]\\
&= \frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\Big[\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j-1}-\frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}\Big]\\
&= \frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\Big[\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j}-\Big(\frac{\sum_{i:i\in j}y_{ijk}}{n_j}\Big)^2\Big]\\
&= \frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}^2}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\sigma_{j,k}^2.
\end{aligned}
\tag{A.16}
$$

The second part of (A.15) can be written as similarly as (A.16). Using (A.13), third part of (A.15) can be written as:

$$
\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\Big)\Big] &= \mathbb{E}\Big(\sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}T_{ijk}T_{i'jk'}}{n_{jk}n_{jk'}}\Big)+\mathbb{E}\Big(\sum_{i:i\in j}\frac{y_{ijk}y_{ijk'}T_{ijk}T_{ijk'}}{n_{jk}n_{jk'}}\Big)\\
&= \sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}y_{ijk}y_{i'jk'}\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk'}}{n_{jk}n_{jk'}}\Big)+\sum_{i:i\in j}y_{ijk}y_{ijk'}\mathbb{E}\Big(\frac{T_{ijk}T_{ijk'}}{n_{jk}n_{jk'}}\Big)\\
&= \sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)},
\end{aligned}
\tag{A.17}
$$

since $\mathbb{E}\Big(\frac{T_{ijk}T_{ijk'}}{n_{jk}n_{jk'}}\Big)=0$.

Now (A.15) can be written as:

$$\begin{aligned}\mathbb{E}(\hat{\tau}^2_{j,kk'}) &= \mathbb{E}\big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\big)^2+\mathbb{E}\big(\sum_{i:i\in j}\frac{y_{ijk'}T_{ijk'}}{n_{jk}}\big)^2-2\,\mathbb{E}\Big[\big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\big)\big(\sum_{i:i\in j}\frac{y_{ijk'}T_{ijk'}}{n_{jk'}}\big)\Big]\\ &= \frac{(\sum_{i:i\in j}y_{ijk})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y^2_{ijk}}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k}\\ &\quad+\frac{(\sum_{i:i\in j}y_{ijk'})^2}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y^2_{ijk'}}{n_j(n_j-1)}+\mathbb{E}\Big(\frac{1}{n_{jk'}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k'}\\ &\quad-2\sum_{i:i\in j}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)},\end{aligned} \tag{A.18}$$

The second part of (A.14) can be written as:

$$\tau^2_{j,kk'}=\big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\big)^2+\big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\big)^2-2\,\big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\big)\big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\big). \tag{A.19}$$

Putting equation (A.18) and (A.19) in (A.14) we have:

$$
\begin{aligned}
\mathrm{Var}(\hat{\tau}_{j,kk'}) &= \mathbb{E}(\hat{\tau}^2_{j,kk'}) - \tau^2_{j,kk'} \\
&= \frac{n_j}{n_j-1}\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)^2 - \frac{\sum_{i:i\in j} y^2_{ijk}}{n_j(n_j-1)} + \mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k} \\
&\quad + \frac{n_j}{n_j-1}\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)^2 - \frac{\sum_{i:i\in j} y^2_{ijk'}}{n_j(n_j-1)} + \mathbb{E}\Big(\frac{1}{n_{jk'}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k'} \\
&\quad - 2\sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)} - \Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)^2 - \Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)^2 + 2\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big) \\
&= \frac{-1}{n_j-1}\Big[\frac{\sum_{i:i\in j} y^2_{ijk}}{n_j} - \Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)^2\Big] + \mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k} \\
&\quad + \frac{-1}{n_j-1}\Big[\frac{\sum_{i:i\in j} y^2_{ijk'}}{n_j} - \Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)^2\Big] + \mathbb{E}\Big(\frac{1}{n_{jk'}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k'} \\
&\quad + 2\Big[\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big) - \sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)}\Big] \\
&= \mathbb{E}\Big(\frac{1}{n_{jk}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k} - \frac{1}{n_j-1}\sigma^2_{j,k} + \mathbb{E}\Big(\frac{1}{n_{jk'}}\Big)\frac{n_j}{n_j-1}\sigma^2_{j,k'} - \frac{1}{n_j-1}\sigma^2_{j,k'} \\
&\quad + 2\Big[\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big) - \sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)}\Big].
\end{aligned}
\tag{A.20}
$$

Now using (A.3), the last part of (A.20) can be written as:

$$
\begin{aligned}
&\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)-\sum_{i:i\in j}\sum_{\substack{i':i'\in j\\ i'\neq i}}\frac{y_{ijk}y_{i'jk'}}{n_j(n_j-1)}\\
&=\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)-\Big[\frac{\sum_{i:i\in j}y_{ijk}\sum_{i:i\in j}y_{ijk'}}{n_j(n_j-1)}-\frac{\sum_{i:i\in j}y_{ijk}y_{ijk'}}{n_j(n_j-1)}\Big]\\
&=\Big(1-\frac{n_j}{n_j-1}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)+\frac{\sum_{i:i\in j}y_{ijk}y_{ijk'}}{n_j(n_j-1)}\\
&=-\frac{1}{n_j-1}\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)+\frac{\sum_{i:i\in j}y_{ijk}y_{ijk'}}{n_j(n_j-1)}\\
&=\frac{1}{n_j-1}\Big[\frac{\sum_{i:i\in j}y_{ijk}y_{ijk'}}{n_j}-\Big(\sum_{i:i\in j}\frac{y_{ijk}}{n_j}\Big)\Big(\sum_{i:i\in j}\frac{y_{ijk'}}{n_j}\Big)\Big]\\
&=\frac{\gamma_{j,kk'}}{n_j-1}.
\end{aligned}
\tag{A.21}
$$

Finally we have:

$$
\begin{aligned}
\mathrm{Var}(\hat{\tau}_{j,kk'})&=\Big[\mathbb{E}\Big(\frac{n_j}{n_{jk}}\Big)\frac{\sigma^2_{j,k}}{n_j-1}-\frac{\sigma^2_{j,k}}{n_j-1}\Big]+\Big[\mathbb{E}\Big(\frac{n_j}{n_{jk'}}\Big)\frac{\sigma^2_{j,k'}}{n_j-1}-\frac{\sigma^2_{j,k'}}{n_j-1}\Big]+2\ \frac{\gamma_{j,kk'}}{n_j-1}\\
&=\frac{1}{n_j-1}\Big[\mathbb{E}\Big(\frac{n_j-n_{jk}}{n_{jk}}\Big)\sigma^2_{j,k}+\mathbb{E}\Big(\frac{n_j-n_{jk'}}{n_{jk'}}\Big)\sigma^2_{j,k'}+2\ \gamma_{j,kk'}\Big].
\end{aligned}
\tag{A.22}
$$

A.1.6 Variance Estimation

First we find the following expectations:

$$
\begin{aligned}
\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}-1}\Big)&=\mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}-1}|n_{jk}\Big)\Big]\\
&=\mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}}{n_{jk}-1}\Big)\\
&=\frac{1}{n_j}\mathbb{E}\Big(\frac{n_{jk}}{n_{jk}-1}\Big)\\
&=\frac{1}{n_j}\mathbb{E}\Big(\frac{n_{jk}-1+1}{n_{jk}-1}\Big)\\
&=\frac{1}{n_j}+\frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big),
\end{aligned}
\tag{A.23}
$$

$$
\begin{aligned}
\mathbb{E}\Big[\frac{T_{ijk}}{n_{jk}(n_{jk}-1)})\Big] &= \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}(n_{jk}-1)}|n_{jk}\Big)\Big] \\
&= \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}}{n_{jk}(n_{jk}-1)}\Big) \\
&= \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big),
\end{aligned} \tag{A.24}
$$

$$
\begin{aligned}
\mathbb{E}\Big[\frac{T_{ijk}T_{i'jk}}{n_{jk}(n_{jk}-1)})\Big] &= \mathbb{E}\Big[\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk}}{n_{jk}(n_{jk}-1)}|n_{jk}\Big)\Big] \\
&= \mathbb{E}\Big(\frac{\frac{n_{jk}}{n_j}\frac{n_{jk}-1}{n_j-1}}{n_{jk}(n_{jk}-1)}\Big) \\
&= \mathbb{E}\Big(\frac{1}{n_j(n_j-1)}\Big) \\
&= \frac{1}{n_j(1-n_j)}.
\end{aligned} \tag{A.25}
$$

Now we show that $\mathbb{E}(\hat{\sigma}^2_{j,k}) = \sigma^2_{j,k}$.

First, note that:

$$
\begin{aligned}
&\sum_{i:i\in j} T_{ijk}\Big(y_{ijk}T_{ijk} - \sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2 \\
&= \sum_{i:i\in j} T_{ijk}(y_{ijk}T_{ijk})^2 - 2\sum_{i:i\in j} T_{ijk}\Big(y_{ijk}T_{ijk}\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big) + \sum_{i:i\in j} T_{ijk}\Big(\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2 \\
&= \sum_{i:i\in j} y^2_{ijk}T_{ijk} - 2\sum_{i:i\in j}\Big(y_{ijk}T_{ijk}\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big) + \sum_{i:i\in j} T_{ijk}\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2 \\
&= \sum_{i:i\in j} y^2_{ijk}T_{ijk} - 2\ n_{jk}\sum_{i:i\in j}\Big(\frac{y_{ijk}T_{ijk}}{n_{jk}}\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big) + n_{jk}\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2 \\
&= \sum_{i:i\in j} y^2_{ijk}T_{ijk} - n_{jk}\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2.
\end{aligned} \tag{A.26}
$$

Now,

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2_{j,k}) &= \mathbb{E}\Big[\frac{n_j-1}{n_j}\sum_{i:i\in j}\frac{T_{ijk}\big(y_{ijk}-\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\big)^2}{n_{jk}-1}\Big]\\ &= \frac{n_j-1}{n_j}\mathbb{E}\Big[\frac{\sum_{i:i\in j}y^2_{ijk}T_{ijk}-n_{jk}\Big(\sum_{i:i\in j}\frac{y_{ijk}T_{ijk}}{n_{jk}}\Big)^2}{n_{jk}-1}\Big]\\ &= \frac{n_j-1}{n_j}\Big[\sum_{i:i\in j}y^2_{ijk}\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}-1}\Big)-\mathbb{E}\Big(\frac{(\sum_{i:i\in j}y_{ijk}T_{ijk})^2}{n_{jk}(n_{jk}-1)}\Big)\Big].\end{aligned} \tag{A.27}$$

By (A.23):

$$\begin{aligned}\sum_{i:i\in j}y^2_{ijk}\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}-1}\Big) &= \sum_{i:i\in j}y^2_{ijk}\Big[\frac{1}{n_j}+\frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\Big]\\ &= \frac{1}{n_j}\sum_{i:i\in j}y^2_{ijk}+\frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j}y^2_{ijk}.\end{aligned} \tag{A.28}$$

By (A.24) and (A.25), it follows that:

$$\begin{aligned}\mathbb{E}\Big(\frac{(\sum_{i:i\in j}y_{ijk}T_{ijk})^2}{n_{jk}(n_{jk}-1)}\Big) &= \mathbb{E}\Big(\frac{(\sum_{i:i\in j}y_{ijk}T_{ijk})^2}{n_{jk}(n_{jk}-1)}\Big)+\mathbb{E}\Big(\frac{\sum_{i:i\in j}\sum_{i':i'\neq i}y_{ijk}y_{i'jk}T_{ijk}T_{i'jk}}{n_{jk}(n_{jk}-1)}\Big)\\ &= \sum_{i:i\in j}y^2_{ijk}\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}(n_{jk}-1)}\Big)+\sum_{i:i\in j}\sum_{i':i'\neq i}y_{ijk}y_{i'jk}\mathbb{E}\Big(\frac{T_{ijk}T_{i'jk}}{n_{jk}(n_{jk}-1)}\Big)\\ &= \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j}y^2_{ijk}+\frac{1}{n_j(n_j-1)}\sum_{i:i\in j}\sum_{i':i'\neq i}y_{ijk}y_{i'jk}\\ &= \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j}y^2_{ijk}+\frac{1}{n_j(n_j-1)}\Big((\sum_{i:i\in j}y_{ijk})^2-\sum_{i:i\in j}y^2_{ijk}\Big)\\ &= \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j}y^2_{ijk}+\frac{1}{n_j(n_j-1)}\Big(\sum_{i:i\in j}y_{ijk}\Big)^2-\frac{1}{n_j(n_j-1)}\sum_{i:i\in j}y^2_{ijk}.\end{aligned} \tag{A.29}$$

Finally, by (A.28) and (A.29), it follows that:

$$
\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{j,k}) &= \frac{n_j-1}{n_j}\Big[\sum_{i:i\in j} y^2_{ijk}\mathbb{E}\Big(\frac{T_{ijk}}{n_{jk}-1}\Big) - \mathbb{E}\Big(\frac{\sum_{i:i\in j} y_{ijk}T_{ijk}}{n_{jk}(n_{jk}-1)}\Big)\Big] \\
&= \frac{n_j-1}{n_j}\Big[\frac{1}{n_j}\sum_{i:i\in j} y^2_{ijk} + \frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j} y^2_{ijk}\Big] \\
&\quad - \frac{n_j-1}{n_j}\Big[\frac{1}{n_j}\mathbb{E}\Big(\frac{1}{n_{jk}-1}\Big)\sum_{i:i\in j} y^2_{ijk} + \frac{1}{n_j(n_j-1)}\Big(\sum_{i:i\in j} y_{ijk}\Big)^2 - \frac{1}{n_j(n_j-1)}\sum_{i:i\in j} y^2_{ijk}\Big] \\
&= \frac{n_j-1}{n_j}\Big[\frac{1}{n_j-1}\sum_{i:i\in j} y^2_{ijk} - \frac{1}{n_j(n_j-1)}\Big(\sum_{i:i\in j} y_{ijk}\Big)^2\Big] \\
&= \frac{n_j-1}{n_j}\Big[\frac{n_j}{n_j-1}\sum_{i:i\in j} \frac{y^2_{ijk}}{n_j} - \frac{1}{n_j(n_j-1)}\Big(\sum_{i:i\in j} y_{ijk}\Big)^2\Big] \\
&= \frac{n_j-1}{n_j}\Big(\frac{n_j}{n_j-1}\sigma^2_{j,k}\Big) \\
&= \sigma^2_{j,k}.
\end{aligned}
\tag{A.30}
$$

Thus the estimate of (A.22) can be written as:

$$
\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'}) = \frac{1}{n_j-1}\Big[\mathbb{E}\Big(\frac{n_j-n_{jk}}{n_{jk}}\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j-n_{jk'}}{n_{jk'}}\Big)\hat{\sigma}^2_{j,k'} + 2\,\hat{\gamma}_{j,kk'}\Big]. \tag{A.31}
$$

The first and second terms in (A.31) is positive due to $\mathbb{E}\Big(\frac{n_j}{n_{jk}}\Big) \geq 1$. If there is no correlation between the kth and k'th treatment, the last term becomes zero.

A.2 Bounds on the Variance

A.2.1 Conventional Bounds on the Variance

It has been known since Splawa-Neyman et al. (1990) that neither unbiased nor consistent variance estimation is generally possible to estimate the variance of the treatment effect using stratification, due to the fact that the joint distribution of the potential outcomes can never be fully recovered from data. Also unbiased variance estimation is not generally possible

when $n_j < \infty$.

We have the the estimator of $\mathrm{Var}(\hat{\tau}_{j,kk'})$ that uses Cauchy-Schwarz and the AM-GM inequalities

$$\hat{\gamma}_{j,kk'} \leq \sqrt{\hat{\sigma}^2_{j,k}\hat{\sigma}^2_{j,k'}} \leq \frac{\hat{\sigma}^2_{j,k} + \hat{\sigma}^2_{j,k'}}{2}. \tag{A.32}$$

Using (A.32) in (A.31) the conventional upper bound of the estimate of the variance is:

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^C &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j}{n_{jk}} - 1\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j}{n_{jk'}} - 1\Big)\hat{\sigma}^2_{j,k'} + \hat{\sigma}^2_{j,k} + \hat{\sigma}^2_{j,k'}\Big] \\ &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j}{n_{jk}}\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j}{n_{jk'}}\Big)\hat{\sigma}^2_{j,k'}\Big].\end{aligned} \tag{A.33}$$

Since $\mathbb{E}(\hat{\sigma}^2_{j,k}) = \sigma^2_{j,k}$ and $\mathbb{E}(\hat{\sigma}^2_{j,k'}) = \sigma^2_{j,k'}$, $\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^C$ is conservative as its bias is nonnegative:

$$\mathbb{E}\big(\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^C - \mathrm{Var}(\hat{\tau}_{j,kk'})\big) = \frac{1}{n_j - 1}[\sigma^2{}_{j,k} + \sigma^2{}_{j,k'} - 2\gamma_{j,kk'}] \geq 0.$$

Again by (A.4), the last part in (A.31) can be bound by $-\{\hat{\sigma}^2_{j,k}\hat{\sigma}^2_{j,k'}\}^{\frac{1}{2}} \leq \hat{\gamma}_{j,kk'} \leq \{\hat{\sigma}^2_{j,k}\hat{\sigma}^2_{j,k'}\}^{\frac{1}{2}}$. Thus we can estimate the bound of the variance of the treatment effect by

$$\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^{N\pm} = \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j - n_{jk}}{n_{jk}}\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j - n_{jk'}}{n_{jk'}}\Big)\hat{\sigma}^2_{j,k'} \pm 2\,\{\hat{\sigma}^2_{j,k}\hat{\sigma}^2_{j,k'}\}^{\frac{1}{2}}\Big]. \tag{A.34}$$

The plus or minus sign is chosen depending on whether an upper or a lower bound estimate is desired. Aronow et al. (2014) showed by simulation that $\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^{N\pm}$ provides often narrower than intervals produced by $\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^C$ and it can be obtained much narrower interval by sharp bounds on the variance.

A.2.2 Sharp Bounds on the Variance

Let us consider estimates for the marginal distribution of y_k and $y_{k'}$ exist and can be used to obtain asymptotically sharp bounds on $\mathrm{Var}(\hat{\tau}_{j,kk'})$ given the information available. Let $F_k(y) = \frac{1}{n_j}\sum_{i:i\in j} I(y_{ijk} \leq y)$ and $F_{k'}(y) = \frac{1}{n_j}\sum_{i:i\in j} I(y_{ijk'} \leq y)$ be the marginal distribution functions of y_k and $y_{k'}$,respectively. Define their left-continuous inverses as $F^{-1}_k(u) = \inf\{y :$

$F_k(y) \geq u\}$ and $F_{k'}^{-1}(u) = \inf\{y : F_{k'}(y) \geq u\}$. We define

$$\begin{aligned}\gamma_{j,kk'}^{S+} &= \int F_k^{-1}(u)F_{k'}^{-1}(u)\mathrm{d}u - \mu_{j,k}\mu_{j,k'},\\ \gamma_{j,kk'}^{S-} &= \int F_k^{-1}(u)F_{k'}^{-1}(1-u)\mathrm{d}u - \mu_{j,k}\mu_{j,k'}.\end{aligned} \tag{A.35}$$

Lemma A.2.1 (Hoeffding). Given only F_k and $F_{k'}$ and no other information on the joint distribution of $(y_k, y_{k'})$, the bound

$$\gamma_{j,kk'}^{S-} \leq \gamma_{j,kk'} \leq \gamma_{j,kk'}^{S+}$$

is sharp. The upper bound is attained if y_k and $y_{k'}$ are comonotonic, i.e., $(y_k, y_{k'}) \sim \{F_k^{-1}(U), F_{k'}^{-1}(U)\}$ for a uniform random variable U on $[0, 1]$. The lower bound is attained if y_k and $y_{k'}$ are countermonotonic, i.e., $(y_k, y_{k'}) \sim \{F_k^{-1}(U), F_{k'}^{-1}(1-U)\}$.

Proof. Let $H(y_k, y_{k'})$ be the joint distribution function of $(y_k, y_{k'})$ and define two other distributions $H^{S+}(y_k, y_{k'}) = \min\{F_k(y), F_{k'}(y)\}$ and $H^{S-}(y_k, y_{k'}) = \max\{0, F_k(y) + F_{k'}(y) - 1\}$. All three distributions have the same marginals F_k and $F_{k'}$. a result by Hoeffding shows that

$$\mathbb{E}_{H^{S-}}(y_k y_{k'}) \leq \mathbb{E}_H(y_k y_{k'}) \leq \mathbb{E}_{H^{S+}}(y_k y_{k'})$$

Since $\{F_k^{-1}(U), F_{k'}^{-1}(U)\} \sim H^{S+}$ and $\{F_k^{-1}(U), F_{k'}^{-1}(1-U)\} \sim H^{S-}$, the lower and upper bounds are equivalent to

$$\begin{aligned}\mathbb{E}_{H^{S+}}(y_k y_{k'}) &= \int_0^1 F_k^{-1}(u)F_{k'}^{-1}(u)\mathrm{d}u,\\ \mathbb{E}_{H^{S-}}(y_k y_{k'}) &= \int_0^1 F_k^{-1}(u)F_{k'}^{-1}(1-u)\mathrm{d}u.\end{aligned}$$

The integrals exist because $|F_k^{-1}(u)|, |F_{k'}^{-1}(u)| \leq \max_{i=1}^{n_j} \max(|y_{ijk}|, |y_{ijk'}|) < \infty$. □

Lemma A.2.1 implies that $[\gamma^{S+}_{j,kk'}, \gamma^{S-}_{j,kk'}]$ is the sharpest interval bound for $\gamma_{kk'}$

$$\begin{aligned}
\mathrm{Var}(\tau_{j,kk'})^{S+} &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j - n_{jk}}{n_{jk}}\Big)\sigma^2_{j,k} + \mathbb{E}\Big(\frac{n_j - n_{jk'}}{n_{jk'}}\Big)\sigma^2_{j,k'} + 2\ \gamma^{S+}_{j,kk'}\Big], \\
\mathrm{Var}(\tau_{j,kk'})^{S-} &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j - n_{jk}}{n_{jk}}\Big)\sigma^2_{j,k} + \mathbb{E}\Big(\frac{n_j - n_{jk'}}{n_{jk'}}\Big)\sigma^2_{j,k'} + 2\ \gamma^{S-}_{j,kk'}\Big].
\end{aligned} \tag{A.36}$$

In practice, we observe neither F_k nor $F_{k'}$, but rather their estimates $\hat{F}_k(y) = \frac{1}{n_{jk}}\sum_{i=1}^{n_j} T_{ijk} I(y_{ijk} \leq y)$, $\hat{F}_{k'}(y) = \frac{1}{n_{jk'}}\sum_{i=1}^{n_j} T_{ijk'} I(y_{ijk'} \leq y)$ and left-continuous inverses

$$\begin{aligned}
\hat{F}^{-1}_k(u) &= \inf\{y : \hat{F}_k(y) \geq u\} = y_{k(\lceil n_{jk} u \rceil)}, \\
\hat{F}^{-1}_{k'}(u) &= \inf\{y : \hat{F}_{k'}(y) \geq u\} = y_{k'(n_{jk} + \lceil n_{jk'} u \rceil)},
\end{aligned}$$

where $y_{k(1)} \leq \cdots \leq y_{k(n_{jk})}$ and $y_{k'(n_{jk}+1)} \leq \cdots \leq y_{k'(n_{jk'}+n_{jk'})}$ are the ordered observed outcomes, and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Substituting $(\hat{F}_k, \hat{F}_{k'})$ for $(F_k, F_{k'})$ in (A.35) yields an interval estimator $[\hat{\gamma}^{S-}_{j,kk'}, \hat{\gamma}^{S+}_{j,kk'}]$ for $\gamma_{j,kk'}$:

$$\begin{aligned}
\hat{\gamma}^{S+}_{j,kk'} &= \int \hat{F}^{-1}_k(u)\hat{F}^{-1}_{k'}(u)\mathrm{d}u - \hat{\mu}_{j,k}\hat{\mu}_{j,k'}, \\
\hat{\gamma}^{S-}_{j,kk'} &= \int \hat{F}^{-1}_k(u)\hat{F}^{-1}_{k'}(1-u)\mathrm{d}u - \hat{\mu}_{j,k}\hat{\mu}_{j,k'}.
\end{aligned}$$

Let the $[0,1]$-partition $\mathcal{P} = \{p_0, p_1, \ldots, p_P\}$ be the ordered distinct elements of $\{0, \frac{1}{n_{jk}}, \frac{2}{n_{jk}} \ldots, 1\} \cup \{0, \frac{1}{n_{jk'}}, \frac{2}{n_{jk'}}, \ldots, 1\}$. Let $y_{k\lceil i \rceil} = y_{k(\lceil n_{jk} p_i \rceil)}$ and $y_{k'\lceil i \rceil} = y_{k'(n_{jk} + \lceil n_{jk'} p_i \rceil)}$. The inverses $\hat{F}^{-1}_k$ and $\hat{F}^{-1}_{k'}$ are piecewise constant since $\hat{F}^{-1}_k(u) = y_{k\lceil i \rceil}$ and $\hat{F}^{-1}_{k'}(u) = y_{k'\lceil i \rceil}$ for $u \in (p_{i-1}, p_i]$. In addition, the symmetry $p_i = 1 - p_{P-i}$ implies that $p_i - p_{i-1} = p_{P+1-i} - p_{P-i}$. Thus $[\hat{\gamma}^{S-}_{j,kk'}, \hat{\gamma}^{S+}_{j,kk'}]$ reduces to

$$\begin{aligned}
\hat{\gamma}^{S+}_{j,kk'} &= \sum_{i=1}^{P}(p_i - p_{i-1}) y_{k\lceil i \rceil} y_{k'\lceil i \rceil} - \hat{\mu}_{j,k}\hat{\mu}_{j,k'}, \\
\hat{\gamma}^{S-}_{j,kk'} &= \sum_{i=1}^{P}(p_i - p_{i-1}) y_{k\lceil i \rceil} y_{k'\lceil P+1-i \rceil} - \hat{\mu}_{j,k}\hat{\mu}_{j,k'}.
\end{aligned} \tag{A.37}$$

where $\hat{\mu}_{j,k}$ and $\hat{\mu}_{j,k'}$ are as defined in (A.8).

Substituting $\hat{\sigma}^2_{j,k}$, $\hat{\sigma}^2_{j,k'}$, and (A.37) for $\{\sigma^2_{j,k}, \sigma^2_{j,k'}, \gamma_{j,kk'}\}$ in the expressions for $\gamma^{S-}_{j,kk'}$ and $\gamma^{S+}_{j,kk'}$, we obtain the interval estimator $[\hat{\gamma}^{S-}_{j,kk'}, \hat{\gamma}^{S+}_{j,kk'}]$ for $\gamma_{j,kk'}$:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^{S+} &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j - n_{jk}}{n_{jk}}\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j - n_{jk'}}{n_{jk'}}\Big)\hat{\sigma}^2_{j,k'} + 2\ \hat{\gamma}^{S+}_{j,kk'}\Big], \\
\widehat{\mathrm{Var}}(\hat{\tau}_{j,kk'})^{S-} &= \frac{1}{n_j - 1}\Big[\mathbb{E}\Big(\frac{n_j - n_{jk}}{n_{jk}}\Big)\hat{\sigma}^2_{j,k} + \mathbb{E}\Big(\frac{n_j - n_{jk'}}{n_{jk'}}\Big)\hat{\sigma}^2_{j,k'} + 2\ \hat{\gamma}^{S-}_{j,kk'}\Big].
\end{aligned} \tag{A.38}$$

Appendix B

Supplement

B.1 Results of SUPPORT Data

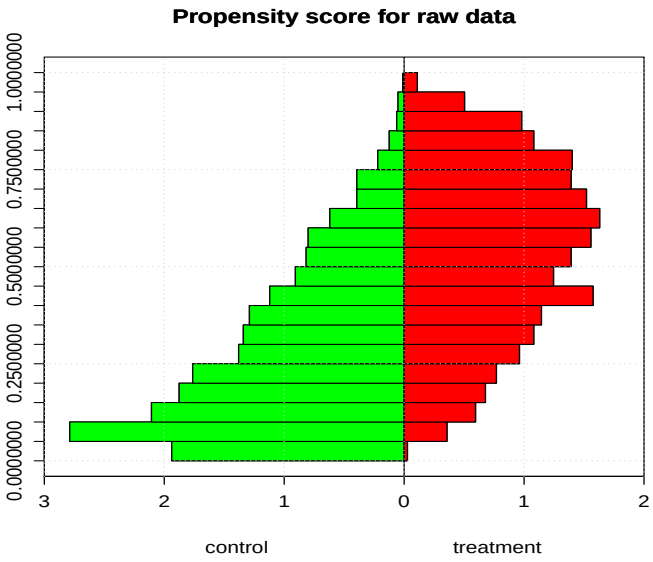


Figure B.1: Figure shows the histogram of the propensity score for the original SUPPORT data

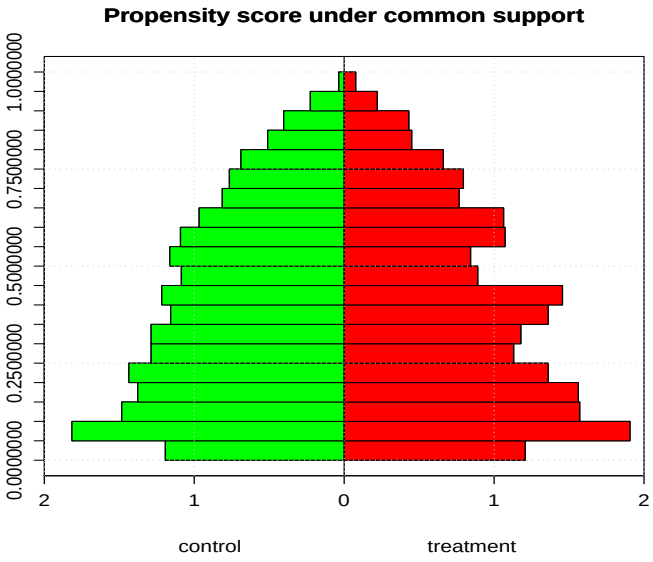


Figure B.2: Figure shows the histogram of the propensity score under largest connected components

B.2 Previous Results of SUPPORT Data

Crump et al. (2009) find the following result for the SUPPORT data. Table B.2 show that there are 870 control subjects that have low propensity score compare to 40 treated subjects. Again, there are 10 control subjects that have high propensity score compare to 87 treated subjects. Clearly, the data present a lack of common support in terms of propensity score.

Table B.1: Covariate imbalance in SUPPORT data

	No RHC	RHC	SMD
n	3551	2184	
age (mean (sd))	61.76 (17.29)	60.75 (15.63)	0.061
sex = Male (%)	1914 (53.9)	1278 (58.5)	0.093
race (%)			0.036
black	585 (16.5)	335 (15.3)	
other	213 (6.0)	142 (6.5)	
white	2753 (77.5)	1707 (78.2)	
edu (mean (sd))	11.57 (3.13)	11.86 (3.16)	0.091
income (%)			**0.142**
> $50k	257 (7.2)	194 (8.9)	
$11-$25k	713 (20.1)	452 (20.7)	
$25-$50k	500 (14.1)	393 (18.0)	
Under $11k	2081 (58.6)	1145 (52.4)	
ninsclas (%)			**0.194**
Medicaid	454 (12.8)	193 (8.8)	
Medicare	947 (26.7)	511 (23.4)	
Medicare & Medicaid	251 (7.1)	123 (5.6)	
No insurance	186 (5.2)	136 (6.2)	
Private	967 (27.2)	731 (33.5)	
Private & Medicare	746 (21.0)	490 (22.4)	
cat1 (%)			**0.583**
ARF	1581 (44.5)	909 (41.6)	
CHF	247 (7.0)	209 (9.6)	
Cirrhosis	175 (4.9)	49 (2.2)	
Colon Cancer	6 (0.2)	1 (0.0)	
Coma	341 (9.6)	95 (4.3)	
COPD	399 (11.2)	58 (2.7)	
Lung Cancer	34 (1.0)	5 (0.2)	
MOSF w/Malignancy	241 (6.8)	158 (7.2)	
MOSF w/Sepsis	527 (14.8)	700 (32.1)	
das2d3pc (mean (sd))	20.37 (5.48)	20.70 (5.03)	0.063
dnr1 = Yes (%)	499 (14.1)	155 (7.1)	**0.228**
ca (%)			**0.107**
Metastatic	261 (7.4)	123 (5.6)	
No	2652 (74.7)	1727 (79.1)	
Yes	638 (18.0)	334 (15.3)	
surv2md1 (mean (sd))	0.61 (0.19)	0.57 (0.20)	**0.198**
aps1 (mean (sd))	50.93 (18.81)	60.74 (20.27)	**0.501**
scoma1 (mean (sd))	22.25 (31.37)	18.97 (28.26)	**0.110**
wtkilo1 (mean (sd))	65.04 (29.50)	72.36 (27.73)	**0.256**
temp1 (mean (sd))	37.63 (1.74)	37.59 (1.83)	0.021
meanbp1 (mean (sd))	84.87 (38.87)	68.20 (34.24)	**0.455**

resp1 (mean (sd))	28.98 (13.95)	26.65 (14.17)	**0.165**
hrt1 (mean (sd))	112.87 (40.94)	118.93 (41.47)	**0.147**
pafi1 (mean (sd))	240.63 (116.66)	192.43 (105.54)	**0.433**
paco21 (mean (sd))	39.95 (14.24)	36.79 (10.97)	**0.249**
ph1 (mean (sd))	7.39 (0.11)	7.38 (0.11)	**0.120**
wblc1 (mean (sd))	15.26 (11.41)	16.27 (12.55)	0.084
hema1 (mean (sd))	32.70 (8.79)	30.51 (7.42)	**0.269**
sod1 (mean (sd))	137.04 (7.68)	136.33 (7.60)	0.092
pot1 (mean (sd))	4.08 (1.04)	4.05 (1.01)	0.027
crea1 (mean (sd))	1.92 (2.03)	2.47 (2.05)	**0.270**
bili1 (mean (sd))	2.00 (4.43)	2.71 (5.33)	**0.145**
alb1 (mean (sd))	3.16 (0.67)	2.98 (0.93)	**0.230**
resp = Yes (%)	1481 (41.7)	632 (28.9)	**0.270**
card = Yes (%)	1007 (28.4)	924 (42.3)	**0.295**
neuro = Yes (%)	575 (16.2)	118 (5.4)	**0.353**
gastr = Yes (%)	522 (14.7)	420 (19.2)	**0.121**
renal = Yes (%)	147 (4.1)	148 (6.8)	**0.116**
meta = Yes (%)	172 (4.8)	93 (4.3)	0.028
hema = Yes (%)	239 (6.7)	115 (5.3)	0.062
seps = Yes (%)	515 (14.5)	516 (23.6)	**0.234**
trauma = Yes (%)	18 (0.5)	34 (1.6)	**0.104**
ortho = Yes (%)	3 (0.1)	4 (0.2)	0.027
cardiohx (mean (sd))	0.16 (0.37)	0.20 (0.40)	**0.116**
chfhx (mean (sd))	0.17 (0.37)	0.19 (0.40)	0.069
dementhx (mean (sd))	0.12 (0.32)	0.07 (0.25)	**0.163**
psychhx (mean (sd))	0.08 (0.27)	0.05 (0.21)	**0.143**
chrpulhx (mean (sd))	0.22 (0.41)	0.14 (0.35)	**0.192**
renalhx (mean (sd))	0.04 (0.20)	0.05 (0.21)	0.032
liverhx (mean (sd))	0.07 (0.26)	0.06 (0.24)	0.049
gibledhx (mean (sd))	0.04 (0.19)	0.02 (0.16)	0.070
malighx (mean (sd))	0.25 (0.43)	0.20 (0.40)	**0.101**
immunhx (mean (sd))	0.26 (0.44)	0.29 (0.45)	0.080
transhx (mean (sd))	0.09 (0.29)	0.15 (0.36)	**0.170**
amihx (mean (sd))	0.03 (0.17)	0.04 (0.20)	0.074

Table B.2: Results from previous study. Table presents the number of subjects that have high and low propensity score for treated and control untis

	$\hat{e}(x) < 0.1$	$0.1 < \hat{e}(x) < 0.9$	$0.9 < \hat{e}(x)$	Total
Controls	870	2671	10	3551
Treated	40	2057	87	2184
All	910	4728	97	5735