Comparing Causal Inference Estimators for Average Treatment Effect of Treated Units in Observational Studies

Kip Brown

Southern Illinois University Edwardsville

March 14, 2018

1 Introduction

There is a variety of methods used for estimating the average treatment effect in observational studies. To obtain unbiased estimates of the average treatment effect, there must be a balance in the distribution of covariates between the treated and control groups. Difficulties arise when a lack of covariate balance occurs between these groups, and when the covariates have a confounding effect on the response. Covariate balance can be obtained using propensity score methods and reweighting schemes. The propensity score is defined by Rubin and Rosenbaum as the probability of treatment assignment conditional on observed baseline characteristics [6]. The

propensity score is used as a balancing score, in order to determine the treatment effect in some experiment. This paper will discuss important theorems behind the propensity score, as well as several methods for determining the average treatment effect by using the propensity score as well as re-weighting schemes. This will be done by analyzing the different methods in Monte Carlo simulations and then a non-randomized study. The study was conducted by the German Breast Cancer Study Group in an attempt to estimate the effect of breast conservation methods versus mastectomies on the quality of life of the patient.

2 Set Up

When an experiment is designed to estimate the effects of a treatment, the researchers would ideally use randomized control trials (RCTs). In RCTs, treatment allocation is done randomly, which theoretically ensures the distribution of the covariates are the same in both the treatment and control groups. RCTs are not always feasible for both ethical and financial reasons, which results in the use of observational studies. In an observational study, the treated and control groups may have systematic differences, resulting in different distributions of the covariates between the two groups. This issue can lead to bias in the estimate for the average treatment effect. This paper will explore multiple methods of balancing the distribution of the covariates between the two groups to address this issue.

2.1 Propensity Score Framework

Suppose there is a random sample of size n from a population. For the i^{th} unit in the sample, let T_i denote which treatment was received, where $T_i = 0$ denotes the i^{th} unit receiving the control treatment, and $T_i = 1$ denotes the i^{th} unit receiving the treatment of interest. Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes of the control treatment and the treatment of interest, respectively. Let

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

denote the outcome observed for the i^{th} unit. The issue is that only one outcome or the other is observed, never both. These responses are used in the estimation of treatment effects. The goal of these studies is to evaluate different estimators for the average treatment effect in the population and the average treatment effect of the treated units. The average treatment effect (`ATE`) is defined as [1]

$$\tau = E[Y_i(1) - Y_i(0)].$$

The average treatment effect for the treated (`ATT`) is defined as,

$$\tau_t = E[Y_i(1) - Y_i(0) \mid T_i = 1], \tag{1}$$

which will be the focus of this paper. Each unit will also have a K-dimensional vector of pre-treatment covariates, denoted X_i. Throughout this paper, assume that

treatment assignment is strongly ignorable. This is made up of three assumptions [6][13].

Assumption 1. *(Unconfoundedness) For any unit* $i = 1, \ldots, n$*,*

$$P(T_i = 1 \mid Y_i(0), Y_i(1), X_i) = P(T_i = 1 \mid X_i) \tag{2}$$

or, using conditional independence notation

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i$$

Assumption 2. *(Probabilistic Assignment) For any unit* $i = 1, \ldots, n$*,*

$$0 < P(T_i = 1 \mid X_i) < 1$$

$$0 < \pi(X_i) < 1$$

Assumption 3. *(Individualistic) For any unit* $i = 1, \ldots, n$*, the probability of treatment assignment can be written as a common function of the ith's unit potential outcome and observed covariates.*

The first assumption above implies that the treatment assignment is conditionally independent of the outcome when conditioned on the covariates. The main idea is that when conditioning on the observed pre-treatment covariates, X_i, the responses $Y(0)$ and $Y(1)$ have no effect on the probability of receiving treatment.

The second assumption is that the probability of a unit receiving the treatment

of interest, conditional on the covariates, is between, but not including, 0 and 1. If the probability was equal to 1, the probability of receiving the control treatment would be 0. That means there would be no units receiving the control treatment at that level of X_i, and as a result, no comparisons could be made. Similar reasoning can be used as to why the probability above cannot equal 0. This assumption is essential when using the propensity score, and will be discussed futher later in the paper.

The third assumption implies that treatment assignment of a given unit depends only on a function of the observed covariates. This foundation leads us into discussing the propensity score.

3 The Propensity Score

As previously mentioned, the propensity score is a key part of many covariate balancing methods. In observational studies, the treated and control groups may have different distributions of covariates. The goal is to balance these distributions, so that the researcher can make unbiased estimates of the treatment effect, which can be done by conditioning on a balancing score.

Definition 3.1. (Balancing Score) A balancing score $b(x)$ is a function of the covariates such that

$$T_i \perp\!\!\!\perp X_i \mid b(Xi).$$

This can also be represented as a probability,

$$P(T_i = 1 \mid X_i, b(X_i)) = P(T_i = 1 \mid b(X_i)). \tag{3}$$

Balancing scores are not unique. For example, one of the most basic balancing scores is the function $b(X_i) = X_i$. Typically, this is not used in practice because X_i has the potential to have a very high dimensionality. Ideally, the balancing score will be low dimensional. This leads to a one dimensional balancing score, the propensity score. The propensity score is defined by Rubin and Rosenbaum. [6]

Definition 3.2. (Propensity Score) The Propensity Score is the conditional probability that a unit with observed covariates, X_i, will be in treatment group 1. The propensity core, $\pi(X_i)$, is then,

$$\pi(X_i) = P(T_i = 1 \mid X_i). \tag{4}$$

In RCT's the propensity score is known, but it is unknown in observational studies, so it must be estimated. This is usually done using a logistic regression model, which will be discussed later. The propensity score must be shown to be a balancing score.

Theorem 1 . *(Propensity Score is a balancing score) The propensity score* $\pi(X_i) = P(T = 1|X_i = x)$ *is a balancing score.*

Proof. It must be shown that the propensity score is a balancing score, which by

equation (3),

$$P(T_i = 1 \mid X_i, \pi(X_i)) = P(T_i = 1 \mid \pi(X_i)). \tag{5}$$

First, lets look at the left side (5). Since $\pi(X_i)$ is a function of X_i, the left side can be written as follows,

$$\begin{aligned} P(T_i = 1 \mid X_i, \pi(X_i)) &= P(T_i = 1 \mid X_i) \\ &= \pi(X_i). \end{aligned}$$

It now must be shown that the right side of (5) is also equal to $\pi(X_i)$.

$$\begin{aligned} P(T_i = 1 \mid \pi(X_i)) &= 1 \cdot P(T_i = 1 \mid \pi(X_i)) + 0 \\ &= 1 \cdot P(T_i = 1 \mid \pi(X_i)) + 0 \cdot P(T_i = 0 \mid \pi(X_i)) \\ &= E_T \left[T_i \mid \pi(X_i) \right] \\ &= E_X \Big[E_T \left[T_i \mid X_i, \pi(X_i) \right] \mid \pi(X_i) \Big] \\ &= E_X \Big[P(T_i = 1 \mid X_i, \pi(X_i)) \mid \pi(X_i) \Big] \\ &= E_X \Big[P(T_i = 1 \mid X_i) \mid \pi(X_i) \Big] \\ &= E_X \Big[\pi(X_i) \mid \pi(X_i) \Big] \\ &= \pi(X_i). \end{aligned}$$

By showing that both sides meet in the middle at $\pi(X_i)$, it has been shown that $P(T_i = 1 \mid X_i, \pi(X_i)) = P(T_i = 1 \mid \pi(X_i))$. Thus, $\pi(X_i)$ is a balancing score. $\square$

Now, it is of interest to show that if the unconfoundedness assumption holds for X, then it also holds for any balancing score.

Theorem 2 . *(Unconfoundedness given any balancing score)*
Suppose Assumption 1 is true. Then, treatment assignment is unconfounded given any balancing score,

$$P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = P(T_i = 1 \mid b(X_i))$$

or, using conditional independence notation

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid b(X_i).$$

Proof. Let Assumption 1 be true. It must be shown that

$$P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = P(T_i = 1 \mid b(X_i)). \tag{6}$$

Beginning with the left side of (6),

$$\begin{aligned} P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) &= 1 \cdot P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) + \\ &\qquad 0 \cdot P(T_i = 0 \mid Y_i(0), Y_i(1), b(X_i)) \\ &= E_T \left[T_i \mid Y_i(0), Y_i(1), b(X_i) \right] \\ &= E_X \Big[E_T \left[T_i \mid Y_i(0), Y_i(1), X_i, b(X_i) \right] \mid Y_i(0), Y_i(1), b(X_i) \Big] \end{aligned}$$

Now lets look at the internal expectation. By conditioning on X_i and letting As-

sumption 1 be true,

$$E_T\left[T_i \mid Y_i(0), Y_i(1), X_i, b(X_i)\right] = E_T\left[T_i \mid X_i, b(X_i)\right].$$

Then, by (3),

$$E_T\left[T_i \mid Y_i(0), Y_i(1), X_i, b(X_i)\right] = E_T\left[T_i \mid b(X_i)\right]$$

Thus,

$$E_X\Big[E_T\left[T_i \mid Y_i(0), Y_i(1), X_i, b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big] = E_X\Big[E_T\left[T_i \mid b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big]$$

But, since the expectation is with respect to X, and by Assumption 1,

$$\begin{aligned} E_X\Big[E_T\left[T_i \mid b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big] &= E_X\Big[E_T\left[T_i \mid b(X_i)\right] \mid b(X_i)\Big] \\ &= E_T\left[T_i \mid b(X_i)\right] \\ &= 1 \cdot P(T_i = 1 \mid b(X_i)) + 0 \cdot P(T_i = 0 \mid b(X_i)) \\ &= P(T_i = 1 \mid b(X_i)). \end{aligned}$$

Thus, $P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = P(T_i = 1 \mid b(X_i))$. So if Assumption 1 holds for X, it also holds for any balancing score. □

The implication of this result is that all bias that was associated with confounding

due to the covariates has been removed.

4 Logistic Regression and Estimating the Propensity Score

In observational studies, the propensity score is unknown, and thus, must be estimated. Logistic regression is one way to do this, and will be the method used in this paper. By definition, the propensity score is

$$\pi(X_i) = P(T_i = 1 \mid X_i) = E[T_i \mid X_i],$$

since T_i is a Bernoulli random variable. The binary logistic regression response function is [11]

$$\pi(X_i) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}, \tag{7}$$

where X_i is vector of pre-treatment observed covariates for the i^{th} unit, and β is the vector of parameters. To find the fitted response function, estimates of β must be found. This is done with most likely estimators (MLE's). Since T_i is a Bernoulli random variable, the likelihood function can then be written as

$$L(\beta) = \prod_{i=1}^{n} \pi(X_i)^{T_i} \cdot (1 - \pi(X_i))^{1-T_i}$$

Then the log-likelihood function for the model is

$$\begin{aligned} l(\beta) = \ln(L(\beta)) &= \ln\Big(\prod_{i=1}^{n} \pi(X_i)^{T_i} \cdot (1-\pi(X_i))^{1-T_i}\Big) \\ &= \sum_{i=1}^{n} \ln\Big(\pi(X_i)^{T_i} \cdot (1-\pi(X_i))^{1-T_i}\Big) \\ &= \sum_{i=1}^{n} \Big(T_i \cdot \ln(\pi(X_i)) + (1-T_i) \cdot \ln(1-\pi(X_i))\Big). \end{aligned} \tag{8}$$

Equation (8) will be referred to again in the covariate balancing propensity score method, which is discussed later in this paper. To derive the MLE's for β, the partial derivative of the log-likelihood function with respect to β is taken. Note that $\pi(X_i)$ is a function of β.

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \Big(\frac{T_i}{\pi(X_i)} \cdot \pi'(X_i) + \frac{1-T_i}{1-\pi(X_i)} \cdot -\pi'(X_i)\Big) \tag{9}$$

This derivative is set equal to 0 and solved for $b_0, b_1, ..., b_p$, which maximize our log-likelihood function. This can be verified with a second derivative test. Then the fitted logistic response function is

$$\hat{\pi}(X_i) = \frac{\exp(X_i'b)}{1+\exp(X_i'b)}$$

The fitted values are then the estimated propensity scores. The linearized model, also referred to as the logit response function, is

$$ln\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = X_i\beta$$

There are multiple methods in which the researcher can evaluate the performance of a model. The deviance, or likelihood ratio statistic, is a measure of adequacy of a given model, and will be used in fitting the propensity score model.

Definition 4.1. Let l_r and l_p be the log-likelihood functions for the reduced model and the proposed model respectively. Then the likelihood ratio statistic, D, is

$$D = 2[l_p - l_r]$$

This statistic will be used to when deciding whether or not to include a covariate into the propensity score model. Let l_b, l_p, and l_n be the log-likelihood functions for the current base model, the proposed model, and the null model, respectively. Let D_b and D_p be the deviance statistics for the base and proposed models respectively. R will return the deviance statistics with respect to the null model such that

$$D_b = 2[l_b - l_n],$$
$$D_p = 2[l_p - l_n].$$

The deviance between the base model and the proposed model is of interest, so

$$\begin{aligned} D_b - D_p &= 2[l_b - l_n] - 2[l_p - l_n], \\ &= 2l_b - 2l_n - 2l_p + 2l_n, \\ &= 2l_b - 2l_n, \\ &= 2[l_b - l_n]. \end{aligned}$$

This result is the likelihood ratio statistic between the current base model and the proposed model, and will be used in the next section.

4.1 Estimating the Propensity Score

Throughout this paper, the propensity score was modeled using the program R. One issue in modeling the propensity score is deciding which covariates should be included, as well as deciding which higher order and interaction terms should be included. This is done in a three step, iterative process. This process is discussed by Rubin and Imbens [13].

4.1.1 Step 1: Selecting Scientifically Significant Covariates

The first step is to include all of the covariates that are deemed to be scientifically significant factors. This is typically done in collaboration with an expert in the field of study. The content used throughout this paper varies in content matter; therefore, significant factors were estimated based on the authors judgment. At this stage, these covariates are included in a linear manner.

4.1.2 Step 2: Selecting Statistically Significant Covariates

At this stage, any additional observed covariates that were not deemed scientifically significant were considered. This step is done by an iterative process. A base model is fit including all of the covariates from step 1. A new model is fit for each

of the remaining variables, such that each of the predictors in the base model are included, plus 1 additional unused variable. The likelihood ratio statistic is calculated for each of the models with respect to the base model. If none of the test statistics are greater than 1, then none of the covariates will be added to the model. If one or more of the test statistics are greater than one, the covariate with the largest test statistic is added to the base model. This process is repeated until none of the test statistics are greater than 1, indicating that no more first order terms should be added to the model.

4.1.3 Step 3: Selecting Quadratic and Interaction Terms

At this stage, the researcher will add quadratic and interaction terms. Higher order terms are possible, but will not be included in the models in this paper. According to the hierarchical approach, only quadratic and interaction terms that include covariates that are already included in the model will be considered. Again, if there are interactions that are deemed to be scientifically significant, they will be included in the model. If not, an iterative approach similar to step 2 is taken. The base model is the final model from step 2. New models are fit with all of the terms thus far plus one additional quadratic or interaction term. The likelihood ratio statistic is calculated for each of additional models. There must be more evidence for a second order term to be included into the model, thus, likelihood ratio statistics will be compared to the cutoff value of 2.71 in this stage. If none of the test statistics are greater than 2.71, then none of the second order terms are added to the model. If one or more test statistics are greater than 2.71, the second order term with the largest test statistic

is added to the model. This process is repeated until none of the test statistics are greater than 2.71, which indicates that the model is complete.

4.2 Covariate Balancing Propensity Score

One of the main issues in many observational studies is that the researches do not know whether or not the propensity score has been modeled correctly. Even if the propensity score is adequately modeled, it is not ensured that the covariate distributions will be balanced as a result. To check the validity of a propensity score model, the researcher checks the resulting covariate balance. Researchers will then change their model and check the resulting covariate balance again, and repeat this process until they have obtained an acceptable covariate balance. The covariate balancing propensity score (CBPS) estimates the propensity score and optimizes covariate balance simultaneously. Traditionally, the parameters for the propensity score are estimated by the MLE method, which maximize the likelihood function. In the CBPS method, the parameters of the likelihood function are estimated by the method of moments, while satisfying a balancing condition. First, look at the likelihood function. The goal is to find estimates of β such that equation (8) is maximized,

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \left[\sum_{i=1}^{n} \Big(T_i \cdot \log(\pi(X_i)) + (1 - T_i) \cdot \log(1 - \pi(X_i)) \Big) \right]$$

Now, recall the partial derivative found in equation (9), which will be denoted here as $s_\beta(T_i, X_i)$,

$$\frac{1}{n} s_\beta(T_i, X_i) = \frac{1}{n} \frac{\partial l}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{T_i}{\pi(X_i)} \cdot \pi'(X_i) + \frac{1 - T_i}{1 - \pi(X_i)} \cdot -\pi'(X_i) \Big).$$

Then, for the `ATT` method, the parameters are estimated such that the following balancing condition is met,

$$\frac{1}{n_1} \sum_{i=1}^{n} \Big(T_i - \frac{(1 - T_i)\pi(X_i)}{1 - \pi(X_i)} \Big) \tilde{X}_i = 0,$$

where $\tilde{X}_i$ is some function of X_i defined by the researcher [10]. In the CBPS method, this is set to be X_i in order to balance the first moment of each covariate, or $\tilde{X}_i = (X_i^T (X_i^2)^T)^T$ in order to balance the first two moments of each covariate. Theoretically, this will result in better covariate balance between the treatment and control groups. This is the condition that separates the CBPS method from traditional propensity score estimation. Even if the model is slightly mispecified, the CBPS method will still obtain a better covartiate balance. The proof of this is outside the scope of this paper.

5 Types of Covariate Balancing Methods

This section will look at 4 methods for estimating the `ATT`: matching, stratification, inverse probability of treatment weighting, and entropy balancing.

5.1 Matching Methods

Matching on the propensity score involves forming matched sets of treated and control units that have similar values of the propensity score. Within these matched sets, the average treatment effect can be estimated. There are multiple methods in finding matched sets, each made up of two parts. First, the researcher must determine which distance metric they would like to implement. There are several ways this can be done, but this paper will discuss mahalanobis distance and absolute propensity score difference.

Definition 5.1. (Mahalanobis Distance): The mahalanobis distance is defined as

$$D_{ij} = (\mathbf{X_i} - \mathbf{X_j})'\Sigma^{-1}(\mathbf{X_i} - \mathbf{X_j}),$$

where $\mathbf{X_i}, \mathbf{X_j}$ are $p \times 1$ vectors of covariates for the i^{th} and j^{th} units respectively, and Σ is the variance-covariance matrix of the covariates [13].

Definition 5.2. (Absolute Propensity Score Difference): This difference is defined as,

$$D_{ij} = |\pi(X_i) - \pi(X_j)|,$$

where $\pi(X_i), \pi(X_j)$ are the estimated propensity scores for the i^{th} and j^{th} units respectively [12].

Next, the researcher needs to decide how they would like to match the units

based off of the chosen distance metric. This can be done in many different ways. This paper will focus on 1 to 1 nearest-neighbor matching for both distance metrics. In 1 to 1 matching, a treated unit is selected. Then, based on which distance metric is being applied, the control unit with the smallest distance from the treated unit are then matched together. Often the researcher will set some caliper for the distance. For example, consider a scenario where the researcher set the caliper to be an absolute propensity score difference of 0.05. Now, suppose a treated unit has a propensity score of 0.85. Then, if there are no control units with an estimated propensity score between 0.80 and 0.90, then no match can be made, and the unit is discarded. There are a few more complex matching methods, but this paper will only be looking at 1 to 1 nearest neighbor matching with replacement, with and without calipers. To estimate the `ATT`, the treated units are matched with control units, which means some control units will be discarded. Once units are matched, the difference of the response is measured. Then a simple average of that difference is the `ATT`.

5.1.1 Efficiency Bounds

The extremes of the propensity score are difficult to analyze, and result in higher variance in the estimates for the `ATT`. A way to compensate for this issue is to implement upper and lower bounds for the propensity score, where any propensity score outside of these bounds are discarded. One way of doing this is to simply define some level α, typically 0.1. As a result, the estimator will only look at propensity

score values such that,

$$\alpha \leq \pi(X_i) \leq 1 - \alpha.$$

A better way of obtaining these efficiency bounds is discussed by Crump [2].

Theorem 3 . *Suppose that treatment assignment is strongly ignorable and the density of* $\mathbb{X}$ *is bounded away from zero and infinity. Suppose also that* $\sigma_w^2(x) = \sigma^2$ *for all* $w \in 0, 1$ *and* $x \in \mathbb{X}$*. Then the optimal subpopulation average treatment effect is* $\tau_{S,\mathbb{A}_H^*}$*, where*

$$\mathbb{A}_H^* = \{x \in \mathbb{X} \mid \alpha \leq \pi(X_i) \leq 1 - \alpha\}.$$

If

$$\sup_{x \in \mathbb{X}} \frac{1}{\pi(X_i)1 - \pi(X_i)} \leq 2E\left[\frac{1}{\pi(X_i)1 - \pi(X_i)}\right],$$

then $\alpha = 0$ *and* $\mathbb{A}_H^* = \mathbb{X}$*. Otherwise,* α *is a solution to*

$$\frac{1}{\alpha(1-\alpha)} = 2E\left[\frac{1}{\pi(X_i)1 - \pi(X_i)} \;\middle|\; \frac{1}{\pi(X_i)1 - \pi(X_i)} \leq \frac{1}{\alpha(1-\alpha)}\right].$$

Removing these units with extreme propensity score values is called trimming. The proof of this is outside the scope of this paper. Another way of obtaining these bounds is by looking at the maximum and minimum estimated propensity scores between the two groups. All treated units with estimated propensity scores less than

the minimum estimated propensity score of the control units are discarded. Similarly, all control units with estimated propensity scores greater than the maximum estimated propensity score of the treated units are discarded. The second and third trimming methods will be applied.

5.2 Stratification on the Propensity Score

Stratification on the propensity score involves splitting the units into mutually exclusive subsets based on their propensity score. Units are first ordered by their propensity scores. By the probabilistic property of the propensity score, these values range from 0 to 1. Now suppose that the range of propensity score values is split into J stratas, or subclasses. As discussed by Imbens and Rubin [13], the intervals are defined such that

$$\cup_{i=1}^{J} = [0, 1),$$

where $b_0 = 0$ and $b_J = 1$. Units within each strata now have very similar propensity scores, and are treated like they have roughly equal propensity scores. Within these stratas, units will now theoretically have a better covariate balance between the treated and control groups. Now, let $n_c(j)$ and $n_t(j)$ be the number of control units and treated units respectively in the j^{th} stratum, and let $q(j)$ be the fraction of units in the j^{th} stratum. Now, within each strata, the strata specific average treatment,

$\tau_{dif}(j)$, can be estimated. This is defined to be

$$\tau_{diff}(j) = \bar{Y}_t(j) - \bar{Y}_c(j),$$

where

$$\bar{Y}_t(j) = \frac{1}{n_t(j)} \sum_{i=1}^{n} T_i \cdot B_i(j) \cdot Y_i \text{ and } \bar{Y}_c(j) = \frac{1}{n_c(j)} \sum_{i=1}^{n} (1 - T_i) \cdot B_i(j) \cdot Y_i.$$

$B_i(j)$ is an indicator variable where 1 indicates the unit is in the j^{th} stratum, and a 0 if not. Once each strata specific treatment effect is estimated, then the overall average treatment effect can be estimated.

$$\tau_{strat} = \sum_{j=1}^{J} q(j) \cdot \tau_{diff}(j)$$

Now, there are multiple ways that the units can be stratified, a common one being splitting the units into quintiles. These quintiles can be split in a way that would result in each stratum containing the same range of propensity score values. Another way to split the quintiles is so the stratum will have roughly equal proportions in each quintile. The units within each quintile will have similar propensity score values, allowing one to make direct comparisons to determine the average treatment effect. It may also be of interest to find the average treatment effect in subpopulations of the data. This can be done when using stratification.

5.3 Inverse Probability of Treatment Weighting Using the Propensity Score

Re-weighting is a process that re-assigns the weight based on the propensity score for each unit. In this re-weighting scheme, the `ATT` can be estimated with the following estimator,

$$\tau_{W,ATT} = \frac{1}{n}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}, \tag{10}$$

where $\pi(X_i)$ is the estimated propensity score, as discussed by Austin [1]. $\tau_{W,ATT}$ is an unbiased estimator for the `ATT`, which is

$$E[\tau_{W,ATT}] = E[Y_i(1) - Y_i(0)|T_i = 1].$$

Proof. Recall, the unconfoundedness assumption says $P(T_i = 1 \mid Y_i(0), Y_i(1), X_i) = P(T_i = 1 \mid X_i)$. Then

$$\begin{aligned} E[\hat{\tau}_{W,ATT}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n} T_i Y_i\right] - E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right]. \end{aligned}$$

First, focus on the left expectation.

$$E\left[\frac{1}{n}\sum_{i=1}^{n} T_i Y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[T_i Y_i\right]$$

.

Now since T is a binary treatment, T can only take on the values 0 or 1. This term only has a value when $T = 1$, so Y will only take on the value $Y(1)$. Now,

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} T_i Y_i\right] &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i(1)|T=1] \\ &= E[Y(1)|T=1]. \end{aligned}$$

Now lets look at the other expectation. By similar logic as above, this term is 0 when $T_i = 0$, so treat Y_i as $Y_i(0)$.

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{(1-T_i)Y_i(0)\pi(X_i)}{1-\pi(X_i)}\right] \\ &= E_X\left[E\left[\frac{(1-T_i)Y_i(0)\pi(X_i)}{1-\pi(X_i)} \mid X_i\right]\right] \\ &= E_X\left[\frac{\pi(X_i)}{1-\pi(X_i)} E\left[(1-T_i)Y_i(0) \mid X_i\right]\right]. \end{aligned}$$

By the unconfoundedness assumption, $1-T_i$ and Y_i can be split into two expectations.

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] = E_X\left[\frac{\pi(X_i)}{1-\pi(X_i)} \cdot E\left[(1-T_i) \mid X_i\right] \cdot E\left[Y_i(0) \mid X_i\right]\right]$$

But since T_i is discrete,

$$\begin{aligned} E[(1-T_i) \mid X_i] &= (1-0) \cdot P(T=0 \mid X_i) + (1-1) \cdot P(T=1 \mid X_i) \\ &= P(T=0 \mid X_i) \\ &= 1 - P(T=1 \mid X_i) \\ &= 1 - \pi(X_i). \end{aligned}$$

Then,

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] &= E_X\left[\frac{\pi(X_i)}{1-\pi(X_i)} \cdot (1-\pi(X_i)) \cdot E[Y_i(0) \mid X_i]\right] \\ &= E_X\left[\pi(X_i) \cdot E[Y_i(0) \mid X_i]\right] \\ &= E_X\left[P(T=1 \mid X_i) \cdot E[Y_i(0) \mid X_i]\right] \\ &= E_X\left[E[T_i \mid X_i] \cdot E[Y_i(0) \mid X_i]\right]. \end{aligned}$$

Now again, by the unconfoundedness assumption,inner expectations together can be pushed together. This results in

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] &= E_X\left[E[T_i \cdot Y_i(0) \mid X_i]\right] \\ &= E[T_i \cdot Y_i(0)] \\ &= E[Y(0) \mid T=1]. \end{aligned}$$

Finally,

$$\begin{aligned} E[\hat{\tau}_{W,ATT}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} T_i Y_i\right] - E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)}\right] \\ &= E[Y(1) \mid T=1] - E[Y(0) \mid T=1] \\ &= E[Y(1) - Y(0) \mid T=1]. \end{aligned}$$

Thus, $\tau_{W,ATT}$ is an unbiased estimator of the average treatment effect. □

It can also be thought of as a re-weight of each individual uint, such that,

$$w_i = T_i + \frac{(1-T_i)\pi(X_i)}{1-\pi(X_i)}.$$

Essentially, T_i and $1 - T_i$ are indicators. Then treated units are not re-weighted and control units are re-weighted as $\frac{\pi(X_i)}{1-\pi(X_i)}$. Once re-weighted, a simple difference in means can estimate the `ATT`.

5.4 Entropy Balancing

Entropy balancing is a re-weighting scheme with a slightly different approach than the other methods discussed. If the researcher is using traditional propensity score methods, then they will be applying an iterative process. The researcher will develop a model and test the resulting covariate balance, then make some adjustment to the model in hopes of improving the covariate balance. This process is repeated until an acceptable balance is obtained. Entropy balancing takes a different approach

to obtaining balanced covariate distributions. Similar to the idea of the CBPS, the researcher can designate 1 or more balancing constraints on the sample moments, typically the first two moments. These constraints will balance the the covariate distributions between the treated and control groups in both mean and variance. To derive these new weights, the loss function and balancing constraints need to be defined. The loss function is defined to be

$$h(w_i) = w_i \cdot \ln(\frac{w_i}{q_i}),$$

where w_i is the new estimated weight and q_i is the base weight, typically $\frac{1}{n}$. The first constraint is the balancing constraint. These constraints are applied in order to balance the covariate distribution between the treated and control groups. The balancing constraint is defined to be

$$\sum_{i|T=0} w_i \cdot c_{ri}(X_i) = m_r \quad \text{with} \quad r \in 1, ..., R.$$

R is the set of balance constraints identified by the researcher, $c_{ri}(X_i)$ is the function that describes the balancing constraints defined, and m_r is the r^{th} order moment of the variable X [9]. Typically, $c_{ri}(X_{ij}) = X_{ij}^r$ or $c_{ri}(X_{ij}) = (X_{ij} - \mu_j)^r$. The literature is unclear on specifically how many moment conditions to balance, but this paper focuses on the first two moments, $r = \{1, 2\}$. The other constraints are the normalizing constraints. These constraints ensure that our new weights sum to

one, and are strictly positive. They are defined as follows:

$$\sum_{i|T=0} w_i = 1 \quad \text{and} \quad w_i \geq 0 \quad \text{for all } i \text{ such that } T = 0.$$

The goal now is to minimize the loss function with the constraints mentioned above. This can be done by using the method of lagrange multipliers.

$$L = \sum_{i|T=0} w_i \cdot \ln(\frac{w_i}{q_i}) + \sum_{r=1}^{R} \lambda_r \Big(\sum_{i|T=0} w_i \cdot c_{ri}(X_i) - m_r \Big) + (\lambda_0 - 1) \Big(\sum_{i|T=0} w_i - 1 \Big).$$

Now, find the partial derivative with respect to w_i and set it equal to 0.

$$\begin{aligned}
\frac{\partial L}{\partial w_i} &= \ln(\frac{w_i}{q_i}) + 1 + \Big(\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) + (\lambda_0 - 1) \\
0 &= \ln(w_i) - \ln(q_i) + \Big(\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) + \lambda_0 \\
\ln(w_i) &= \ln(q_i) - \Big(\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) - \lambda_0 \\
w_i &= q_i \cdot \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) \cdot \exp(-\lambda_0)
\end{aligned}$$

The weights must sum to 1, so a new w_i^* is obtained such that,

$$\begin{aligned} w_i^* &= \frac{w_i}{\sum_{i|T=0} w_i} \\ &= \frac{q_i * \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i)\Big) * \exp(-\lambda_0)}{\sum_{i|T=0} q_i * \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i)\Big) * \exp(-\lambda_0)} \\ &= \frac{q_i * \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i)\Big)}{\sum_{i|T=0} q_i * \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i)\Big)}. \end{aligned}$$

Typically, equalizing the first two moments are of interest, so the weight simplifies to

$$w_i^* = \frac{q_i * \exp(-\lambda_1 c_{1i}(X_i) - \lambda_2 c_{2i}(X_i))}{\Sigma_{\{i|T=0\}} q_i * \exp(-\lambda_1 c_{1i}(X_i) - \lambda_2 c_{2i}(X_i))},$$

Re-weighting with $R = 2$ ensures that when each unit gets re-weighted, the sample mean and sample variance between the treated and control groups will be equal. This should result in a better estimate of the `ATT`.

6 Evaluating the Methods

This section explores the performance of each estimation method using the traditional propensity score and the CBPS. The methods are applied in two different settings, a simulation study and empirical data. This paper focuses on an

observational study conducted on breast cancer treatments.

6.1 Simulations

Four simulation designs were considered, each of which applied all of the methods discussed in this paper. Each design design used 3 pre-treatment variables for 1000 iterations.

Design	Description
A	50 treatment units, 100 control units, equal variance-covariance matrices
B	250 treatment units, 250 control units, equal variance-covariance matrices
C	50 treatment units, 100 control units, unequal variance-covariance matrices
D	250 treatment units, 250 control units, unequal variance-covariance matrices

For design A and B, the variance-covariance matrices are as follows.

$\Sigma_{T=1} = \Sigma_{T=0} =$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Design C and D will have different variances between the two groups, with variance-covariance matrix,

$\Sigma_{T=1} =$

$$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1.5 \end{bmatrix}$$

and

$\Sigma_{T=0} =$

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

For each design, the means will be,

$$\mu_{T=1} = [0, 0, 0]' \quad \mu_{T=0} = [0.4, 0.4, 0.4]'.$$

First, the raw estimate for the `ATT` was collected, each of the methods discussed in this paper was tested. A key for the method abbreviations is located in the appendix. The outcome variable was simulated as $Y_i = X_{1i} + X_{2i} + X_{3i} + \epsilon$, where ϵ is distributed $N(0, (0.5)^2)$.

6.2 Design A Results

	RAW	PSM	PSMC	PSTMC	PSTMO	MD	STRT1	STRT2
Bias	-121.26	-2.75	0.53	0.44	0.72	-18.26	-11.96	-12.66
MSE	155.96	1.76	1.53	1.66	1.61	4.84	3.05	3.18
	WPS	CBMat1	CBMat2	CBMat3	CBMat4	CBWPS	EB	
Bias	-0.53	-3.95	-1.20	-1.15	-1.36	0.04	-0.95	
MSE	3.08	1.84	1.69	1.82	1.77	0.24	0.26	

Table 1: Design A Results

Table 1 shows the results of the simulation from the first setting. The RAW estimate is by far the most biased and also has the highest mean squared error (MSE). The PSM method performs fairly well both in terms of bias and MSE. This method

is greatly improved when a caliper is set, which can be seen in the performance of the remaining three propensity score methods. PSMC, PSTMC, and PSTMO all perform fairly similarly. With the exception of CBWPS, these methods perform better than the remaining methods. Besides the raw estimate, the MD method performs the worst, both in terms of bias and MSE. Here, both of the stratification methods also perform very poorly. The WPS method is very accurate, but not quite as precise as some of the other methods. The next four methods shown are the same matching methods, except using the covariate balancing propensity score. These methods turn out to perform slightly worse than the normal propensity score. CBWPS, on the other hand, performs the best out of all of the methods in this design, both in terms of bias and MSE. Entropy balancing is slightly worse in terms of bias, but very good in MSE. In this setting, CBWPS is the best method.

6.3 Design B Results

	RAW	PSM	PSMC	PSTMC	PSTMO	MD	STRT1	STRT2
Bias	-120.17	-1.52	-0.10	-0.24	0.06	-12.09	-10.33	-12.27
MSE	146.74	0.26	0.19	0.19	0.18	1.70	1.38	1.84
	WPS	CBMat1	CBMat2	CBMat3	CBMat4	CBWPS	EB	
Bias	-0.02	-1.93	-0.53	-0.58	-0.43	0.01	-0.12	
MSE	1.25	0.27	0.19	0.20	0.20	0.07	0.07	

Table 2: Design B Results

In design B, each estimator performs significantly better, which was to be expected with the larger sample size. PSTMO and WPS both perform much better

than in design A. The MSE for each method is much smaller in each of the CB-Mat methods. Similar to design A, CBWPS outperforms the other methods by a considerable amount, both in terms of bias and MSE.

6.4 Design C Results

	RAW	PSM	PSMC	PSTMC	PSTMO	MD	STRT1	STRT2
Bias	-121.05	-23.04	0.24	-2.58	0.82	-43.61	1.36	-13.60
MSE	156.31	9.32	1.42	2.50	1.57	22.23	1.55	3.36
	WPS	CBMat1	CBMat2	CBMat3	CBMat4	CBWPS	EB	
Bias	-49.79	-25.07	-1.03	-2.28	-1.33	-0.09	-0.83	
MSE	28.36	10.64	1.91	2.65	2.22	0.36	0.37	

Table 3: Design C Results

There are some very interesting results in this section. PSM, MD, WPS, and CBMat1 all perform much worse in this setting, but there is great improvement in STRT1. Again, CBWPS outperforms the other methods.

6.5 Design D Results

	RAW	PSM	PSMC	PSTMC	PSTMO	MD	STRT1	STRT2
Bias	-119.56	-15.78	-0.56	-1.45	-0.18	-34.54	-9.54	-12.63
MSE	145.44	3.60	0.21	0.38	0.19	12.75	1.31	1.96
	WPS	CBMat1	CBMat2	CBMat3	CBMat4	CBWPS	EB	
Bias	-57.87	-15.62	-0.46	-0.77	0.50	-0.09	-0.20	
MSE	34.45	3.61	0.26	0.28	0.23	0.10	0.10	

Table 4: Design D Results

The result of this design is similar to the others. The propensity score matching methods that are using calipers perform well. The EB method performs very well in this setting as well. Similar to the other three settings, CBWPS outperforms the other methods.

6.6 Simulations Conclusion

The most apparent conclusion is that CBWPS is the best estimator in the experimental designs that were tested. This is the method that uses the inverse probability weighting method when using the covariate balancing propensity score. Entropy balancing and matching, with a caliper and trimming, also performed very well in each of the 4 settings. In designs A,B, and C, the matching methods that used the traditional propensity score outperformed the CBPS. There were very good results when using the 1-1 nearest neighbor matching with a caliper of 0.1 method, and applying the third trimming method in design B. Mahalanobis distance matching performed very poorly in each of the 4 designs. Stratification also performed fairly poorly in each setting. The only exception was in design setting C, where stratification with equal proportions in each strata performed relatively well. Now, these methods will be applied to an observational study.

7 German Breast Cancer Study Group Study

During the 1980's in Germany, mastectomies were the common choice of treatment for breast cancer. An alternative method was breast conservation surgery. This method removes only a portion of the breast, as opposed to the entire breast in a mastectomy. The German Breast Cancer Study Group conducted an observational study for the purpose of estimating the average treatment effect of the mastectomies versus breast conservation methods on patients emotional and physical status. Both emotional and physical status are obtained via a self reported survey. The data was obtained from the R package "nonrandom" and is obtained using the data("stu1") command. The data obtained consists of $n = 646$ patients, a subset of the original data, the head of which and the summary statistics are shown below. The data contains $n_0 = 479$ mastectomy patients and $n_1 = 167$ breast conservation patients. A description of the variables is located in the appendix.

Table 5: Head of the stu1 data

	klinik	tmass	therapie	alter	tgr	age	ewb	pst	mp
1	3	-7.76	0	-11.91	1	1	63.46	81.25	0
2	3	-4.76	0	-4.91	1	1	90.38	93.75	0
3	4	-3.76	0	-14.91	1	1	73.08	93.75	1
4	6	-7.76	0	-0.91	1	1	75.00	81.25	1
5	6	-3.76	0	-1.91	1	1	34.62	56.25	1

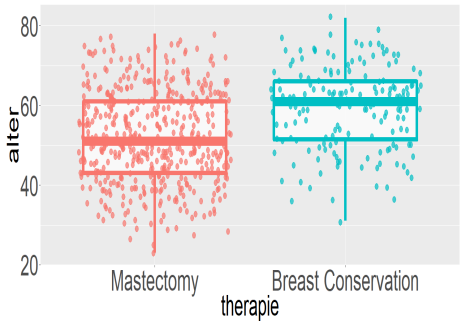
Table 6: Continuous Predictor Variables

Table 7: Sample Means

	tmass	alter
Breast Conservation	14.47	59.41
Mastectomy	13.51	52.00

Table 8: Sample Standard Deviations

	tmass	alter
Breast Conservation	4.40	11.50
Mastectomy	3.64	10.40



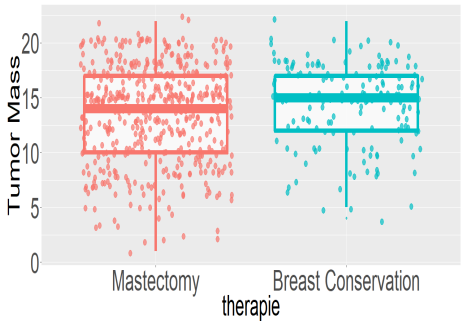


Table 9: Sample Conditional Proportions of Categorical Variables

	mp		tgr		age	
	< 15	≥ 15	≤ 10mm	> 10mm	≤ 55	> 55
Breast Conservation	0.5449	0.4551	0.1796	0.8204	0.3353	0.6647
Mastectomy	0.4405	0.5595	0.2714	0.7286	0.6138	0.3862

To obtain an unbiased estimate of treatment effects, the pre-treatment covariates need to be balanced between the treatment and control groups. The summary statistics above show that the variable "alter" is significantly different between the breast conservation and mastectomy groups. The categorical variable proportions differ between the two groups, most significantly between the variable "age". With these differences, the raw difference in mean responses could introduce a significant amount of bias.

7.1 Estimating the Propensity Score Model

Recall, to estimate the propensity score,

1. Include scientifically significant predictor variables,
2. Include statistically significant predictor variables,
3. Selecting quadratic and interaction terms.

Due to the small number of pre-treatment covariates, step 1 was skipped. Jumping directly to step 2, the iterative process found the following likelihood ratio statistics.

	Step			
$tmass$	6.53	4.28	5.11	-
$alter$	52.07	-	-	-
mp	5.42	6.44	-	-
tgr	5.87	4.14	4.39	0.25
age	38.96	0.05	0.03	0.01

Table 10: Likelihood Ratio Statistics

After three iterations, the variables alter, mp, and tmass are the three linear terms in the logistic regression model. By the hierarchical approach, only the three variables in the model will be considered in step 3.

	Step	
$alter^2$	0.86	0.76
$alter * many\_pat$	0.38	0.38
$alter * tmass$	0.19	0.17
$many\_pat * tmass$	0.06	0.22
$tmass^2$	3.98	-

Table 11: Likelihood Ratio Statistics

In step 3, $tmass^2$ was identified as a significant predictor to the model. The resulting estimated logit propensity score function is,

$$\ln\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = -0.756 + 0.059 \cdot alter - 0.495 \cdot many\_pat + 0.041 \cdot tmass - 0.010 \cdot tmass^2.$$

The CBPS is esimated with CBPS package with the CBPS function in R. The logit CPBS function is,

$$\ln\left(\frac{\pi_{CB}(X_i)}{1-\pi_{CB}(X_i)}\right) = -0.780 + 057 \cdot alter - 0.476 \cdot mp + 0.044 \cdot tmass - 0.009 \cdot tmass^2$$

7.2 ATT Estimates for pst and ewb

Estimates for the `ATT` of getting a mastectomy versus breast conservation methods on physical and emotional status were calculated for the raw estimates and the three best performing methods from the simulations section.

	RAW	PSTMO	CBWPS	EB
ATT	-1.59	-1.67	-1.15	1.73

Table 12: `ATT` Estimates for pst

	RAW	PSTMO	CBWPS	EB
ATT	0.09	0.44	0.19	0.33

Table 13: `ATT` Estimates for ewb

The CBWPS method was clearly the best estimator from the simulation section. Also shown in the simulation section, the raw estimate can be very biased. The raw

estimate for the `ATT` of mastectomies on physical status is an average increase of -1.59 units. The inverse probability weighting method using the covariate balancing propensity score estimates that the difference is less than that, being an a decrease of 1.15 units. The entropy balancing method finds an interesting result. Where the raw, matching, and IPW estimates found a decrease in physical status, the entropy balancing method finds an increase of 1.73 units. The raw estimate for the `ATT` of mastectomies on emotional status is nearly 0, implying that patients are, on average, reporting the same emotional status. The other three methods shown estimate a very slight increase in emotional status. In these two scenarios, the raw estimate is fairly close to the estimation methods discussed in this paper. The raw estimate is biased when there is a large imbalance between the predictor variables, and those predictor variables have a significant confounding effect. This could be an explanation for the similarities in the raw estimate and the estimators in this section.

8 Conclusion

The propensity score is the probability of receiving treatment conditional on the observed, pre-treatment covariates. This value is used as a balancing score, so the treated and control groups in observational or non randomized studies can be compared. These comparisons can be made in a variety of ways, including matching, stratification, weighting, and entropy balancing. Once these comparisons are made, the average treatment effect can be calculated. These estimation methods perform differently throughout different settings. The inverse probability weighting using the

covariate balancing propensity score performed the best in the limited designs that were explored. The researcher should be cautious before accepting a result obtained from a single estimation method.

8.1 Future Research

The purpose of this paper was to explore causal inference estimators in only a few experimental designs. Future research can be applied to exploring a wider variety of scenarios and a deeper look at the features of each of the methods, to gain a better understanding of which estimators should be used in which setting. R codes used for simulation and estimation can be found at https://github.com/kbrown1224/Thesis-Codes.

9 Appendix

	Description
klinik	categorical variable the clinic center
tmass	numeric variable indicating tumor size in mm
alter	numeric variable indicating age
therapie	binary treatment variable, 0 if the patient received a mastectomy, 1 if the patient received breast conservation
tgr	categorical variable indicating tumor size, 1 if the tumor size $\leq$ 10 mm, 0 if the tumor size $>$ 10 mm
age	categorical variable indicating age, 1 if the patient is $\leq$ 55 years old, 0 if the patient is $>$ 55 years old
ewb	numeric response variable indicating emotional status
pst	numeric response variable indicating physical status
mp	categorical variable indicating how many patients that units clinic had, 1 if $\geq$ 15 patients, 0 if $<$ 15 patients

Table 14: Description of stu1 Dataset

	Description
Raw	Raw difference in the average response value, Y
PSM	Propensity score matching, with 1-1 nearest neighbor matching without a caliper
PSMC	Propensity score matching, with 1-1 nearest neighbor matching with caliper = 0.1
PSTMC	Propensity score matching, with 1-1 nearest neighbor matching, with caliper = 0.1, and trimming according to Crump's lemma
PSTMO	Propensity score matching, with 1-1 nearest neighbor matching, caliper = 0.1, and trimming according to 3rd trimmning method discussed
MD	Mahalanobis distance matching
STRT1	Stratification, quintiles with roughly equal proportions in each strata
STRT2	Stratification, quintiles with equal ranges of propensity score
WPS	Inverse probability weighting
CBMat1	PSM using CBPS
CBMat2	PSMC using CBPS
CBMat3	PSTMC using CBPS
CBMat4	PSTMO using CBPS
CBWPS	WPS using CBPS
EB	Entropy balancing

Table 15: Description of Method Abbreviations

References

[1] Austin, P. (2011), "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research,* 46: 399 - 424.

[2] Crump et. al. (2009), "Dealing with limited overlap in estimation of average treatment effects" *Biometrika,* Vol. 96, No. 1, 187 - 199.

[3] Dawid, A. (1979), "Conditional Independence in Statistical Theory (with discussion)" *J.R. Statist Soc.,* B41, 1-31.

[4] R. Dehejia and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluationg the Evaluation of Training Programs," *Journal of the American Statistical Association,* Vol. 94, No. 448 1053 - 1062.

[5] K. Hirano and G. Imbens (2001), "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catherization," *Health Services & Outcomes Research Methodology,* 2: 259 - 278.

[6] P. Rosenbaum and D. Rubin. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* Vol. 70, No. 1, 41 - 55.

[7] P. Rosenbaum and D. Rubin. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score" *Journal of the American Statistical Association,* Vol. 79, No. 387 516 - 524.

[8] J. Hahn. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica,* Vol. 66, No. 2, 315 - 331.

[9] J. Hainmueller. (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Oxford Journals,* Vol. 20, No. 1, 25 - 46.

[10] K Imai and M Ratkovic. (2014) "Covariate balancing propensity score," *Journal of the Royal Statistical Society,* Vol. 76, Part 1, 243 - 263.

[11] M. Kutner, C. Nachtsheim, and J. Neter. (2004), "Applied Linear Regression Models", *McGraw-Hill/Irwin* 4th Ed., 570 - 572.

[12] E. Stuart. (2010), "Matching Methods for Causal Inference: A Review and A Look Forward", *Statistical Science,* Vol. 25, No. 1, 1 - 21.

[13] G. Imbens and D. Rubin. (2015) "Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction", *Cambridge University Press*

[14] S. Senn, E. Graf, and A. Caputo. (2007) "Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure", *Wiley InterScience* Vol. 26, 5529 - 5544.