ABSTRACT

Title of Dissertation: THE USE OF RANDOM FORESTS IN PROPENSITY SCORE WEIGHTING

Yating Zheng, Doctor of Philosophy, 2023

Dissertation directed by: Laura M. Stapleton, Professor
Department of Human Development and Quantitative Methodology
Quantitative Methodology: Measurement and Statistics Program

An important problem of social science research is the estimate of causal effects in observational studies. Propensity score methods, as effective ways to remove selection bias, have been widely used in estimating causal effects in observational studies. An important step of propensity score methods is to estimate the propensity score. Recently, a machine learning method, random forests, has been proposed as an alternative to the conventional method of logistic regression to estimate the propensity score as it requires less stringent assumptions and provides less biased and more reliable estimate of the treatment effect. However, previous studies only covered limited conditions with a small number of covariates and medium sample sizes, leaving the generalizability of the results in doubt. In addition, previous studies have seldom explored how to choose the hyper-parameters in random forests in the context of propensity score methods. This dissertation, via a simulation study, aims to 1) make a more comprehensive comparison between the use of random forests and logistic regression to determine which model performs better under what conditions, 2) explore the effects of the hyperparameters on the estimate of the treatment effect in propensity score weighting. An empirical study is also used as an illustration

about how to choose the hyperparameters in random forests using propensity score weighting in practical settings.

THE USE OF RANDOM FORESTS IN PROPENSITY SCORE WEIGHTING

by

Yating Zheng

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Laura M. Stapleton, Chair
Professor Jeffrey Harring
Professor Hong Jiao
Associate Professor Yan Li
Associate Professor Tracy Sweet

Acknowledgements

This study would not have been completed without the help of my advisor, Dr. Laura Stapleton. I would like to express my sincere gratitude to her for all the guidance and support throughout my doctoral studies. I would also like to thank my committee members, Dr. Harring, Dr. Sweet, Dr. Jiao and Dr. Li, for their valuable insights and suggestions. In addition, I owe my gratitude to the instructors who taught me wonderful Statistics and Measurement classes and instilled in me the spirit of never stop learning. Finally, I would like to thank my husband, parents, and all of my friends at University of Maryland College Park, who gave me priceless memories during my Ph.D. study.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

An important objective of much social science research is to identify the causal relation between a treatment or intervention and the subsequent outcomes. The gold standard for making causal inference is the randomized experimental design, in which all the subjects are randomly assigned to the treatment conditions (Shadish, 2002). Through the mechanism of randomization, the treatment and control groups are made equal in expectation in all the pre-treatment characteristics, both observed and unobserved. Thus, any observed difference in the outcomes between the treatment and control groups can be assumed to be due to the treatment condition.

However, it is not always feasible to use a randomized experimental design due to ethical or practical concerns. For example, researchers cannot force students to attend one school instead of another or randomly assign students to participate in some special educational programs. In such conditions, usually called observational studies, it is the subjects who choose their treatment conditions. This mechanism of treatment assignment may introduce selection bias, which means systematic differences in pre-treatment characteristics between participants in different treatment conditions. The selection bias may lead to biased estimates of treatment effects or causal effects. Inu other words, in observational studies, the treatment and control groups may differ in some pre-treatment characteristics, thus the post-treatment differences may not be necessarily due to the treatment effect. Thus, in observational studies, a very important task of making valid causal inferences is to control or adjust for selection bias (Rosenbaum, 1996).

Propensity score (PS) methods, as effective ways to remove selection bias, have been widely used in observational studies. An important step of propensity score methods is to estimate the propensity score, which is the probability of receiving a specific treatment condition given the observed covariates (Rosenbaum & Rubin, 1983). The conventional way to estimate

the propensity score is via logistic regression (Stuart, 2010; Lee et al., 2010; Thoemmes & Kim, 2011), which assumes that the functional form between the treatment status and the predictors is properly specified. A violation of the assumption may lead to biased estimates of the treatment effect. Recently, some non-parametric machine learning methods, such as classification tree, bagging, random forests and boosting, have been proposed as alternatives to logistic regression in estimating the PSs. Comparison studies showed that the use of random forests could provide less biased and more reliable estimates of the treatment effect than logistic regression (Arpino & Cannas, 2018; Lee et al., 2010). However, there are some limitations of these studies. First, in these studies, the researchers just used the default values of the hyperparameters in random forests in the existing software packages, where a hyperparameter is a parameter whose value is set before the model training process to help the model achieve good performance. There is a lack of understanding regarding the effects of the choice of the hyperparameters on the estimates of the propensity score and the treatment effect. Second, these studies only explored conditions with a small number of covariates and medium sample sizes, leaving the generalizability of the results in doubt. This dissertation 1) explores the effects of the choice of the hyperparameters in a well-performing machine learning method, random forests, on the estimates of the treatment effect in the context of propensity score weighting, and 2) makes a more comprehensive comparison between the use of random forests and logistic regression to determine which model performs better under what conditions. A simulation study is used to conduct the analysis. Results from the simulation study show that, in general, without hyper-parameter tuning, the performance of random forests is not comparable to that of logistic regression, which is indicated by the larger ASMD (average of the standardized mean difference on the covariates), the larger relative bias of the TE (treatment effect) estimates, and the greater variance of the TE estimates

by random forests. When appropriate hyper-parameters are used, the performance of random forests improves substantially in balancing covariates though it still is not as good as logistic regression. Despite the performance of random forests is not as good as logistic regression in balancing the covariates, with appropriate hyper-parameters, on average, random forests can provide more accurate TE estimates than logistic regression. In addition to the simulation study, an empirical study is also included as a tutorial about when to use which method and how to choose the hyperparameters in random forests in practical settings. Findings from this study adds to the literature on the comparison of random forests versus logistic regression in the context of propensity score methods and provides some guidance about how to choose the hyperparameters in applications of random forests to obtain a less biased and more reliable estimate of the treatment effect.

This dissertation proposal starts with an overview of propensity score methods, including the development, theoretical framework, assumptions and implementations, as well as an introduction of random forests and its application in the context of propensity score weighting. Chapter 3 describes the detailed research design of the simulation study. Chapter 4 discusses the results of the simulation study and Chapter 5 uses an empirical example to validate the results of the simulation study. Chapter 6 summarizes the findings and discusses the implications and limitations.

Chapter 2. Theoretical Framework and Literature Review

As mentioned earlier, the main objective for the application of PS methods is to identify the causal relation between the treatment and the outcome in observational studies. I start this chapter with an introduction of Rubin's model for causal inference. Then I provide an overview of the PS methods, including the development, theoretical framework, assumptions, and implementation. Next, I introduce random forests, a non-parametric machine learning method, its application in PS methods as well as the limitations. Finally, I conclude the chapter with two research questions that are explored in the study.

2.1 Rubin's Model for Causal Inference

To better understand causal effects, it is important to discuss Rubin's model for causal inference. Through Rubin's causal model (1974), the estimation of causal effects can be regarded as a comparison of potential outcomes. In particular, the causal effect δ for subject *i* can be defined as

$$\delta_i = Y_i(1) - Y_i(0) \,, \tag{1}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes if subject i receives the treatment or control condition respectively.

However, as stated by Holland (1986), the "fundamental problem of causal inference" is that for each subject only one of these potential outcomes can be observed. The potential outcome which cannot be observed is usually called *counterfactual*. Specifically, if subject i is assigned to the treatment group, then only $Y_i(1)$ can be observed and $Y_i(0)$ is the counterfactual. The same logic applies when subject i is assigned to the control group, then only $Y_i(0)$ can be observed and $Y_i(1)$ is the counterfactual. Therefore, it is impossible to estimate an individual-level causal effect.

One solution is to calculate the aggregated causal effect, the average causal effect between the treatment and control groups (Morgan & Winship, 2007). In randomized experimental designs, the randomization mechanism makes the treatment and control groups equal in expectation in all the pre-treatment characteristics, thus the causal effect can be unbiasedly estimated using the difference in the averages of the outcome variable between the two groups

$$\delta = E[Y(1)] - E[Y(0)]\,, \tag{2}$$

However, in observational studies, the treatment assignment is not randomized, which may bring in selection bias, the systematic differences between the treatment and control groups on pre-treatment characteristics. Because of the selection bias, the estimation of the causal effect using Equation 2 can be biased. To obtain unbiased estimates of causal effects, the bias due to selection needs to be removed. One solution that has been proposed to remove this selection bias is propensity score methods.

2.2 An Overview of Propensity Score Methods

The section below introduces the development of PS methods, the theoretical framework, the assumptions, and the implementation procedures.

2.2.1 Development of propensity score methods.

Various methods have been explored to remove selection bias and make valid causal inferences in observational studies, including matching (Cochran & Chambers, 1965; Rubin, 1973; Rosenbaum, 1989), regression (Hahn, 1998; Heckman et al., 1997), instrumental variables estimation (Hausman et al., 2012; Mullahy, 1997), regression discontinuity analysis (Lee & Lemieux, 2010; Lee & Munk, 2008), structural equation modeling (Grace, 2008; MacCallum & Austin, 2000) and propensity score (PS) methods (Rosenbaum & Rubin, 1983; Hirano & Imbens,

2003). Among these methods, matching and regression were used quite often before the surge of propensity score methods in the first decade of the 21st century. Matching works through pairing the treated and control subjects which have similar values on all the observed covariates that are related to the outcome. One problem with this method is that it is hard to match subjects on multiple covariates (Guo & Fraser, 2010). Another commonly used method is regression, which estimates causal effects through calculating the average group differences in outcomes conditional on relevant covariates. A concern of using regression is that it relies heavily on extrapolation, which may lead to biased estimates of causal effects (Arpino & Cannas, 2015; Drake, 1993). This problem is substantially serious when the treatment and control subjects differ systematically on the covariates, with the pre-treatment characteristics related to the outcomes.

Rosenbaum and Rubin (1983) introduced a promising alternative, which is actually a set of methods based on the PS, including PS matching, weighting and sub-classification. With the advantages of being able to easily construct matched sets on multiple covariates and avoid the problem of extrapolation in regression, PS methods became popular. Over the last two decades, PS methods have been widely used in many research fields to evaluate the effectiveness of various treatments and programs, in medicine (Wolfe & Michaud, 2004), education (Hong & Raudenbush, 2006), and in the policy realm (Arceneaux et al., 2010) for example.

2.2.2 Theoretical framework of propensity score methods.

A propensity score is defined as the probability of receiving the treatment given the observed covariates (Rosenbaum & Rubin, 1983). The key idea of PS methods is to make the treatment and control groups as comparable as possible in the pre-treatment covariates to facilitate an unbiased estimate of the treatment effect. Different PS methods use different

approaches to achieve this goal. Specifically, PS matching works through pairing the treatment and control subjects with similar propensity scores and discarding those without appropriate matches. Then treatment effects can be estimated using the mean difference in the outcome between the matched groups (Czajka et al., 1992; Rosenbaum & Rubin, 1985; Rubin, 1973). Subclassification, also known as stratification, works through stratifying all subjects into multiple subsets based on the estimated propensity scores, then estimating the treatment effect within each subset, and finally calculating the weighted average treatment effect across subsets as the overall treatment effect (Austin, 2011; Rosenbaum & Rubin, 1984). In PS weighting, the treatment and control groups are made equal in expectation on the pre-treatment covariates through changing the weights of the subjects. These approaches to using the propensity score to obtain unbiased estimates of the treatment effect hinge on important, necessary, assumptions as addressed next.

2.2.3 Assumptions of propensity score methods.

All of these methods operate under two key assumptions, specifically the strongly ignorable treatment assignment and the stable unit treatment value assumptions.

1. *Strongly ignorable treatment assignment assumption* (Rosenbaum & Rubin, 1983). This assumption involves two components, specifically, that 1) the treatment assignment (T), or more precisely the treatment selection, to be independent of the potential outcomes $(Y(0), Y(1))$ given the observed covariates (X), and 2) each subject has a positive probability of receiving each treatment:

$$\{Y(0), Y(1)\} \perp T | X \,, \tag{3}$$

$$0 < P(T = 1 | X) < 1 \quad \text{for all } X \,. \tag{4}$$

The first component implies that, conditional on observed covariates X, the treatment selection is random just as in randomized experimental designs. In other words, there is no unobserved covariate influencing the treatment selection. Thus, the treatment and control groups are equal in expectation in all the pre-treatment characteristics, both observed and unobserved. Sensitivity analysis can be used to examine the extent to which the inference about the causal effect is robust to potential unobserved confounders (Liu et al., 2013; Rosenbaum, 2002). More details about sensitivity analysis will be provided in section 2.2.4.2. The second component implies that each subject has a positive probability of receiving the treatment, meaning that the propensity score distributions of the treatment and control groups should have some overlap (also called common support).

2. *Stable unit treatment value assumption (SUTVA)* (Cox, 1958; Rubin, 1978). This assumption requires that the outcomes of a subject are not affected by the treatment selection of any other subjects. In practice, this assumption can be violated. For example, when evaluating the effects of a math intervention program a student who does not participate in the intervention program may have a higher math score if he/she gets some help from a friend who joins the intervention program. But if that friend had not been in the program, then he/she would not have scored as highly. However, the plausibility of this assumption can be improved by better research design such as reducing interactions between the treatment and control groups (Stuart, 2010).

If the assumptions are met, researchers can use PS methods. The section below addresses how to implement PS methods step by step.

2.2.4 Implementation of propensity score methods.

Before describing the implementation procedures, I would like to give a brief introduction about the types of treatment effects that can be estimated using PS methods as the estimation methods differ slightly for different types of treatment effects.

2.2.4.1 Types of treatment effects.

For most of the PS methods, two types of treatment effects can be estimated: the average treatment effect (ATE) and the average treatment effect on the treated (ATT). They are expressed as follows:

$$ATE = E[Y(1)] - E[Y(0)] , \quad (5)$$

$$ATT = E[Y(1)|T = 1] - E[Y(0)|T = 1] . \quad (6)$$

The ATE refers to the treatment effect on all the subjects, both treated and control (Imbens, 2004). In other words, it is the average difference in the outcomes if the whole population receives the treatment versus if the whole population receives the control. The ATT is the treatment effect for those in the treatment group only (Imai et al., 2007; Kurth et al., 2006). In other words, it is the average difference in the outcomes if all the treated subjects receive the treatment versus if none of the treated subjects receives the treatment.

The expected values of ATE and ATT will be the same only when the treatment effects are homogeneous, varying only randomly across subjects, otherwise not (Harder et al., 2010). Homogeneous treatment effects mean that the treatment effects are consistent for subjects with different characteristics. Using the ATE or the ATT is usually determined by research interest. Specifically, if the researcher is interested in the treatment effect on the whole population, ATE is usually used; otherwise, if the researcher is interested in the treatment effect on those who are likely to receive the treatment, ATT is the right choice. While PS matching can only estimate the

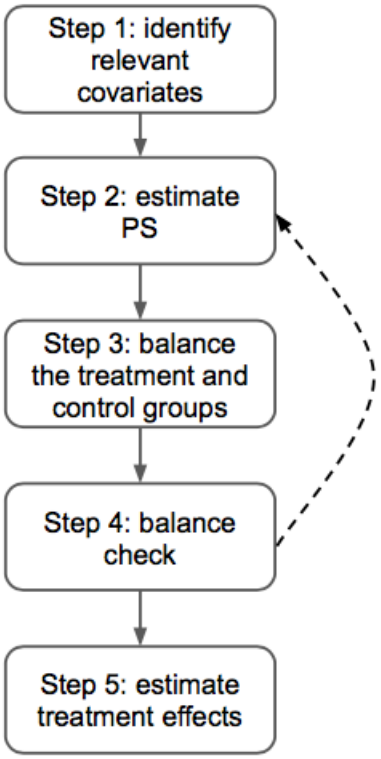
ATT, the use of PS weighting and sub-classification can estimate either the ATE or ATT, depending on interest.

2.2.4.2 Implementation procedures.

In general, all of the PS methods include five steps: 1) identify pre-treatment covariates that could potentially bias estimates of the treatment effect, 2) calculate propensity scores (using logistic regression or a non-parametric approach) on the identified covariates, 3) condition the propensity scores between the treatment and control groups, 4) check the conditioning quality (e.g., balance check), 5) estimate the treatment effects (Stuart, 2010). Figure 1 shows the flow of the implementation of PS methods.

Figure 1

The steps to implement propensity score methods



Step1. Covariate selection.

The selection of appropriate covariates is important because it influences all the steps afterwards and the accuracy of the estimates of the treatment effects. Previous studies suggest including all covariates that are related to the treatment selection and the outcome as PS methods heavily rely on the strongly ignorable treatment assignment assumption, which assumes that there are no unobserved differences between the treatment and control groups conditional on observed covariates (Rubin & Thomas, 1996; Stuart, 2010). If important covariates (especially covariates that are related to the outcome) are omitted, the estimated PS may be biased, which may influence the conditioning quality, the balance of the treatment and control groups on the pre-treatment covariates, and then introduce bias into the estimates of the treatment effects (Rosenbaum & Rubin, 1983; Shadish & Steiner, 2010; Steiner et al., 2011; Stuart, 2010). However, one type of covariate should not be included: covariates that can be affected by the treatment selection. Including this type of covariate may lead to substantial bias in the estimate of the treatment effect (Rosenbaum & Rubin, 1983; Frangakis & Rubin, 2002; Stuart, 2010).

Step 2. PS estimation.

After identifying the related covariates, the next step is to estimate the PSs. There are various methods to estimate the PSs. A conventional way is to utilize logistic regression. The log-odds of the propensity score is regressed on all selected pre-treatment covariates:

$$\log\left(\frac{e_i}{1-e_i}\right) = \beta_0 + \sum_{m=1}^{M} \beta_m X_{mi} \,, \tag{7}$$

where e_i is the propensity score of subject i, the probability of being in the treatment group for subject i; β_0 is the intercept of this model; β_m is the regression coefficient of the covariate X_m; X_{mi} is the value of subject i on covariate X_m.

Logistic regression is easy to implement and quite interpretable. However, to get

unbiased estimates of the PSs and the treatment effects, an important assumption needs to be met, the correctness of the functional form. In practice this assumption can be violated. Common misspecifications of the functional form include omitting important covariates, and specifying a linear relation between the covariates and the log odds of the treatment selection while the true relation is non-linear. Failing to meet the assumption of the correctness of the functional form may lead to biased estimates of the PSs, substantial difference in the pre-treatment covariates between the treatment and control groups, and subsequent biased estimates of treatment effects (Lee et al., 2010; Westreich et al., 2010). To enhance the performance of logistic regression, researchers often include all the variables relevant to the treatment and/or the outcome in the PS model and add polynomial terms or interactions between covariates (if a non-linear relation is hypothesized) (Rubin & Thomas, 1996). However, it is hard to identify the appropriate polynomial and interaction terms especially when there is a large number of covariates.

Recently, some non-parametric machine learning methods have been introduced to estimate the PSs, such as decision trees, bagging, random forests, and boosting (Austin, 2011; McCaffrey et al., 2004). These methods do not assume any functional form between the treatment assignment and the covariates, can select critical covariates automatically, and can capture non-linear and non-additive relations between the log odds of the propensity scores and the pre-treatment covariates (Cham, 2013; Lee et al., 2010). Several studies have demonstrated that these machine learning methods provide less biased estimates of the treatment effects especially when non-linearity and non-additivity exist (Austin, 2012; Lee et al., 2010; Setoguchi et al., 2008). Examining how these machine learning methods work is a key component of this study. I discuss the details of these methods in Section 2.3.

Step 3. PS conditioning.

Once the PSs are estimated, the next step is to use the estimated PSs to balance the treatment and control groups, making the treatment and control groups as comparable as possible on the pre-treatment covariates. Different PS methods use different strategies to achieve this goal.

Matching. PS matching achieves the goal of balancing the treatment and control groups on the pre-treatment covariates through discarding unmatched cases, usually unmatched cases in the control group. There are many different kinds of matching mechanisms depending on the matching algorithm, the restriction of distance, 1:1 or 1:k matching, and matching with or without replacement.

Subclassification. Subclassification divides subjects into strata based on the values of the estimated PSs. Subjects within each stratum in the treatment and control groups will have similar propensity scores and therefore similar distributions of covariates (Rosenbaum & Rubin, 1983). Within each stratum, a subclass treatment effect can be estimated by comparing the group mean differences in the outcomes. The overall treatment effect can be estimated using the weighted average of the treatment effects across strata. Depending on the research interest (to estimate the ATE or the ATT), there are two weighting methods:

$$ATE: w_j = \frac{n_j}{N}, \tag{8}$$

$$ATT: w_j = \frac{n_{jT}}{n_j}, \tag{9}$$

where w_j is the weight of the treatment effect for subclass j; n_j, n_{jT} and N represent the sample size of subclass j, the number of treated subjects in subclass j and the total sample size, respectively.

Cochran (1968) found that stratifying a sample into five subgroups based on one

covariate could eliminate 90% of bias due to that covariate. Rosenbaum and Rubin (1984) extended Cochran's work to PS subclassification and found similar results. It is a common practice to create five subclasses when using PS subclassification (Rosenbaum & Rubin, 1984; Yang et al., 2016; Zanutto et al., 2005).

Weighting. The method of PS weighting was first introduced by Rosenbaum and Rubin (1983). The key idea of PS weighting is to make the treatment and control group as comparable as possible on the pre-treatment covariates through changing the weights of subjects. There are two types of weighting methods: inverse probability of treatment weighting (IPTW) and weighting by the odds (WBO). IPTW is used to estimate the ATE while WBO is used to estimate the ATT:

$$w_{IPTW.i} = \frac{T_i}{e_i} + \frac{1-T_i}{1-e_i} \ , \tag{10}$$

$$w_{WBO.i} = T_i + (1 - T_i)\frac{e_i}{1-e_i} \ , \tag{11}$$

where $w_{IPTW.i}$ and $w_{WBO.i}$ represent the weight of subject *I* using IPTW and WBO respectively; T_i is the treatment assignment indicator with $T_i = 1$ representing that subject i is in the treatment group and $T_i = 0$ representing that subject *i* is in the control group; e_i is the estimated PS for subject i.

Using IPTW, subjects in the treatment group will be weighted by the inverse of the estimated PS while subjects in the control group will be weighted by the inverse of the probability of not receiving the treatment condition (Czajka et al., 1992; Imbens, 2004). With WBO, subjects in the treatment group receive a weight of 1 while those in the control group are weighted by the odds of the estimated PS (Harder et al., 2010; Hirano et al., 2003). Compared with other PS methods, PS weighting is much easier to implement (no need to match the subjects, nor divide subjects into strata, calculate the treatment effect within each stratum and the

aggregate treatment effect across strata) and is conducted on the entire sample without discarding any subjects, which is helpful to ensure the representativeness of the sample to the population of interest (Morgan & Todd, 2008; Rosenbaum, 1987). Yet PS weighting is not without limitation. When extreme PSs exist, it is likely to get extreme weights which may lead to biased TE estimates and excessive variance of the TE estimates (Hirano & Imbens, 2001; Li & Thomas, 2019).

Step 4. Balance check.

The goal of balance checking is to assess whether the treatment and control groups are comparable on the relevant covariates after matching, subclassification or weighting (Caliendo & Kopeining, 2008; Pan & Bai, 2005). A commonly used balance diagnostic is the absolute standardized mean difference (ASMD) between the treatment and control groups on each of the covariates:

$$ASMD_m = \frac{\bar{X}_{mT} - \bar{X}_{mC}}{\sigma_{mT}}, \quad (12)$$

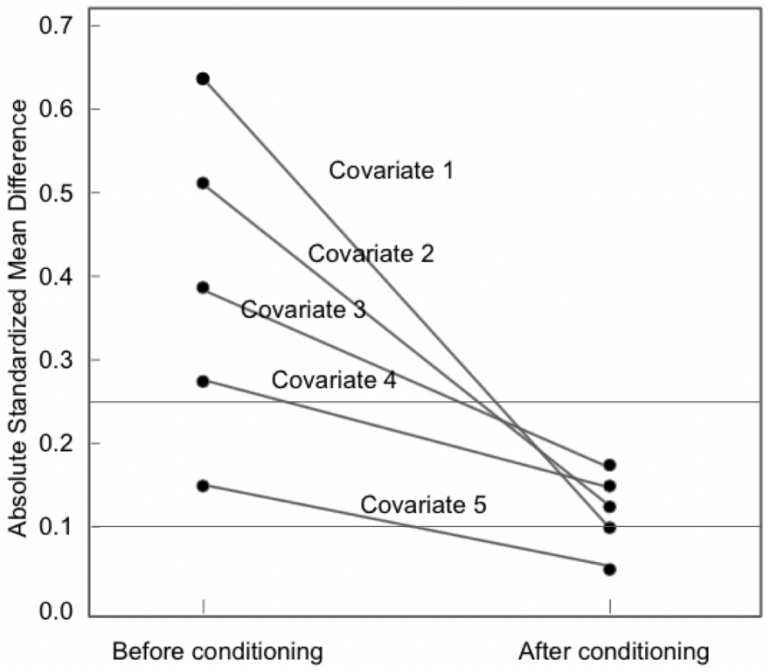
Where $ASMD_m$ is the absolute standardized mean difference on the covariate X_m; $\bar{X}_{mT}$ and $\bar{X}_{mC}$ represent the average values of treated and control subjects on the covariate X_m respectively; σ_{mT} is the standard deviation of the values of the treated subjects on the covariate X_m.

Covariates with ASMD smaller than 0.25 are usually considered balanced (Ho et al., 2007). Some researchers suggest using an even stricter cut-off value of 0.10 (Harder et al., 2010). For matching, the ASMD is directly calculated using Equation 12. For IPTW and WBO, the weighted group means and the standard deviation of the weighted covariate values of the treated subjects are used to calculate the ASMD. For subclassification, the ASMD within each stratum is calculated using Equation 12, then the weighted mean of the ASMDs across strata is calculated as the overall ASMD. Graphical plots like Figure 2 can provide a quick assessment of the

covariate balance. If a covariate is not balanced and it is strongly correlated with the outcome, it is problematic as it may yield substantial bias on the estimate of the treatment effect (Stuart, 2010). In this case, some other propensity score models should be tried (go back from step 4 to step2 as shown in Figure 1), usually by adding polynomial terms or interactions between covariates (Dehejia & Wahba, 1999; Smith & Todd, 2005). Once all of the critical covariates (those strongly correlated with the outcomes) are balanced, researchers can go to the next step to estimate the treatment effects.

Figure 2

Plot for balance check



Step 5. Treatment effect estimation.

As mentioned earlier, PS weighting and subclassification can estimate both the ATT and the ATE while PS matching can only estimate the ATT. For PS matching, the treatment effect

(ATT) can be directly calculated using the group mean differences in the outcomes. For PS subclassification, no matter the ATE or the ATT, the treatment effect is calculated using the weighted average treatment effect across strata:

$$\delta = \sum_{j=1}^{J} w_j \delta_j \ , \tag{13}$$

where δ is the overall treatment effect, δ_j is the treatment effect of stratum j, and w_j is the weight of stratum j. For the ATE, each subclass treatment effect is weighted by the proportion of the subclass size over the total sample size (Equation 8) while for the ATT each subclass treatment effect is weighted by the proportion of treated subjects in this subclass (Equation 9).

In PS weighting, the treatment effect can be calculated using the weighted group mean differences in the outcomes:

$$\delta = \frac{\sum_{i \epsilon T} w_i y_i}{\sum_{i \epsilon T} w_i} - \frac{\sum_{i \epsilon C} w_i y_i}{\sum_{i \epsilon C} w_i} \ , \tag{14}$$

where $i \ \epsilon \ T$ represents subject i in the treatment group, $i \ \epsilon \ C$ represent subject i in the control group, w_i is the weight of subject, and y_i is the outcome value of subject i. To estimate the ATE, each subject is weighted by the inverse of the probability of being in the chosen group (Equation 10). To estimate the ATT, treated subjects always receive a weight of 1 while control subjects are weighted by the odds of the PS (Equation 11).

In addition to these three commonly used methods, researchers have developed a "doubly-robust" method, which is a combination of conditioning and regression analysis (Funk et al., 2011; Stuart & Rubin, 2008). It regresses the outcomes on the covariates and the treatment condition on the matched sample or uses weighted regression of the outcome on the covariates. In this way, it can control for the remaining covariate differences after PS conditioning (if a lack of balance is found) and further reduce bias in the estimates of the treatment effects (Cochran & Rubin, 1973; Rubin, 1973).

The above five steps illustrate the general procedure of PS methods. However, as mentioned earlier in Section 2.2.3, for PS methods to provide an unbiased estimate of the treatment effect, the strongly ignorable treatment assignment assumption needs to be met. That's to say, there are no unobserved covariates influencing the treatment assignment conditional on the observed covariates. What if unobserved confounders do exist? Will that bias the estimate of the treatment effect? Sensitivity analysis is available to assess the robustness of the inference about the treatment effect if unobserved confounders exist. Various approaches can be used to conduct sensitivity analysis, including Rosenbaum's (1998), Greenland's (1996), Harding's (2003), Lin et al.'s (1998), and VanderWeele and Arah's (2011). Among these approaches, Rosenbaum's approach is used most often for matching methods like PS matching. The key idea of this approach is to find the thresholds of the associations between the unobserved confounder and the treatment assignment and/or the associations between the unobserved confounder and the outcome that would explain away the treatment effect (Gastwirth et al., 1998). For other PS methods other than matching, VanderWeele and Arah's approach is recommended because of its good performance and flexibility (Liu et al., 2013; Shen et al., 2011). The key idea of VanderWeele and Arah's approach is to approximate the true relation between the treatment assignment and the outcome adjusting for the hypothetical unobserved confounder (VanderWeele & Arah, 2011). Through the sensitivity analysis, researchers can have a better idea about the validity of the estimated treatment effect. However, very few applied researchers checked this after getting the estimated treatment effect.

In this section, I introduced the development and theoretical framework of PS methods, the important assumptions of using PS methods to estimate causal effects, and how to implement various PS methods step by step. In the next section, 2.3, I will introduce the theoretical

framework of random forests (an alternative method to estimate the PSs), its applications in the context of PS methods, as well as the limitations.

2.3 An Overview of Random forests

As discussed earlier, a very important step in the use of PS methods is to estimate the PSs. The conventional way to estimate the PSs is by the use of logistic regression. However, an assumption must be made regarding the correctness of the functional form to obtain unbiased estimates of the PSs (Stoltzfus, 2011). In reality, the relation between the log odds of the treatment assignment and the predictors may be non-linear and non-additive. Although logistic regression can address a non-linear and non-additive relation through the addition of polynomial terms and interactions between covariates, this approach to modeling requires iterative trials to find the appropriate polynomial terms and interactions, which is labor-intensive and time-consuming (Diamond & Sekhon, 2013). Moreover, applied researchers usually do not follow the iterative procedure (Austin, 2009). Recent studies have drawn attention to some machine learning methods, which can automatically select covariates, and capture nonadditivites and nonlinearities in the PS model. Multiple machine learning methods have been explored to estimate PSs, including decision trees, bagging, random forests, and generalized boosted models (Cham, 2013; McCaffrey et al., 2004; McCaffrey et al., 2013; Shadish & Steiner, 2010). Among all these methods, the method of random forests was comparatively parsimonious and was found to perform very well (Arpino & Cannas, 2018). This section gives an overview of the theoretical framework of random forests, and its application in PS methods, as well as its limitations in current applications.

2.3.1 Theoretical Framework of Random Forests.

Random forests, originated by Breiman (2001), is an ensemble machine learning method for classification. Since its introduction, it has been widely used in many fields and has shown to be very effective in many real-world classification problems (Orlandi et al., 2016). The key idea of this method is to build a strong classifier through combining multiple weak classifiers, which are single decision trees. Thus, to explain the theoretical framework of random forests, I will first introduce the mechanism of its basis, decision trees.

2.3.1.1 Decision trees.

Decision trees, also called classification and regression trees (CART), are non-parametric methods for constructing prediction models from data. The method of decision trees is the basis of all other tree-structured methods such as bagging, random forests, and boosting as I will discuss in Section 2.3.1.2 (Strobl et al., 2009). A decision tree uses a recursive algorithm to estimate a function which describes the relation between a set of independent variables and a dependent variable (McCaffrey et al., 2004). Based on the type of the dependent variable, continuous or categorical, there are two kinds of decision trees, regression trees for continuous outcomes and classification trees for categorical outcomes. In the PS model, the outcome is the treatment assignment, which is a binary variable, thus a classification tree is used.

The process of a classification tree starts with the whole sample (usually called root node) and splits the sample into two subgroups (also called child nodes) by a covariate and a corresponding cut-off value. The split can occur based on any covariate with any cut-off value. Among all the possible choices of the covariate and the cut-off value, the algorithm selects the one that can reduce the disorder of the data as much as possible. A common measure of disorder is *entropy*, which is calculated as follows:

$$E(node) = -p_T \cdot \log_2 p_T - p_C \cdot \log_2 p_C \ , \tag{15}$$

where p_T is the proportion of the subjects of the treatment group in the node, and p_C is the proportion of the subjects of the control group in the node.

After each split, the entropy is supposed to be less than that before split. The loss in the entropy before and after a split is called *information gain*, which is calculated as:

$$IG(split) = E(p\_node) - \left(E\left(c_{node_1}\right) + E\left(c_{node_2}\right)\right) , \tag{16}$$

where $E(p\_node)$, $E(c\_node_1)$, and $E(c\_node_2)$ are the entropies of the parent node, child node 1, and child node 2 respectively.

At each split, the choice of the covariate and the corresponding cut-off value aims to maximize the information gain. To achieve this goal, the algorithm searches through all the covariates, and all the corresponding cut-off values, computes the information gains, then chooses the covariate and cut-off value which can create the maximum information gain. In a classification tree, one covariate can be used repeatedly at multiple splits with different cut-off values (Breiman, 2001; Moisen, 2008; Strobl et al., 2009).

After the split, the subgroups will be more homogeneous in the outcome. This process continues until every terminal node just contains one kind of subject or a pre-determined criterion is met. Usually, the criterion can be a pre-determined number of splits (usually called maximum depth of a tree), or a pre-determined node size (usually called minimum terminal node size). This early stopping strategy is called *pruning* and is used to avoid overfitting, which means that the model works perfectly for the current sample but cannot be generalized to the whole population.

After a tree model is built on the sample (also called training dataset), it can be used to make predictions on new datasets. Specifically, a subject from a new dataset goes through the

tree until reaching a terminal node. The prediction result of this subject, either class membership or probability of having a specific class membership, is determined by the condition of the terminal node which this subject belongs to. In conventional applications for which the goal is to predict the class membership, a subject gets the same class membership as the majority of the subjects in its terminal node (called majority vote). In the context of PS methods, the goal is not to predict the class membership but to estimate the PSs. The PS of a subject is estimated by the proportion of the treated subjects in its terminal node.

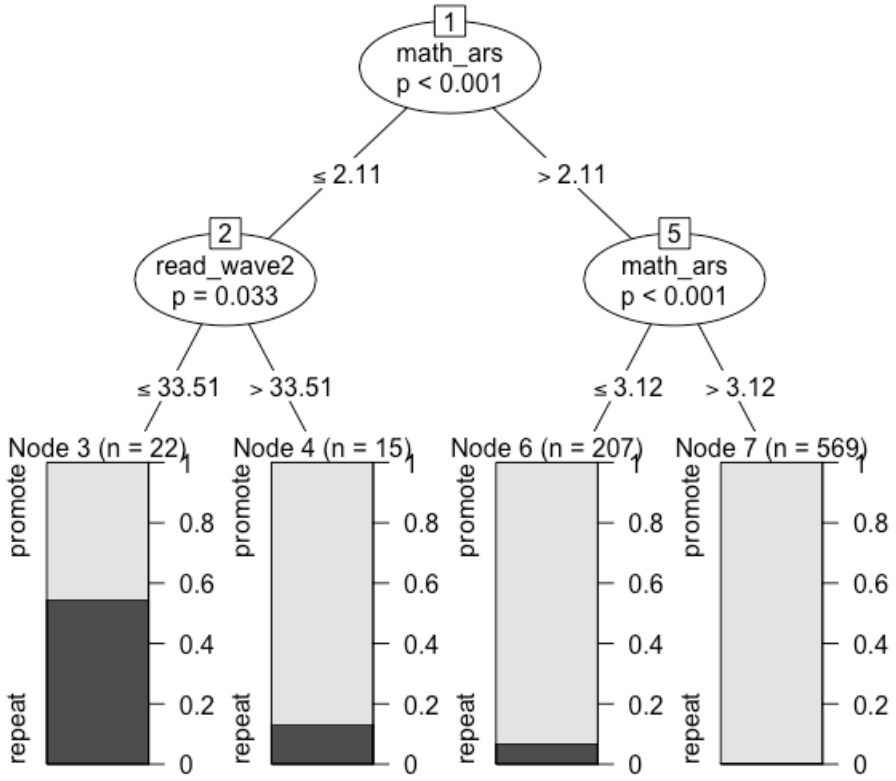
This algorithm is very straightforward and quite interpretable. Most importantly, it can automatically select the most influential covariates (using information gain), and can adaptively capture non-linear and non-additive effects (Lee et al., 2010). The example below demonstrates the work flow of a classification tree and how it accommodates interaction and polynomial terms.

The example is based on a previous study by Hong and Raudenbush (2005) about the effects of kindergarten retention on students' reading and math abilities. In this study, the authors explored the ECLS-K 98-99 dataset and identified 38 covariates that they believed influenced students' retention/promotion status at the end of kindergarten. A subset of the dataset in this study is used to illustrate the work flow of a classification tree. The goal is to predict first-time kindergartners' retention/promotion status at the end of an academic year from influential factors. To simplify the illustration, instead of using all the 39 covariates, I only use two most influential (based on information gain) covariates, which are continuous variable *read_wave2* (students' reading score in the 2nd semester of the academic year), and continuous variable *math_ars* (teachers' report of students' math ARS score). The classification tree derived from the sample data is presented in Figure 3. It shows that the entire sample of 813 students is first split

into two groups (Node 2 and 5) based on whether or not the student's math score is higher than 2.11. Then on the left branch, the group of students with math scores lower than or equal to 2.11 is further split into two subgroups (Node 3 and 4) based on students' reading scores. These two subgroups vary in the percentage of students who are retained at kindergarten. Similarly, on the right branch, the group of students with math score higher than 2.11 is further split into two subgroups (Node 6 and 7) based on the same covariate math_ars with a different cut-off value. From this example, we can see that the effect of reading score on the outcome, students' kindergarten retention/promotion status, differs for students with different math scores. Specifically, for students with math scores lower than or equal to 2.11 there is an effect of reading score, while for those with math scores higher than 2.11 there is no such effect. In this way, the classification tree accommodates an interaction term. The way to accommodate polynomial terms is similar as a polynomial term can be regarded as the multiplication of the same covariate several times. The right branch of the tree shows an effect of the quadratic term math_ars. As described earlier, the PS is estimated by the proportion of treated subjects in each terminal node. For example, in this case, the percentages of retained students in Node 3, 4, 6, and 7 are 54.5%, 13.3%, 6.8% and 0.5% respectively, thus the PS is 0.545 for any student in Node 3, 0.133 for any student in Node 4, 0.068 for any student in Node 6, and 0.005 for any student in Node 7.

Figure 3

An example of a classification tree

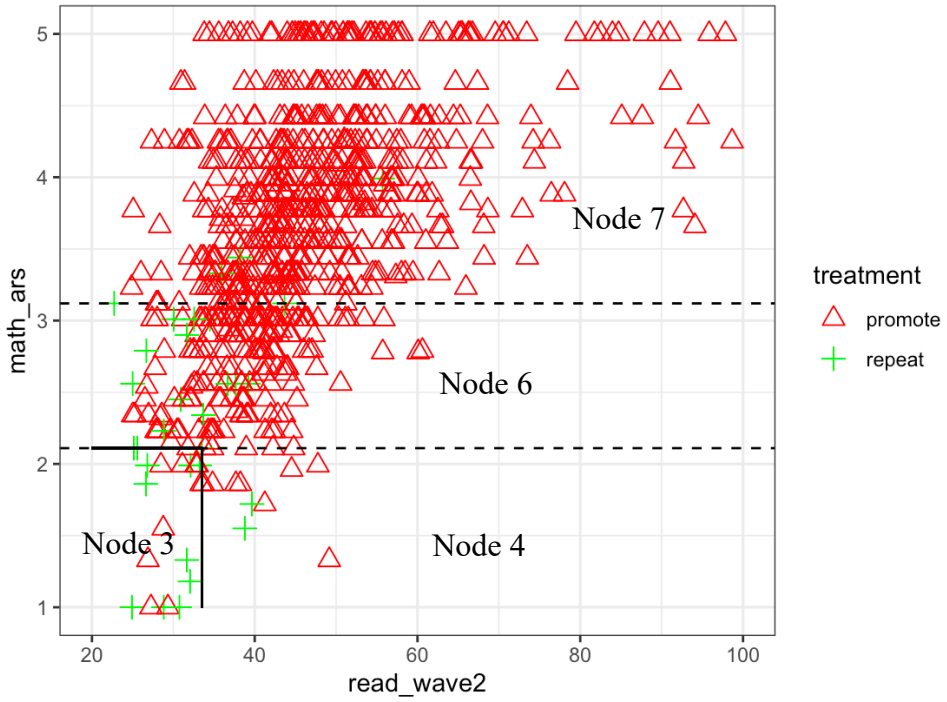


The tree structure, as shown in Figure 3, provides one way to present the work flow of a tree model. Another way to demonstrate the work flow is a rectangular partition of the covariate space as you can see in Figure 4. Firstly, the entire sample is split into two areas (Node 3 + Node 4 and Node 6 + Node 7) based on the covariate math_ars. Then each of these two areas is further split into sub-areas, making each sub-area more homogeneous in terms of the treatment status. Finally, the treatment condition of a student is determined via majority vote and the PS is determined by the proportion of treated students in the sub-area where this student belongs to. Compared with the tree representation in Figure 3, the partition representation in Figure 4 can better present the decision boundary, which is represented by the rectangular solid line. Students

below the line are predicted to be retained at kindergarten while those above the line are expected to be promoted to 1st grade.

Figure 4

Partition plot of a classification tree



From the above partition plot (Figure 4), we can see that a potential problem of a single classification tree is that the decision boundary is rectangular which may lead to inaccurate prediction of the class membership and the PS if the true decision boundary is not orthogonal to the coordinate axes (Dietterich, 2002). Another potential problem of a single classification tree is its instability to small changes in the datasets (Strobl et al., 2009). From the above description of the recursive partitioning process of classification trees, we can see that any change of the datasets may lead to a change of the choice of the splitting covariate and the cut-off value, thus

influencing how the subjects are split into new nodes, and all the splits afterwards. A small change of the dataset may lead to a very different tree structure. A solution to the above problems is to use some ensemble methods, using the average over an ensemble of multiple trees, rather than a single tree. Both empirical and simulation studies indicate that ensemble methods, such as bagging and random forests, can provide more reliable and accurate predictions than single trees (Bauer & Kohavi, 1999; Breiman, 1996; Dietterich, 2000). In the next section, I provide a detailed description of one of these ensemble methods, random forests.

2.3.1.2 Random forests.

The random forests method involves multiple classification trees and works as follows.

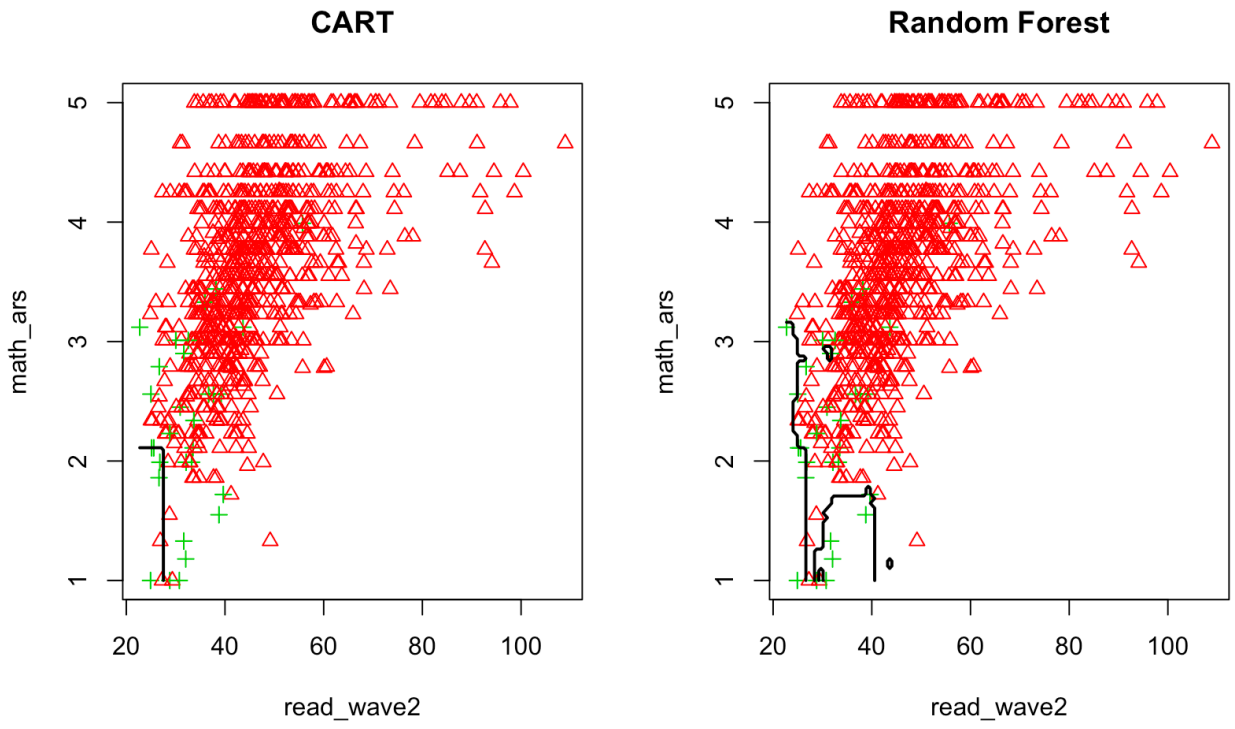
1. Draw multiple random samples from the original data.
2. Fit a classification tree on each of the samples. Different from the traditional classification tree, at each split the splitting covariate is random selected from only a subset of the covariates instead of all the covariates.
3. For each subject, estimate the class membership or the PS from each classification tree.
4. For each subject, use majority vote to determine the final class membership from the multiple estimated class memberships, or take an average of the multiple estimated PSs from the multiple trees to obtain the final estimate of the PS.

Through combining the estimates of multiple trees, random forests can approximate a smooth decision boundary which may be better aligned with the true decision boundary than the rectangular decision boundary provided by a single classification tree. Thus, random forests can better differentiate subjects of different class memberships than a single classification tree. In the previous example of kindergarten retention, the decision boundary of random forests is much

smoother than that of a single classification tree as you can see in Figure 5. And the AUC (area under the receiver operating characteristic curve), a measure of classification performance when the size of one class is much larger than that of the other class, increases from 0.71 to 0.80 when changing the model from a single classification tree to random forests. Besides, the smooth decision boundary of random forests also indicates its flexibility in accommodating different kinds of non-linear relations between the outcome and the covariates.

Figure 5

Decision boundaries of a tree and random forests



The superiority of random forests over logistic regression in recovering different kinds of non-linear relations between the log odds of the class membership or treatment assignment and the covaraites is also demonstrated using the following simulation examples. As we know, when

there are multiple covariates it is hard to directly visualize the relation between the outcome and the predictors. Alternatively, Friedman (2001) developed a method called partial dependence plots which can visualize the effect of each covariate on the outcome while averaging all the other covariates. Since its development, partial dependence plots have been widely use in many studies to show the relation between the outcome and covariates (Couronne et al., 2018; MacCaffrey et al., 2004; Strobl et al., 2009). Usually, in a partial dependence plot, the x-axis represents a covariate while the y-axis represents the estimated probability of receiving the treatment condition based on a specific model. Here, to illustrate random forests' ability in accommodating different kinds of non-linear relations, I generated six simulated datasets based on the following models:

Main effects only: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 1 + 3x_1 - x_2$,

Interaction terms: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 1 + x_1 - x_2 + 2x_1 * x_2$,

Polynomial terms: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -2 + 3x_1^2 + x_2$,

Interaction and polynomial terms: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 1 + 3x_1^2 - x_2 + 2x_1 * x_2$,

Exponential terms: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -e^{x_1} + x_2$,

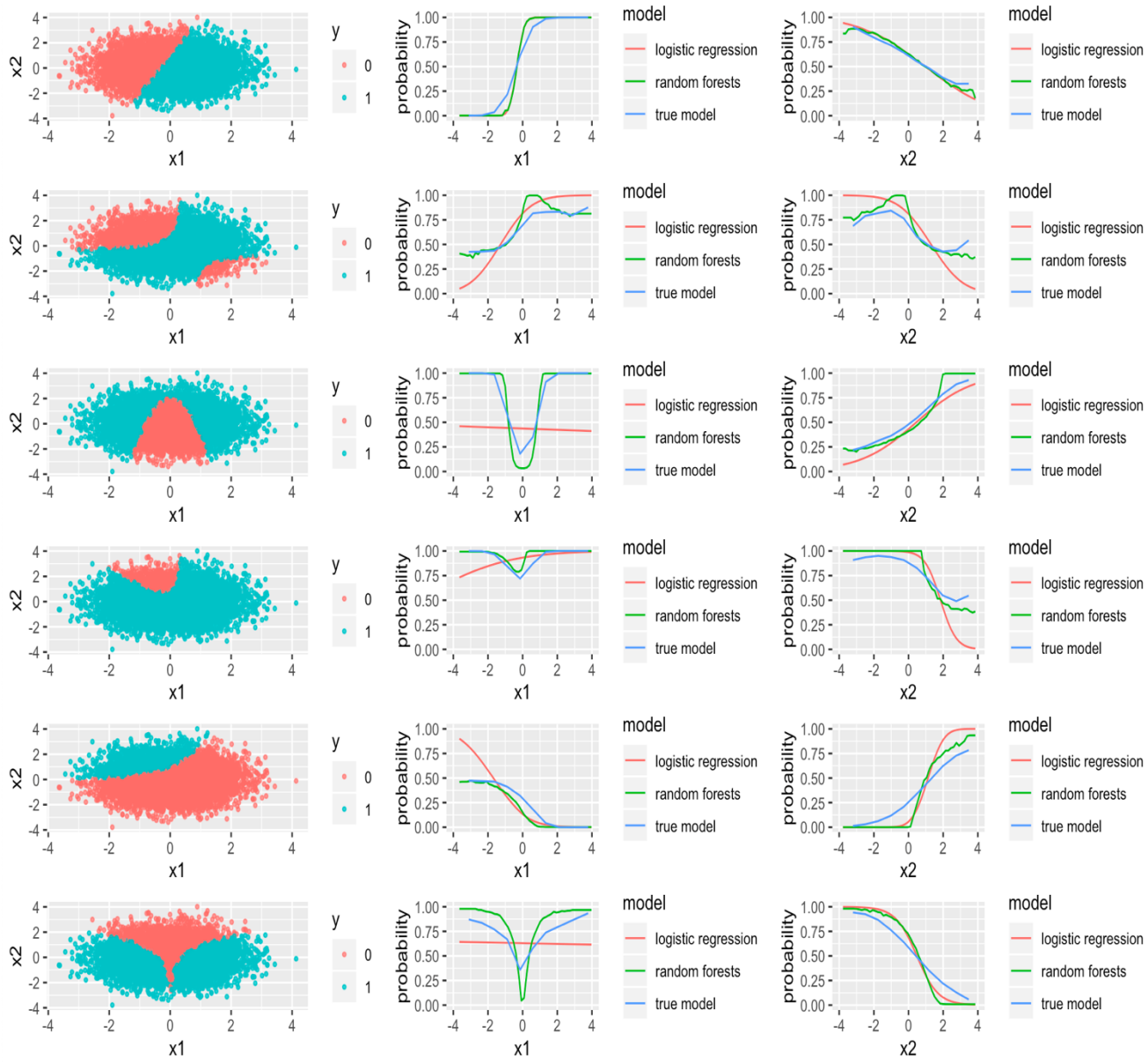
Log terms: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 1 + \log(|x_1|) - x_2$,

Where x1 and x2 are randomly generated from a standard normal distribution and the sample size is 10,000. For each dataset, a main effects only logistic regression model and a random forests model (with default hyper-parameters) are fitted and used to estimate the probabilities of receiving the treatment condition. Figure 6 shows the distribution of the data points (the first column) and the partial dependence plots of x1 and x2 respectively (the second and the third

columns). We can see that, in general, the relation between the covariate and the probability estimated by the random forests model is more aligned with the true relation, especially when nonlinearity exists in the datasets. This indicates that random forests can better deal with different kinds of non-linear relations than logistic regression, or more accurately the most commonly used main effects only logistic regression models.

Figure 6

Partial Dependence Plots



Note. From top to bottom, the scenarios are main effects only, interaction terms, polynomial terms, both interaction and polynomial terms, exponential terms, and log terms.

In this section, I introduced the theoretical framework of random forests and explained how it accommodates different kinds of non-linear relations. The section below provides some

details about its implementation as well as how to choose the hyperparameter values in random forests to improve model performance.

Different packages in various software have been developed to conduct random forests estimation, including the R packages party (Hothorn et al., 2006) and randomForest (Breiman & Cutler, 2001), the Python package Scikit-Learn (Pedregosa et al., 2011), and Stata. In this dissertation, I focus on the R packages (especially randomForest) as they are used often by applied researchers.

In conventional applications of making classification, for random forests to achieve a good performance (higher classification accuracy), it requires good performance of each classification tree and diversity between trees (Breiman, 2001). Several hyperparameters are known to influence the performance of random forests: the subject sampling method, the maximum depth of each tree or the minimum terminal node size, the number of covariates to select from at each split, the number of trees, and the splitting criterion, among which the first three are found to be most influential (Lin & Jeon, 2006; Orlandi et al., 2016; Probst et al., 2019; Rijin & Hutter, 2018; Segal et al., 2004). The section below explicates these three key hyperparameters, their impacts on the model, as well as how to choose the values of these hyperparameters.

Subject sampling method. Random sampling of the subjects can be conducted with or without replacement and with a designated proportion of the original sample size. The proportion of subjects to be sampled is important because it influences the strength of each classification tree and the diversity between trees. Ideally, the proportion should be neither too large nor too small as an extremely large proportion will reduce the diversity between trees while a too small proportion will weaken the performance of each classification tree. In practice, researchers have

different strategies. When Breiman first introduced the method of random forests, he used sampling with replacement (also called bootstrap sampling) with the original sample size N. The commonly used R package randomForest adopted this practice and set bootstrap sampling as the default setting. Hothorn et al. (2006) used sampling without replacement (also called subsampling) with 0.632 of the original sample size as an approximation to bootstrap sampling with the original sample size because theoretically, on average, a bootstrap sample of size N contains about 0.632*N different subjects. Besides, subsampling is more efficient than bootstrap sampling. Thus, Horthorn and his colleagues made subsampling with 0.632 of the original sample size as the default setting in their R package party. Strobl et al. (2007) used a simulation study to compare the performance of bootstrap sampling with the original sample size and subsampling with 0.632 of the original sample size and found that subsampling with a size of 0.632 of the original sample size can reduce the covariate selection bias towards continuous covariates and categorical covariates with multiple levels. Friedman and Hall (2007) investigated different sampling schemes and found that subsampling with half of the original sample size is equivalent to bootstrap sampling with the original sample size.

Number of covariates from which to select at each split. In random forests, at each split, the splitting covariate is selected from a subset of the covariates instead of all the covariates as this approach can increase the diversity between trees and thus improve the performance of random forests (Breiman, 2001). Regarding the number of covariates from which to randomly sample at each split (usually denoted as mtry), the optimal value differs for different datasets, but in general $mtry = \sqrt{p}$ (round up) is a reasonable setting in most empirical studies with p being the total number of covariates (Hastie et al., 2009; Liaw & Wiener, 2002; Strobl et al., 2008). The R package randomForest followed this suggestion and set $\sqrt{p}$ as the default value of mtry.

Another commonly used value of mtry is log2p + 1 (round up), where p is the total number of covariates. This choice was first proposed by Breiman (2001) and was later justified by Bernald et al. (2007) through an empirical study on a well-known hand-written digit recognition dataset. Bernald and his colleagues tried various values of mtry and found that an interval of mtry values could achieve best accuracies; this interval contains the two values $mtry = \sqrt{p}$ and mtry = log2p + 1.

The maximum depth of each tree/the minimum terminal node size. These two hyperparameters are both used to determine the size of the trees in random forests. Regarding whether and when to prune the trees, researchers have different opinions. Originally, when Breiman (2001) developed random forests, he suggested growing each tree to maximal depth, continuing the splitting process until each terminal node contains purely one kind of subject because 1) trees with maximal depth can reduce bias, 2) the overfitting issue of fully-grown trees is not a problem in random forests. Breiman examined 16 empirical datasets from the UCI machine learning repository and found that there was no overfitting issue of random forests with fully-grown trees. For about three years, researchers followed Breiman's suggestion and used random forests without pruning (Liaw & Wiener, 2002). However, in 2004, this notion was criticized by Segal (2004). Segal revisited the datasets Breiman used and found that these datasets are hard to overfit using tree-structured models. By investigating the performance of random forests on some real-world and simulated datasets, Segal found that maximally sized trees did overfit. This finding was supported by Lin and Jeon (2006). They examined the classification accuracy of random forests models with various terminal node sizes on some simulated datasets and found that pruning trees could enhance prediction accuracy, especially when the number of covariates is small and the sample size is large.

In the context of PS methods, pruning trees is even more important because the goal is to estimate the probability instead of predicting the class membership. If we allow trees to grow until purity, that is, until every terminal node contains one kind of subject, the probability estimate will be either 0 or 1. As a consequence, a large number of trees may be needed to obtain reliable probability estimates aggregated over the multiple trees (Dankowski & Ziegler, 2016; Kruppa et al., 2013). However, on the other hand, if we stop the splitting process too early and prune the trees too much, the performance of each tree may be poor to provide accurate probability estimates. To choose the optimal size of the trees, researchers used minimum terminal node size more often than maximum depth of trees because 1) the depth of trees needed differs for datasets of different sample sizes and complexity levels, 2) setting the maximum depth of trees cannot fully avoid pure terminal nodes and extreme probability estimates (Lin & Jeon, 2006; Dankowski & Ziegler, 2016). For example, in an asymmetric tree where the left branch of the tree continues growing while the right branch reaches purity very quickly, setting the maximum depth of the tree can only avoid extreme probability estimates in the left branch but does not work for the right branch. About the value of the minimum terminal node size, the default value in the R package randomForest is 1. As discussed above, it may not be a good choice for probability estimation. Recently, researchers found that the convergence properties of random forests required the terminal node size to increase with the sample size (Dankowski & Ziegler, 2016; Lin & Jeon, 2006). The terminal node size of 10% of the total sample size is advocated by some recent studies (Dankowski & Ziegler, 2016; Kruppa et al., 2013; Malley et al., 2012).

In addition to the hyperparameters described above, in random forests there are some other hyperparameters, such as the number of trees, and the criterion to select the splitting

covariate. For the number of trees, theoretically the value should be as large as possible as the performance of random forests is found to be monotonically improving with the number of trees (Probst & Boulesteix, 2018). However, in practice, researchers usually set the value to be a few hundred because 1) the performance usually reaches a plateau when the number of trees increases to a few hundred, 2) adding more trees will increase the computation time. Oshiro et al (2012) and Probst and Boulesteix (2018) suggested setting the value to be about 100 while Liaw and Weiner (2002) used 500 trees and set it as the default value in their R package randomForest. For the splitting criterion, in general there are two types of decision rules. The first and most common type of splitting criterion is to search for a covariate and a corresponding cut-off value that maximizes the reduction of the impurity measure (e.g., entropy) after the split. This type of splitting criterion has been widely used by researchers and has been adopted by most random forests packages, including the prevailing R package randomForest and the Python package Scikit-Learn. The second type of splitting criterion is based on two-stage conditional significance tests with the first stage selecting the splitting covariate and the second stage selecting the cut-off value (Hothorn et al., 2006). Specifically, at the first stage, a hypothesis test is conducted for each covariate where the null hypothesis is that the covariate is not related with the class membership given the other covariates. The covariate with the smallest p-value is selected. At the second stage, another hypothesis test is conducted on the selected covariate with various cut-off values where the null hypothesis is that the proportions of treated subjects are equal between the two nodes after the split. Similarly, the cut-off value with the smallest p-value is selected. This splitting criterion is adopted in the R package party but is used less often. A possible reason is that the computation time is much larger than that using the first splitting criterion because of

the huge number of tests. Besides, I doubt about the reasonability of this criterion as the smallest p-value does not necessarily mean most influential covariate.

In summary, in this section I discussed the various hyperparameters in random forests, their impacts on model performance, and commonly used values for each hyperparameter. Among all these hyperparameters, researchers found that the number of covariates to select from at each split, minimum terminal node size, and subject sampling method are most influential. Rijin and Hutter (2018) assessed the importance of the hyperparameters in random forests using a method called functional ANOVA, which analyzes the importance of hyperparameters based on a regression model that yields predictions of the model performance (classification accuracy or AUC) for different hyperparameter values. They conducted an analysis on 100 datasets from the OpenML repository and found that the top three influential hyperparameters in random forests are minimum terminal node size, number of covariates to sample from at each split, and subject sampling method (with or without replacement) as these three hyperparameters explain most of the variance in the model performance. Probst et al. (2019) used tunability to measure the importance of the hyperparameters in random forests and got similar findings. In this study, tunability was defined as the difference in the model performance (AUC) between a reference hyperparameter configuration and the best possible hyperparameter configuration for a given dataset, where the reference hyperparameter configuration is the global optimal hyperparameter configuration based on an extensive number of empirical experiments on 38 datasets from the OpenML respository. To assess the tunability of one hyperparameter, the authors fixed other hyperparameters, tried multiple values for the target hyperparameter, calculated the tunability of this hyperparameter on each of the 38 datasets, and then took an average across datasets. Using this method, Probst and his colleagues found that tuning the hyperparameter mtry provided the

biggest performance improvement (AUC), which was followed by subject sampling method (sampling proportion and with or without replacement) and minimum terminal node size. Thus, in this study, I focus on tuning these three hyperparameters and exploring their impacts on the estimates of the PSs and subsequently the estimates of the treatment effect.

Regarding the optimal values of the hyperparameters, there is no one configuration that fits all datasets as the optimal values depend on characteristics of the datasets and also interactions between hyperparameters. For example, as discussed earlier, researchers found that a comparatively large terminal node size is needed if the sample size is large (Bernald et al., 2009; Dankowski & Ziegler, 2016; Lin & Jeon, 2006). Another example is that theoretically more trees are needed for lower sampling proportions, higher terminal node sizes, and smaller mtry values as such settings will increase the diversity between trees (Probst et al., 2019). Basically, the optimal hyperparameter values differ for different datasets.

Then how to obtain the optimal hyperparameter values for a given dataset? In the conventional application of classification, the optimal values of the hyperparameters are usually determined through out-of-bag estimation, a process similar to cross-validation. Specifically, researchers usually try multiple sets of hyperparameter values (based on characteristics of the dataset and researchers' experience) and choose the set with the best performance (smallest classification error or largest AUC) on the out-of-bag subjects, which are subjects not used to train the model. In detail, it works as follows: 1) for each tree, a model is built on a subset of the whole sample, either a bootstrap sample or a subsample, 2) the fitted model is used to predict the class memberships for the out-of-bag subjects, 3) for each out-of-bag subject, the final class membership is determined through aggregating the results over multiple trees, 4) the misclassification rate for the out-of-bag subjects (usually called out-of-bag error) is calculated.

Researchers usually repeat this process for every set of hyperparameter values and choose the set with the lowest out-of-bag classification error as the optimal setting of the hyperparameters for a given dataset.

However, in the context of PS methods, the strategy of using out-of-bag estimation to choose the optimal hyperparameter values is not applicable for the following reason. For PS methods, the ultimate goal is not to predict the class membership but to estimate the true treatment effect, thus it does not make sense to use the conventional classification-based metrics, such as misclassification error or AUC, to determine the optimal hyperparameter values. Then can we use the accuracy and/or precision of the estimated treatment effect to evaluate the performance of different hyperparameter values and decide what values to use? Apparently, the answer is no as true treatment effects are usually unknown in empirical studies. Through a simulation study which allows the true treatment effect to be known, this study aims to explore the effects of the three most influential hyperparameters in random forests, the subject sampling method, the number of covariates from which to select at each split, and the minimum terminal node size, on the estimate of the treatment effect, and provide some guidance about how to choose the values of these hyperparameters in the context of PS methods.

2.3.1.3 Other machine learning methods.

Besides random forests, there are some other ensemble methods, such as bagging and boosting. Bagging is a simpler version of random forests. The only difference between bagging and random forests is that in bagging at each split the covariate is selected from all the covariates instead of a random subset (Breiman, 1999). Boosting works through building multiple trees sequentially, with each tree learning from the residuals of the previous tree (Ridgeway, 1999). In addition to these tree-structured models, some other machine learning methods, such as support

vector machine, neural network, etc., have also been explored in making classifications and estimating the PSs. Among all these methods, random forests have been proven to perform very well in providing less biased and more reliable estimate of the treatment effect (Lee et al., 2010; Li, 2016; Nichols & McBride, 2017).

2.3.2 The use of random forests in PS methods and its limitations

As discussed earlier, random forests can automatically select influential covariates and capture nonlinear and nonadditive relations, which can avoid the laborious covariate selection process and potential violation of the model specification assumption in logistic regression (Arpino & Cannas, 2018; Diamond & Skehon, 2013). Comparison studies have shown that random forests usually provide higher classification accuracy than logistic regression (Couronne et al., 2018). Because of these advantages, the method of random forests has been widely used in making classification in many fields such as medicine, biology, and economics (Amini et al., 2016; Cavicchioli et al., 2019; Pham, 2017).

Recently, researchers proposed using random forests as an alternative to logistic regression in estimating the PSs. Results from both empirical and simulation studies indicated the superiority of random forests over logistic regression as well as other machine learning methods in producing accurate and reliable estimates of the PSs and treatment effects. Lee et al. (2010) used a simulation study to compare the performance of logistic regression, classification trees, pruned classification trees, random forests, and boosting in the context of PS weighting at varying degrees of nonlinearity and non-additivity (manipulated through changing the number of interaction and quadratic terms) with varying sample sizes. They found that random forests and boosting provided treatment effect estimates with smaller relative bias and smaller standard error than logistic regression and all of the other explored machine learning methods. And the

superiority of random forest over logistic regression is more salient with increasing nonlinearity and non-additivity. Arpino and Cannas (2018) extended the comparison to more machine learning methods (neural network and naïve bayes) and more PS methods (both PS matching and PS weighting), and obtained similar results. They found that 1) for PS matching, the treatment effect estimate using random forests has the lowest relative bias across all the conditions with different degrees of non-linearity and non-additivity (manipulated through changing the number of interaction and quadratic terms) and with different sample sizes, 2) for PS weighting, the treatment effect estimate using random forests usually has the lowest relative bias, 3) in general the treatment effect estimate using PS weighting has smaller mean squared error than that using PS matching. Also using a simulation study, Nichols and McBride (2017) found that the use of random forests was more robust to the omission of important covariates, interaction terms, and polynomial terms than logistic regression using PS weighting. Zhao et al. (2016) applied random forests-based PS matching to a national health and nutrition survey data and found that it could produce well balanced treatment and control groups.

However, there are some limitations in the current applications of random forests in the context of PS methods. I highlight five of these limitations here. First of all, in the studies that compared the performance of random forests versus logistic regression, the researchers only used the main effects only logistic regression model, which was not able to accommodate a non-linear relation (Arpino & Cannas, 2018; Lee et al, 2011; Nichols & McBride, 2017). However, in practice, experienced researchers could deal with non-linear relations through properly adding interaction and/or polynomial terms, which would probably improve the model performance in balancing the treatment and control groups in the baseline covariates and would possibly lead to more accurate estimates of the treatment effect (Hong & Radenbush, 2005). More research needs

to be done to compare the performance of random forests versus logistic regression with properly identified interaction and/or polynomial terms. Secondly, in the above simulation studies, when specifying non-linear relations in the true PS model the authors mainly used interaction and/or polynomial terms. However, in practice, some other types of non-linear relation may exist, such as log or exponential forms. It is unclear whether the superiority of random forests versus logistic regression in estimating treatment effects still exist if there are some other types of non-linear relation in the true PS model. Besides, in the above simulation studies, when manipulating the degree of nonlinearity, the authors did not explicitly quantify the nonlinearity degree but implicitly specified it through changing the number of interaction and/or quadratic terms. Such practices may have two potential issues: 1) more interaction or quadratic terms does not necessarily mean a higher degree of nonlinearity as the coefficients of these terms also play a role in determining the nonlinearity degree, 2) such practices are not applicable to manipulate nonlinearity degree across different types of nonlinear relations, such as models with interaction or quadratic terms versus models with log or exponential terms. What's more, the previous simulation studies only covered limited scenarios, with limited cases of sample sizes and covariate dimensionalities, or more accurately limited cases of number of subjects per covariate. Most of the simulation studies followed the data generating structure by Setoguchi and his colleagues (2008) and used 10 covariates. Regarding the sample size, Arpino and Cannas (2018), and Lee et al. (2010) used 500, 1000, and 2000, which means the number of subjects per covariate is 50, 100, and 200. Cham (2013) used a different way to generate the data. Instead of using the structure by Setoguchi and his colleagues, he generated the data based on an empirical dataset. He set the sample size to be 600 and 2000, and set the total number of covariates to be 64, making the number of subjects per covariate to be about 10 and 30. A recent study by Ploeg

et al. (2014) compared the performance of random forests and logistic regression in making classification with varying numbers of subjects per variable, and found that the method of random forests needs more subjects per variable (sometimes more than 200) to achieve a stable estimate than the method of logistic regression. Although in PS methods, the purpose of a random forests model is to estimate the PSs instead of making classification, it would also be of interest to explore whether the superiority of random forests over logistic regression still exists in different conditions of number of subjects per covariate.

Last but not least, in most of these studies, researchers used the default values of the hyperparameters in the R package randomForest. There is a lack of exploration on the effects of the hyperparameters on the estimates of the treatment effects. To the best of my knowledge, Cham (2013) is the only study that explored the effects of hyperparameters on the estimates of the treatment effects in the context of PS methods. In that study, Cham manipulated the type of nonlinear relation in the true PS model, the sample size and the magnitude of the true treatment effect, and explored the effects of three hyperparameters, the decision rule to select the splitting covariate (significance test VS gini index), the number of covariates to sample from at each split (mtry = p, $mtry = \sqrt{p}$, and optimal mtry based on a pilot study), and the method for estimating PSs (using the out-of-bag sample or the whole sample), on the estimate of the treatment effect in the conditions of different sample sizes with different sizes of treatment effect. Specifically, he generated simulated data based on an empirical example (Im et al., 2013) and set four true PS models with various kinds of non-linearity, which were main effects only logistic regression model, logistic regression with two-way interaction effects, logistic regression with quadratic effects, and logistic regression with both two-way interaction and quadratic effects. The outcome was a continuous variable with the values being specified through a regression on the covariates

and the true treatment effect. Two true treatment effects were examined, which were 0 and 0.73. Two levels of sample size were used, 600 and 2000. Cham used bias and standard error of the estimated treatment effects to evaluate the performance of different hyperparameter settings and found that, in general, using the decision rule of significance test, a subsample of the covariates at each split, and the out-of-bag sample to estimate the PSs led to more accurate (less bias) and precise (smaller standard error) estimate of the treatment effect. However, the best hyperparameter configuration (lowest bias) among all the specified settings differs for different conditions. For example, for sample size 600 using significance test as the splitting criterion, subsampling covariates, and out-of-bag estimation yields the most unbiased estimate of the treatment effect while for sample size 2000 the configuration of gini index, all covariates, and whole-sample estimation works best. Besides, there are some flaws of Cham's study. First, two of the three hyperparameters Cham explored, the decision rule to select the splitting covariate and the PS estimating method, were not the top influential hyperparameters (Rijin & Hutter, 2018). Second, Cham only compared the performance of random forests models with different hyperparameters but did not compare the performance of random forests versus logistic regression, thus no guidance can be provided about when to use random forests over logistic regression. Third, it is unclear whether Cham's approach appropriately obtained the estimates of the PSs given his use of out-of-bag sample or total sample instead of the in-bag sample to estimate the PS.

2.4 Research Questions

This dissertation compares the performance of random forests versus logistic regression on the estimates of the treatment effects in the context of PS methods at various sample sizes, with various numbers of covariates, and with different types and varying levels with nonlinearity

and non-additivity. Different from the previous comparison studies which only compared the performance of random forests versus main effects only logistic regression model, this study also compares the performance of random forests versus logistic regression with interaction and/or polynomial terms. Moreover, unlike previous studies which only explored the nonlinear relation which is composed of interaction and/or polynomial terms, this study expands the analysis to other types of nonlinear relations, such as log or exponential forms. And the nonlinearity degree will be measured and manipulated through a more scientific way (details are provided in Section 3.1). In addition, this study explores more conditions with various sample sizes and various numbers of covariates, or in other words different numbers of subjects per covariate, to check whether the previously found superiority of random forests over logistic regression still exists for different numbers of subjects per covariate. Results from this study provide a fairer and more comprehensive comparison of random forests versus logistic regression on the estimates of the treatment effects in the context of PS methods.

Another important objective of this study is to explore the effects of the hyperparameters in random forests on the estimates of the treatment effects in the context of PS methods. With regard to the hyperparameters, I focused on minimum terminal node size, number of covariates to select from at each split, and subject sampling method as they were found to be the most influential hyperparameters in random forests (Rijin & Hutter, 2018). Among all the PS methods, I focused on PS weighting because it relies more on accurate estimates of PSs than PS matching and subclassification (Lee et al., 2010). The effects of interest were examined in terms of the accuracy and precision of the estimates of the treatment effect. A simulation study was conducted to answer the following questions:

1. Under what conditions does the method of random forests perform better than logistic regression in estimating the treatment effect?
2. In what way do the values of the hyperparameters in random forests influence the estimates of the treatment effect? Will tuning influential hyperparameters improve the accuracy and precision of random forests in estimating treatment effect?

The answers to these questions add to the literature about the comparison of random forests versus logistic regression in estimating PSs and provide some guidance about when to use which method. Additionally, it also provides some suggestions about how to choose the values of the hyperparameters in random forests to obtain an optimal performance in PS weighting.

In addition, an empirical data analysis was conducted as an illustration of the choice of the values of the hyperparameters and its impacts on the estimates of the treatment effect. Chapters 3 and 4 explicate the simulation used to evaluate the two research questions. Chapter 5 discusses the empirical study for the demonstration analysis.

Chapter 3. Simulation Study Design

This chapter describes the design of the simulation study to address the primary research questions. First, I start with an explanation about the factors manipulated, both between-cell and within-cell factors, and what levels were used for each factor. Then, I introduce the data generation, model fitting, and estimation process. Next, I describe the metrics used to evaluate the performance across different conditions. Finally, I discuss the implications and limitations of this stimulation study.

3.1 Manipulated Factors

Two kinds of factors were manipulated, between-cell factors and within-cell factors.

3.1.1 Between-cell factors.

For the between-cell factors, I focus on the number of subjects per covariate, number of covariates, type and degree of nonlinearity between the covariates and the treatment assignment ~~of the PS model~~ as well as the number of nonlinear terms in the PS model. Among these factors, the number of subjects per covariate and the nonlinearity degree of the PS model have been found to influence the superiority of random forests approach over logistic regression. Number of covariates and types of nonlinearity of the PS model are included as between-cell factors to provide a more comprehensive comparison of random forests versus logistic regression in different conditions. The number of nonlinear terms is also included because in the data generation process, I did a pilot study and found that the number of nonlinear terms also influenced the relative performance of random forests over logistic regression when holding other factors constant (Appendix A). Below I explicate the levels that I used for each factor as well as the rationale.

Number of subjects per covariate. As discussed earlier, more subjects per covariate are needed to achieve a stable estimate of the classification using random forests (Ploeg et al., 2014). In this study, I used three levels, 20, 200, and 500, among which 200 was used by previous simulation studies (Arpino & Cannas, 2018; Lee et al., 2010). I added the condition of 20 subjects per covariate as it was a widely-advocated minimal criterion for logistic regression (Harrell, 2001; Peduzzi et al., 1996). The condition of 500 subjects per covariate was included because it represented some commonly occurring situations in empirical studies (Baser, 2006; Dagostino, 1998; Rosenbaum & Rubin, 1984).

Number of covariates. The number of covariates was manipulated because 1) as discussed in Section 2.3.1.2, the total number of covariates may influence the choice of the values for some important hyperparameters in random forests, such as the number of covariates from which to select at each split, and the minimum terminal node size (or in other words, pruning trees or not), 2) in practical applications of PS methods the number of covariates varies (Thoemmes & Kim, 2011). Thoemmes and Kim (2011) reviewed 111 empirical studies from 2003 to 2009 that used PS methods, and found that the 25th percentile, median, and 75th percentile number of covariates are 8, 16, and 24. Based on this and the widely used data generation structure by Setoguchi and his colleagues (2008) (to be discussed in Section 3.2), I set the number of covariates to be 10, 20, and 40. For these conditions, 7, 14, and 28 of the covariates are treatment predictors respectively (the remaining covariates are outcome predictors). More details about how the covariates are used in the data generation and PS estimation process are presented in Section 3.2 and Section 3.3 respectively.

Type of nonlinearity. As discussed before, one advantage of the use of random forests over logistic regression is that random forests can automatically address nonlinear relations

between the outcome and the predictors. In this study, I manipulated the nonlinearity type of the true PS model using a similar way as what Lee et al. (2010) and Setoguchi and his colleagues (2008) used with some modifications. Specifically, I used logistic regression models as the true PS models. The type of nonlinearity was manipulated through adding different kinds of nonlinear terms to the main effects only logistic model. In addition to the previously explored nonlinear relations that are composed of interaction and/or polynomial terms, I extended the analysis to the nonlinear relation with exponential terms, which is another common nonlinear relation in many empirical studies of social and natural science research. For example, the death risk due to cardiovascular disease was found to rise exponentially with increasing age (Goldstein et al., 2017). Global population grows exponentially with increasing food supply (Hopfenberg & Pimente, 2001). Besides these three types of nonlinear relations, a condition defined by a linear relation defined by main effects only logistic regression was also included.

Degree of nonlinearity. The degree of nonlinearity was also manipulated as previous studies indicated that the superiority of random forests over logistic regression was more salient with increasing nonlinearity of the true PS model. Different from previous studies which manipulated the nonlinearity degree through changing the number of interaction and/or polynomial terms, I used a method to quantify the nonlinearity degree. In this study, nonlinearity degree was measured as the average of the absolute difference between the true PSs and the estimated PSs based on the best fitting main effects only logistic regression model (Equation 17), where the best fitting main effects only logistic regression model was determined via fitting various main effects only logistic regression models with different sets of covariates and comparing the model fit of these models (using ANOVA tests).

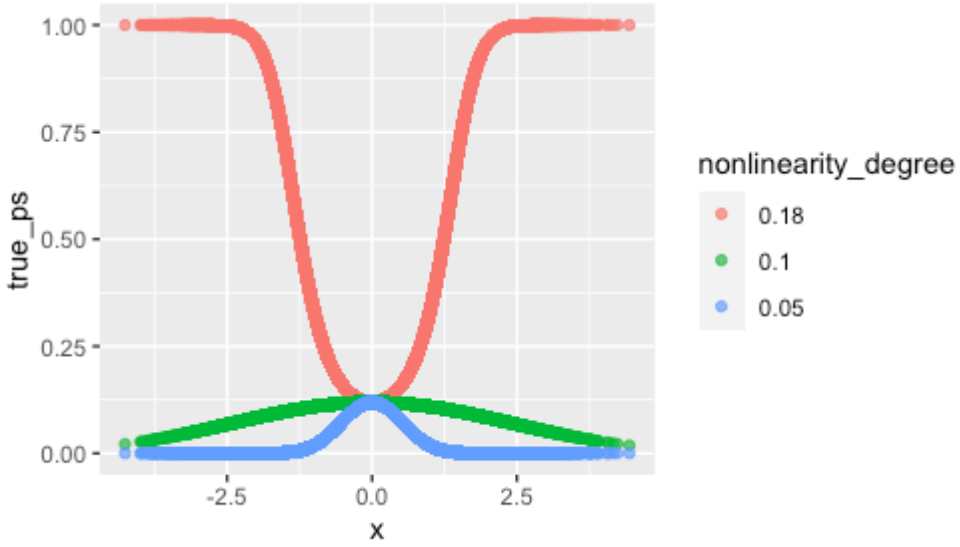
$$nonlinearity\ degree = \frac{\sum |e_{i.est} - e_i|}{n}, \tag{17}$$

Where e_i is the true propensity score of subject i, $e_{i.est}$ is the propensity score of subject i that is estimated from the best fitting main effects only logistic regression model, and n is the total number of subjects. This method was adapted from some previous studies about measures of nonlinearity, in which nonlinearity was usually measured as the average deviation of the response curve from an ideal straight line (Emancipator & Kroll, 1993; Li, 2012). For each of the nonlinearity types, I used three levels of nonlinearity degree: small, medium, and high level of nonlinearity degree. The specific values for each level of nonlinearity were determined through multiple trials to ensure sufficient difference between two adjacent nonlinearity degree levels, practical feasibility of the true PS model, and better comparison between different nonlinearity types and better comparison with previous studies. Specifically, for each nonlinearity type, I set the high-level nonlinearity degree as the 99th percentile rank (quadratic: 0.18; interaction: 0.27; exponential: 0.05). For the low-level nonlinearity, I set it to be 0.05 for the nonlinearity types of quadratic and interaction as 0.05 was at the lower end of the nonlinearity degree range for these two types of nonlinearity degree and it enabled better comparison with previous studies (the study of Lee et al. also included a condition with nonlinearity degree of 0.05). About the medium level nonlinearity, I set it to be 0.1 for the nonlinearity types of quadratic and interaction as 0.1 was in the middle of the nonlinearity degree range for these two types of nonlinearity degree and using the same medium level nonlinearity degree enabled better comparison between these two nonlinearity types. For the nonlinearity type of exponential, I set the low and medium level nonlinearity degree to be 0.01 and 0.03 respectively due to the small range of the nonlinearity degree for this nonlinearity type. Figures 7 and 8 show examples of these three levels of nonlinearity for the nonlinearity types of quadratic and exponential respectively (no figure for the nonlinearity type of interaction is displayed due to its high dimensionality of the covariates).

In the simulation study, the magnitude of nonlinearity degree was manipulated through adjusting the coefficients of the true PS model.

Figure 7

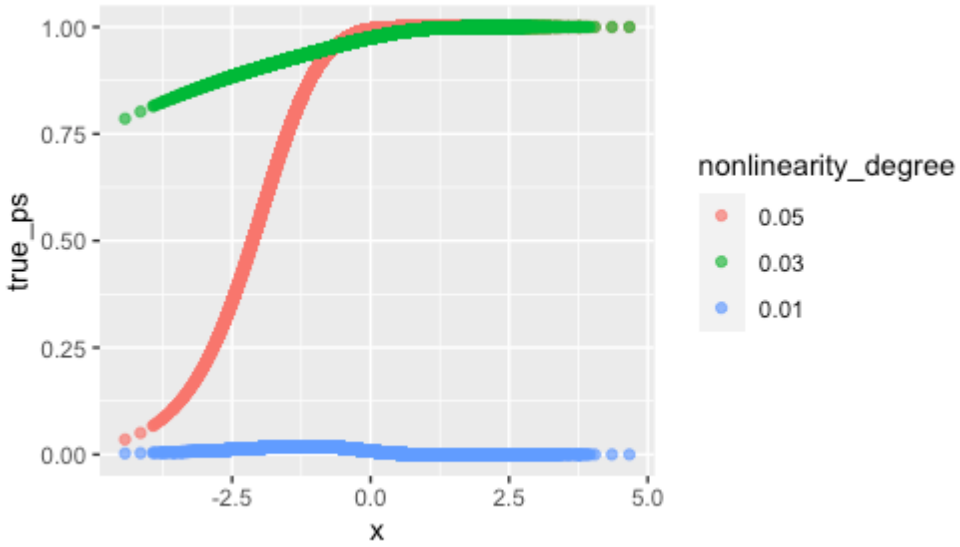
Example of different levels of nonlinearity degree for the nonlinearity type of quadratic



Note. For the nonlinearity degree of 0.18, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -2 + 1.3x^2$; for the nonlinearity degree of 0.1, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -2 - 0.1x^2$; for the nonlinearity degree of 0.05, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -2 - 2x^2$.

Figure 8

Example of different levels of nonlinearity degree for the nonlinearity type of exponential



Note. For the nonlinearity degree of 0.05, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 2.46 + 1.31x + 2.73e^{x}$; for the nonlinearity degree of 0.03, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = 2.84 + 0.35x + 0.80e^{x}$; for the nonlinearity degree of 0.01, the true PS is $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = -2.02 + 0.88x - 2.80e^{x}$.

Number of nonlinear terms. In the data generation process, to determine the number of nonlinear terms, I first conducted a pilot study to check whether the number of nonlinear terms had an impact on the relative performance of random forests over logistic regression while holding other factors constant. Results from the pilot study showed that the number of nonlinear terms did matter (Appendix A). Thus, I also included it as another between-cell factor. About the levels, I chose to use one and three nonlinear terms for the condition of 10 covariates. For the conditions of 20 and 40 covariates, the number of nonlinear terms increased proportionally. This

choice was to represent the ratio of linear versus nonlinear terms in some practical studies and also ensured sufficient differences between levels (Hong, 2005).

In summary, there are 171 (3*3*(3*3*2+1)) between-cell conditions (Table 1). Within each of these conditions, I generated the data, and then imposed different estimation models on each dataset. The estimation models included three logistic regression models (a main effects only logistic regression model, a correctly specified logistic regression model, a logistic regression model with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates) and 27 random forests models with different hyperparameters, which were manipulated as within-cell factors. In the next section, I describe the values that I used for each of the within-cell factors.

Table 1

Summary of the simulation conditions

	Factors	Values
Between-cell factors (3 * 3 * (3 * 3 * 2 + 1) = 171)	Number of subjects per covariate	20, 200, 500
	Number of covariates	10, 20, 40
	Nonlinearity types	Main effects only, model with quadratic terms, model with interaction terms, model with exponential terms
	Nonlinearity degree	quadratic: 0.05, 0.10, 0.18 interaction: 0.05, 0.10, 0.27 exponential: 0.01, 0.03, 0.05
	Number of nonlinear terms	10%, 30% total number of covariates
Within-cell factors (3 * 3 * 3 + 3= 30)	RF: node_size, minimum terminal node size	1, 5%, 10% sample size
	RF: mtry, number of covariates to select from at each split	The total number of covariates (p), Sqrt(p), log2p + 1
	RF: sampling_method, subject sampling method	With replacement with full sample(N), Without replacement with 0.632*N, Without replacement with 0.5*N
	Logistic regression	Main effects only, LR with the quadratic term of most unbalanced covariate and its interaction with all other covariates, LR correctly specified

3.1.2 Within-cell factors.

For the within-cell factors, summarized in Table 1, I primarily focused on three hyperparameters in random forests: minimum terminal node size, number of covariates from

which to select at each split, and the subject sampling method, because they were found to be the most influential hyperparameters in random forests (Rijin & Hutter, 2018). Below I explain the values that I used for each hyperparameter and the rationale.

Minimum terminal node size. Based on some previous studies, I use three values, 1 subject, 5%, and 10% of the sample size. Minimum terminal node size of 1 was used because it is the default value in the commonly used R package randomForest, and was often used by applied researchers (Arpino & Cannas, 2018; Austin, 2012; Lee et al., 2010). The condition of 10% of the sample size was included because it has been advocated by some recent studies (Kruppa et al., 2013; Dankowski & Ziegler, 2016). The condition of 5% of the sample size was included to fill the gap between 1 subject and 10% of the total sample size to make more comprehensive comparison.

Number of covariates from which to select at each split (mtry). As discussed in Section 2.3.1.2, in general, choosing the splitting variable from a subset of the total number of covariates can increase the diversity between trees and enhance the performance of random forests. For this factor, I used three levels: the total number of covariates (p), the square root of the total number of covariates ($\sqrt{p}$), and log2p + 1 (round up). The total number of covariates was used because 1) it can be used as a benchmark to evaluate the influence of this hyperparameter, 2) Strobl et al.'s (2008) simulation study found that increasing the number of covariates from which to choose (from 1 to sqrt(p) to 2/3*p) could reduce covariate selection bias. The square root of the total number of covariates ($\sqrt{p}$) is included because 1) it is the default value in the R package randomForest, and 2) it was suggested by some previous simulation studies (Liaw & Wiener, 2002; Hastie et al., 2009). log2p+1 was also included as it

was another commonly used value of mtry and was found to perform well in some classification problems (Bernald et al., 2007).

Subject sampling method. As discussed earlier, using a subsample of the subjects to build each classification tree can increase the diversity between trees and enhance the performance of random forests. In this study, I used sampling with replacement (also known as bootstrap sampling) with the original sample size, sampling without replacement with 0.632 times the original sample size, and sampling without replacement with 0.5 times the original sample size. Sampling with replacement with the original sample size is the default setting in the R package randomForest. Sampling without replacement with 0.632 times the original sample size is a recommended setting by some researchers (Liaw & Wiener, 2018). The condition of sampling without replacement with 0.5 times the original sample size was also included as it was found to perform equally well as bootstrap sampling with the original sample size but more efficient.

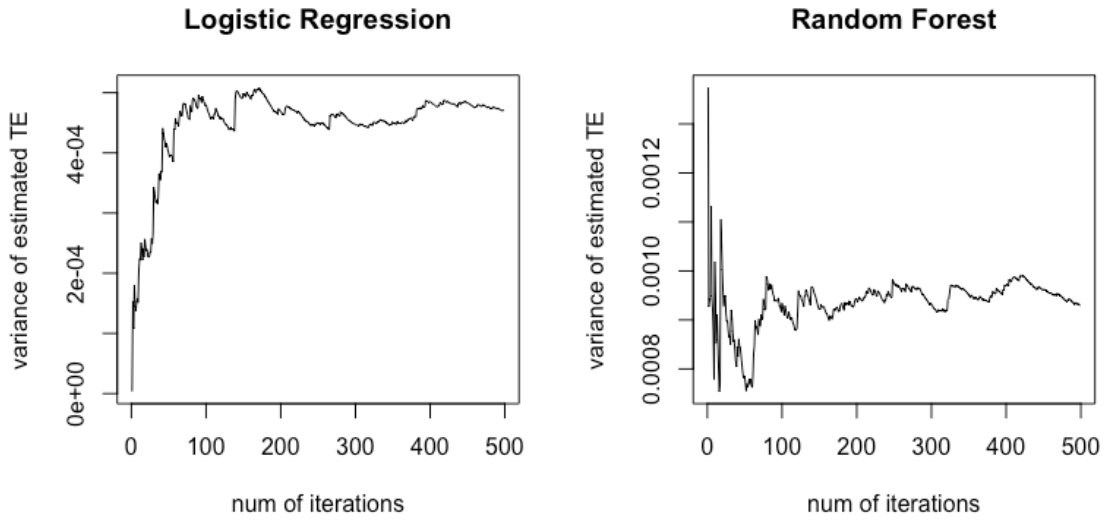
In addition to the above 27 conditions of random forest models with different hyperparameter values, I also included 3 conditions of logistic regression models, specifically a main effects only logistic regression model, a correctly specified logistic regression model, a logistic regression model with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates. The main effects only logistic regression model represents a common practice of applied researchers in estimating PSs while the correctly specified logistic regression model represents the best possible results a logistic regression model could achieve although in practice it is not quite possible to perfectly specify the true model (Bai, 2011; Craycroft, 2016; Shadish et al., 2008; Thoemmes & Kim, 2011). And the logistic regression model with the quadratic term of the most unbalanced covariate and its interaction

with all the other covariates represents one of the approaches that experienced researchers would possibly use when they find that nonlinearity may exist through checking covariate balance.

In summary, within each of the 171 between-cell conditions, there are 30 treatment effect estimates from random forests and logistic regression models. Therefore, in total, there are 5130 (171 * 30) conditions. To determine the number of iterations needed, I conducted a pilot study. Basically, for each condition, I ran 500 iterations, calculated the variance of the estimated TEs after each iteration and found that for both random forests and logistic regression models the TE estimate converged around 200 iterations (Figure 9). Thus, for each of the 5130 conditions, I ran 200 iterations. Table 1 shows a summary of these conditions.

Figure 9

Variance of estimated treatment effect across iterations



Note. This figure shows the convergence plot for the condition of 10 covariates, 200 subjects per covariate, nonlinearity type of quadratic terms, high-level nonlinearity degree, three nonlinear terms. For other conditions, the convergence plot looks similar to this one.

3.2 Data Generation

The data were generated using the simulation structure of Setoguchi et al. (2008) with some modifications. The structure used by Setoguchi et al. has been utilized in several follow-up studies by other researchers, thus using this structure allows my results to be somewhat comparable to the findings of other researchers (Arpino & Cannas, 2018; Lee et al., 2010). Below is a description of the data generation process.

1. Generate the covariates. For each simulated dataset, covariates were generated from a standard normal distribution. Then correlations were induced between some covariates. After that, some covariates were transformed into categorical variables. For example, for the condition that involved 10 covariates, first, 10 covariates were generated from a standard normal distribution. Among these 10 covariates, three covariates only affected the treatment assignment (X5, X6, X7), three only affected the outcome (X8, X9, X10), and the other four (X1, X2, X3, X4) affected both. Then correlations were added between X1 and X5, X2 and X4, X3 and X8, X4 and X9 (Figure 10). Next, covariates X1, X3, X5, X6, X8, and X9 were transformed into binary variables. Specifically, if a value was larger than the average of that covariate, it was categorized as 1 otherwise 0, which was exactly what Setoguchi and his colleagues did. For the conditions with 20 covariates, the data generation process was similar (Figure 11). Specifically, the numbers of treatment predictors, outcome predictors, and the covariates associated with both the treatment assignment and the outcome were all doubled compared with the conditions with 10 covariates. Correlations were added not only between X1 and X5, X2 and X4, X3 and X8, X4 and X9, but also between X11 and X15, X12 and X14, X13 and X18, X14 and X19. The correlation parameter between X11 and X15 was the same as the correlation between X1 and X5,

and the correlation between X12 and X14 was the same as the correlation between X2 and X4, so on and so forth. For the conditions with 40 covariates, data were generated in a similar way.

Figure 10

Simulation structure for conditions with 10 covariates

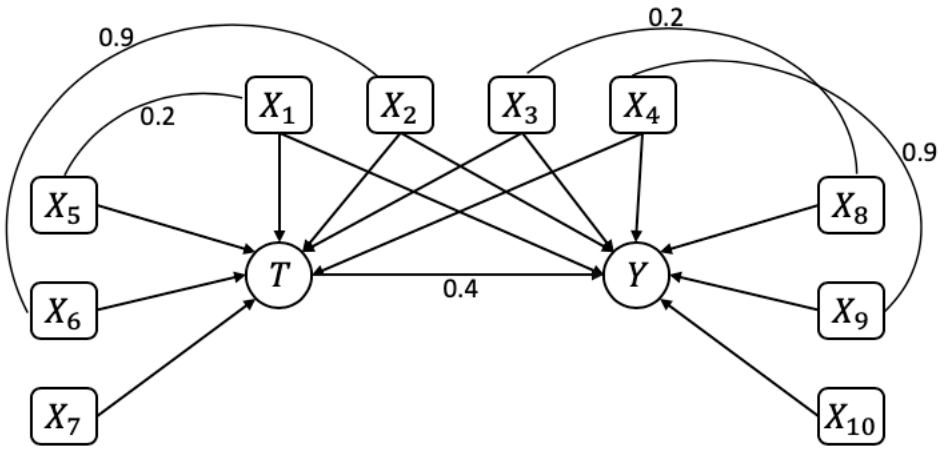
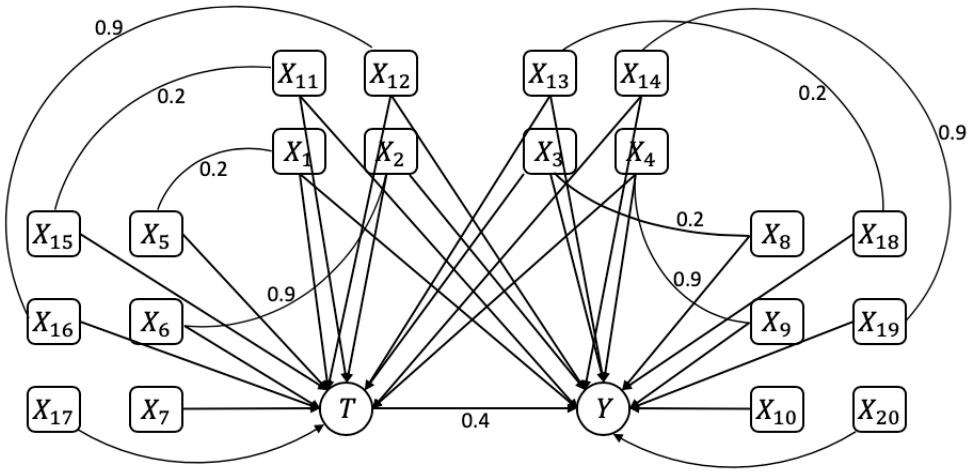


Figure 11

Simulation structure for conditions with 20 covariates



2. Generate the treatment assignment. The true PS was generated using a logistic regression $P(T = 1|X) = (1 + \exp(-\beta f(X))^{-1}$, where f(X) represents the true PS models (Table 2). Then each true PS was compared with a value which was randomly drawn from a uniform distribution between 0 and 1 to obtain the true treatment assignment. If the true PS was larger than the random value, the subject was in the treatment condition otherwise in the control condition. The intercept and coefficients of the covariates were manipulated to achieve the specified nonlinearity degree and make the ratio of treatment versus control group size about 1:1, which represented a common practice in some previous studies (Lee et al., 2010: Setoguchi et al., 2008). Detailed information about the values of the intercept and the coefficients of the covariates can be found in Appendix B.

Table 2

The True Propensity Score Models

Number of Covariates	Model	Functional Form
10	M	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$
	Q	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{Q2} X_2^2$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{Q2} X_2^2 + \beta_{Q4} X_4^2 + \beta_{Q7} X_7^2$
	I	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{I12} X_1 X_2$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{I12} X_1 X_2 + \beta_{I34} X_3 X_4 + \beta_{I67} X_6 X_7$
	E	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{E2} \exp(X_2)$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{E2} \exp(X_2) + \beta_{E4} \exp(X_4) + \beta_{E7} \exp(X_7)$
20	M	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17}$
	Q	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{Q2} X_2^2 + \beta_{Q12} X_{12}^2$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{Q2} X_2^2 + \beta_{Q4} X_4^2 + \beta_{Q7} X_7^2 + \beta_{Q12} X_{12}^2 + \beta_{Q14} X_{14}^2 + \beta_{Q17} X_{17}^2$
	I	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{I12} X_1 X_2 + \beta_{I1112} X_{11} X_{12}$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{I12} X_1 X_2 + \beta_{I34} X_3 X_4 + \beta_{I67} X_6 X_7 + \beta_{I1112} X_{11} X_{12} + \beta_{I1314} X_{13} X_{14} + \beta_{I1617} X_{16} X_{17}$
	E	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{E2} \exp(X_2) + \beta_{E12} \exp(X_{12})$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_{E2} \exp(X_2) + \beta_{E4} \exp(X_4) + \beta_{E7} \exp(X_7) + \beta_{E12} \exp(X_{12}) + \beta_{E14} \exp(X_{14}) + \beta_{E17} \exp(X_{17})$

40	M	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37}$
	Q	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{Q2} X_2^2 + \beta_{Q12} X_{12}^2 + \beta_{Q22} X_{22}^2 + \beta_{Q32} X_{32}^2$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{Q2} X_2^2 + \beta_{Q4} X_4^2 + \beta_{Q7} X_7^2 + \beta_{Q12} X_{12}^2 + \beta_{Q14} X_{14}^2 + \beta_{Q17} X_{17}^2 + \beta_{Q22} X_{22}^2 + \beta_{Q24} X_{24}^2 + \beta_{Q37} X_{37}^2 + \beta_{Q42} X_{42}^2 + \beta_{Q44} X_{44}^2 + \beta_{Q47} X_{47}^2$
	I	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{I12} X_1 X_2 + \beta_{I1112} X_{11} X_{12} + \beta_{I2122} X_{21} X_{22} + \beta_{I3132} X_{31} X_{32}$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{I12} X_1 X_2 + \beta_{I34} X_3 X_4 + \beta_{I67} X_6 X_7 + \beta_{I1112} X_{11} X_{12} + \beta_{I1314} X_{13} X_{14} + \beta_{I1617} X_{16} X_{17} + \beta_{I2122} X_{21} X_{22} + \beta_{I2324} X_{23} X_{24} + \beta_{I2627} X_{26} X_{27} + \beta_{I3132} X_{31} X_{32} + \beta_{I3334} X_{13} X_{14} + \beta_{I3637} X_{36} X_{37}$
	E	$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{E2} \exp(X_2) + \beta_{E12} \exp(X_{12}) + \beta_{E22} \exp(X_{22}) + \beta_{E32} \exp(X_{32})$
		$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \beta_4 X_{14} + \beta_5 X_{15} + \beta_6 X_{16} + \beta_7 X_{17} + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \beta_4 X_{24} + \beta_5 X_{25} + \beta_6 X_{26} + \beta_7 X_{27} + \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} + \beta_4 X_{34} + \beta_5 X_{35} + \beta_6 X_{36} + \beta_7 X_{37} + \beta_{E2} \exp(X_2) + \beta_{E4} \exp(X_4) + \beta_{E7} \exp(X_7) + \beta_{E12} \exp(X_{12}) + \beta_{E14} \exp(X_{14}) + \beta_{E17} \exp(X_{17}) + \beta_{E22} \exp(X_{22}) + \beta_{E24} \exp(X_{24}) + \beta_{E27} \exp(X_{27}) + \beta_{E32} \exp(X_{32}) + \beta_{E34} \exp(X_{34}) + \beta_{E37} \exp(X_{37})$

3. Generate the outcome value. The continuous outcome Y was generated using a linear combination of the treatment assignment T and the covariates that were associated with the outcome such that $Y = \sum \alpha_p X_p - 0.4T$, where the true treatment effect was -0.4. Details about the true outcome model can be found in Table 3. The values of the regression coefficients in the outcome model are as follows: $\alpha_0 = -3.85, \alpha_1 = 0.3, \alpha_2 = -0.36, \alpha_3 = -0.73, \alpha_4 = -0.2, \alpha_5 = 0.71, \alpha_6 = -0.19, \alpha_7 = 0.26$.

Table 3

The True Outcome Models

Number of covariates	Functional Form
10	$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} - 0.4T$
20	$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} + \alpha_1 X_{11} + \alpha_2 X_{12} + \alpha_3 X_{13} + \alpha_4 X_{14} + \alpha_5 X_{18} + \alpha_6 X_{19} + \alpha_7 X_{20} - 0.4T$
40	$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} + \alpha_1 X_{11} + \alpha_2 X_{12} + \alpha_3 X_{13} + \alpha_4 X_{14} + \alpha_5 X_{18} + \alpha_6 X_{19} + \alpha_7 X_{20} + \alpha_1 X_{21} + \alpha_2 X_{22} + \alpha_3 X_{23} + \alpha_4 X_{24} + \alpha_5 X_{28} + \alpha_6 X_{29} + \alpha_7 X_{30} + \alpha_1 X_{31} + \alpha_2 X_{32} + \alpha_3 X_{33} + \alpha_4 X_{34} + \alpha_5 X_{38} + \alpha_6 X_{39} + \alpha_7 X_{40} - 0.4T$

After generating the data, the quality of the generated data was checked through examining the distribution for each of the covariates, the correlations between the covariates, the ratio of treated subjects versus control subjects, the nonlinearity degree of the true PS models, and the expected difference between the treatment and control groups conditional on the covariates. Results indicated that they were aligned with the specified values. Specifically, the continuous covariates followed the standard normal distribution, the correlations between the

specified correlated variables were around 0.2 or 0.9 though, for some cases, the correlations were attenuated a little due to categorizing the covariates, the ratio of treated subjects versus control subjects was about 1:1, the nonlinearity degree was aligned with the specified nonlinearity degree, and the expected difference between the treatment and control group was about -0.4 conditional on the covariates that were related to the outcome. All of these indicated that the data were generated appropriately.

3.3 Estimation Process

After generating the data, I estimated the PSs using logistic regression models and random forests models with different hyperparameters (as defined in the within-cell conditions). As mentioned earlier, for logistic regression, three models were used including a main effects only model, a correctly specified logistic regression model, and a logistic regression model with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates. For random forests, the R package randomForest was used with different combinations of hyperparameters. Specifically, the minimum terminal node size was set to 1 subject, 5% and 10% of the sample size; the number of covariates to select from at each split was set to p, Sqrt(p), and log2p+1, where p was the total number of the covariates; the subject sampling method included sampling with replacement with the original sample size, sampling without replacement with 0.632 times the original sample size, and sampling without replacement with 0.5 times the original sample size. Thus, in total there were 27 (3*3*3) random forests models. Then the weights of the control subjects were adjusted using the weighting method of weighting by the odds (Equation 11), which was also used by Lee, Lessler, and Stuart (2010). Finally, the treatment effect (ATT) was calculated using the weighted average difference in the outcome between the treatment and control groups (Equation 14).

3.4 Performance Metrics

The performance of the various models was evaluated using the following measures.

Average of the absolute standardized mean difference (ASMD) on the covariates. ASMD between the treatment and control groups on each of the covariates was calculated using Equation 12. Then the average value of the ASMD was calculated across all covariates and across replications. A lower ASMD indicates that on average the treatment and control groups are more similar on the covariates. Theoretically, better balance between the treatment and control groups will lead to a less biased estimate of the treatment effect.

Relative bias in the estimate of the treatment effect. This is a direct measure of the accuracy of the estimates. It was calculated as follows:

$$\theta_{bias} = \frac{1}{200}\sum_{r=1}^{200}\left(\hat{\theta}_r - \theta_{true}\right), \tag{18}$$

where θ_{bias} is the bias of the estimated treatment effect across all of the 200 replications, $\widehat{\theta_r}$ is the estimated treatment effect for the rth replication, and θ_{true} is the true treatment effect, which is -0.4 in the study.

Empirical variance of the estimates of the treatment effect. This is a measure of the precision of the estimates. It was calculated as follows:

$$var\left(\hat{\theta}\right) = \frac{1}{200}\sum_{r=1}^{200}\left(\hat{\theta}_r - \bar{\hat{\theta}}\right)^2, \tag{19}$$

where $var\left(\hat{\theta}\right)$ is the empirical variance of the treatment effect estimates across all of the 200 replications, $\widehat{\theta_r}$ is the estimated treatment effect for the rth replication, and $\bar{\hat{\theta}}$ is the average of the estimated treatment effects across all of the 200 replications. A small variance indicates better reliability of the model.

In addition to the point estimates of ASMD, average bias, and average precision, distribution of ASMD and bias for each model was checked using boxplots, where a small variance indicates better reliability of the model. Using all the above metrics, I first compared the estimations of random forests with different hyperparameter values versus the three logistic regression models to see if random forest models lead to better covariate balance and provide more accurate and reliable estimate of the treatment effect than logistic regression models. Then I made between-cell comparisons to see whether the trend was consistent across different between-cell conditions. Besides that, I also made within-cell comparisons to investigate the impact of random forests hyper-parameter values on the estimates of the treatment effect and identified the optimal choice of hyperparameter values. In addition, I examined the impacts of the interactions between two between-cell conditions, between two within-cell conditions, and across between-cell and within-cell conditions. In addition to checking the summary statistics, I also conducted ANOVA tests to see whether the impacts of these factors were significant on the comparative performance of random forests over logistic regression.

Chapter 4 Simulation Study Results

To answer the primary research questions, the simulation study results are presented in three major sections: covariate balance between treatment conditions, relative bias of the estimated treatment effect, and empirical variance in the estimates of the treatment effect.

4.1 Covariate balance between treatment conditions.

Covariate balance was measured using the average of the absolute standardized mean difference (ASMD). I checked ASMD on all covariates across all between-cell conditions and replications. In addition to the commonly used average ASMD, I also used median ASMD across between-cell conditions and replications as the ASMD distribution was skewed due to outliers and the median ASMD was less sensitive to those outliers (Figure 12). Table 4 shows that, before PS weighting, the covariates were unbalanced between the treatment and control groups as the average ASMD (0.280) and the median ASMD (0.269) were greater than the commonly used covariate balance threshold of 0.25. After PS weighting, in general, covariate balance was achieved for all the examined PS models including main effects only logistic regression (lr_main), logistic regression with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates (lr_largest_ASMD), correctly specified logistic regression (lr_correctly_specified), and random forests(rf). This was supported by the fact that the average ASMD and median ASMD were smaller than the covariate balance threshold of 0.25 for each of the examined PS models. Among all the examined PS models, it seems that lr_main performed best in balancing covariates as it had the smallest average ASMD (0.061) and the smallest median ASMD (0.042), which was followed by lr_largest_ASMD (average ASMD 0.069, median ASMD 0.042) and lr_correctly_specified (average ASMD 0.100, median ASMD 0.056), and then random forests without hyper-parameter tuning (average ASMD 0.247, median

ASMD 0.181). Selecting appropriate hyper-parameter settings could improve the performance of random forests in balancing covariates (rf_best average ASMD 0.142, median ASMD 0.145, where rf_best represented the smallest ASMD provided by random forest models for each between-cell condition and each replication), but the performance was still not as good as logistic regression models. To identify what hyper-parameter settings led to better covariate balance, I conducted ANOVA tests and post-hoc pairwise analysis. Results showed that, in general, minimum terminal node size of 1 led to better covariate balance compared with other settings of minimum terminal node size. For the hyper-parameter of number of covariates per split (mtry), sqrt(p) seemed to be a better choice in terms of balancing covariates while for sampling method there was no salient difference between different sampling methods (Table 5). In addition, I also checked the ASMD distribution for each examined PS model. Figure 12 shows that the ASMDs of rf_best and lr_main were much less dispersed with fewer outliers than those from other methods.

In summary, using PS weighting, both logistic regression and random forests could achieve satisfying performance on balancing covariates and, on average, logistic regression performed better than random forest approaches. The performance of random forests could be improved by selecting appropriate hyper-parameter settings, such as using a minimum terminal node size of 1 and using sqrt(p) as number of covariates per split, but, on average, logistic regression models, especially the main effects only logistic regression model, still performed better on balancing covariates. Although logistic regression performed better than random forests on balancing covariates, this did not necessarily yield more reliable and valid estimates of the treatment effects. The results in the section below address this issue.

Table 4

ASMD before and after PS weighting

Weighting	Model Type	Mean	Median
Before	NA	0.280	0.269
After	lr_main	0.061	0.042
	lr_largest_asdm	0.069	0.042
	lr_correctly_specified	0.100	0.056
	rf	0.247	0.181
	rf_best	0.142	0.145

Table 5

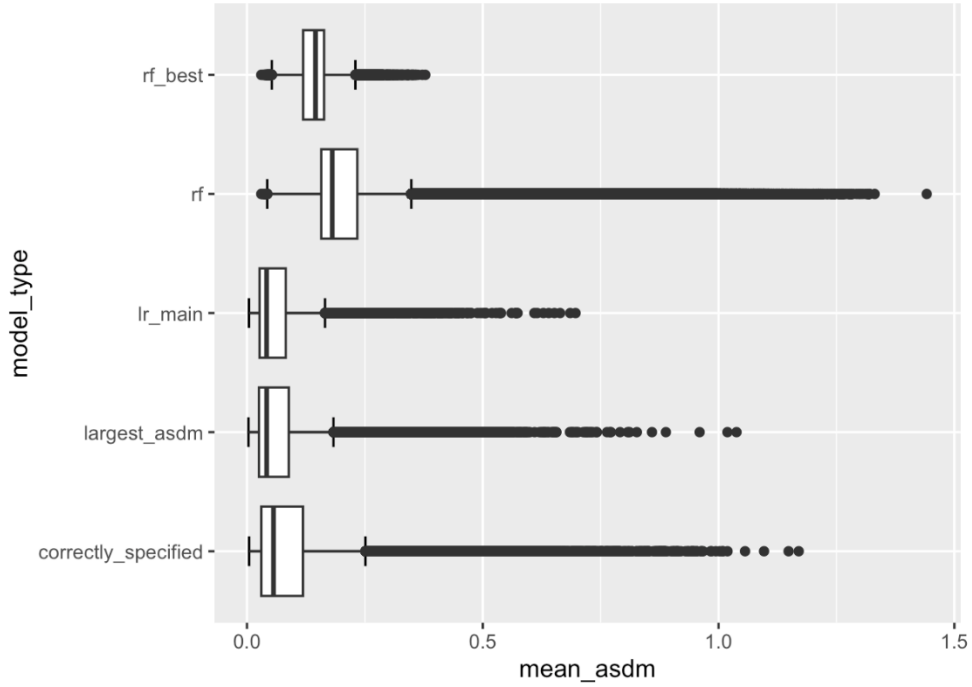
Pairwise t-test of ASMD by rf hyper-parameter

RF Hyper-parameter	Group 1	Group 2	Group 1 Mean	Group 2 Mean	P-value	Effect Size (Cohen's d)
minimum terminal node size	0.05*n	0.1*n	0.269	0.289	<0.001	0.098
	0.05*n	1	0.269	**0.183**	<0.001	0.585
	0.1*n	1	0.289	**0.183**	<0.001	0.691
mtry	log2(p)+1	p	0.231	0.297	<0.001	0.360
	log2(p)+1	sqrt(p)	0.231	**0.212**	<0.001	0.116
	p	sqrt(p)	0.297	**0.212**	<0.001	0.479
sampling method	w/o replace 0.5*n	w/o replace 0.632*n	0.245	0.252	<0.001	0.038
	w/o replace 0.5*n	with replace n	0.245	0.243	<0.001	0.012
	w/o replace 0.632*n	with replace n	0.252	0.243	<0.001	0.049

Note. P-values in the table are Bonferroni adjusted. Bold values represent the comparatively better performance (smaller ASMD) between the two groups.

Figure 12

Distribution of the average ASMD across replications by model type



4.2 Relative bias of the estimated treatment effects

The performance of logistic regression and random forests models on estimating the treatment effect was evaluated from two aspects, accuracy and reliability. For accuracy, it was measured using relative bias of the estimated treatment effect (TE) (Equation 18). The section below presents the summary statistics of overall relative bias, relative bias by between-cell conditions, relative bias by within-cell conditions, and hypothesis testing results about factors that significantly influence the comparative performance of random forests over logistic regression on accurately estimating the TE.

4.2.1 Summary Statistics of Relative Bias

Overall relative bias. For each examined PS model, I calculated the average and median absolute relative bias across conditions and replications. Results showed that, without hyper-parameter tuning, the average and median absolute relative biases of random forests (average absolute relative bias 52.2%, median absolute relative bias 36.5%) were much larger than those of the logistic regression models (Table 6). However, by selecting appropriate hyper-parameter settings, the accuracy of the treatment effect estimates by random forests improved substantially. And the performance of the best performing random forests model (rf_best, average of the absolute relative bias 14.8%, median absolute relative bias 7.1%, where rf_best represented the smallest absolute relative TE bias by random forests for each between-cell condition and each condition) was even better than the best performing logistic regression model lr_main (average absolute relative bias 17.7%, median absolute relative bias 10.8%). And the variance of the absolute relative TE bias was also greatly reduced through using appropriate hyper-parameter settings (Figure 12). To further examine under what conditions with what hyper-parameter setting random forests model was more likely to perform better than logistic regression, I also checked the absolute relative TE bias by between-cell condition and by within-cell condition. More details are provided in the following section.

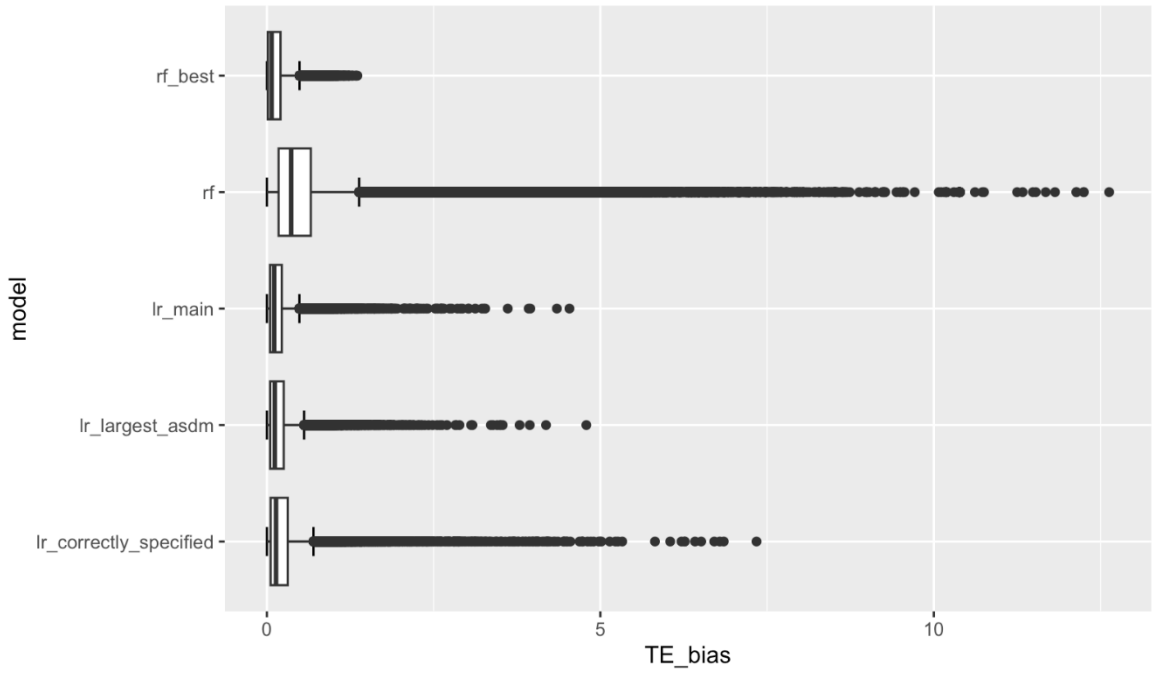
Table 6

Summary statistics of the relative bias (in percentage) on the estimates of the treatment effects

Model	Mean	Median
lr_main	17.7	10.8
lr_largest_asdm	19.7	11.8
lr_correctly_specified	26.8	13.7
rf	52.2	36.5
rf_best	14.8	7.1

Figure 13

Treatment effect absolute relative bias distribution by model



Relative bias by between-cell condition. As discussed in section 3.1, in total, there were 171 between-cell conditions. For each condition at each replication, I compared the least biased TE estimate by random forests (rf_best) to the least biased TE estimate from logistic regression models and found that, among the 34,200 cases, about 48% rf_best produced less relative bias

than any logistic regression model and, for the remaining 52% cases, logistic regression produced less relative bias than rf_best. Although, in general, there were more conditions where logistic regression performed better than random forests, it did not necessarily mean logistic regression was superior to random forests in providing more accurate TE estimates for the following two reasons. First, a detailed analysis of the TE relative bias for different levels of between-cell factors indicated that for some conditions, for example the conditions of 20 covariates, 20 subjects per covariate, nonlinearity type of exponential and high-level nonlinearity degree, rf_best had a higher chance than logistic regression models to provide less biased TE estimates (Table 7). Second, even for the 52% cases where logistic regression produced less biased TE estimates than rf_best, it is still possible for random forests to perform better than logistic regression with other hyper-parameter settings that were not examined in this study due to time limitation and restriction in computing resources. For example, for the condition of 10 covariates, 200 subjects per covariate, with 1 interaction term and medium-level nonlinearity degree, using the examined hyper-parameter settings rf_best had an average absolute relative bias of 39.1% across 200 replications, which was much larger than the average absolute relative bias of 14.7% for the best performing logistic regression model. However, for the same condition, in an investigation using post-hoc simulation, using another set of hyper-parameters, minimum terminal node size of 0.3*n, mtry of sqrt(p), sampling without replacement, sampling size of 0.5*n, the average relative bias of random forests could reach about 8.7%, much smaller than the counterpart of any logistic regression model. This indicates that the choice of hyper-parameter settings in random forests is extremely important. In the following paragraph, I talk more about under what hyper-parameter settings, random forests model is more likely to perform better than logistic regression on TE estimation, based on conditions examined in the simulation.

Table 7

Percentage of conditions when lowest TE bias by random forest < lowest TE bias logistic regression models, by manipulated level of between-cell condition

	Number of covariates			
	10	20	40	
Number of conditions	11400	11400	11400	
Percentage of conditions when lowest rf TE bias < lowest lr TE bias	48.3%	**56.9%**	38.9%	
	Number of subjects per covariate			
	20	200	500	
Number of conditions	11400	11400	11400	
Percentage of conditions when lowest rf TE bias < lowest lr TE bias	**60.6%**	45.8%	37.7%	
	Nonlinearity type			
	main effects only	quadratic	interaction	exponential
Number of conditions	1800	10800	10800	10800
Percentage of conditions when lowest rf TE bias < lowest lr TE bias	42.7%	43.9%	49.4%	**51.7%**
	Nonlinearity degree			
	no	low	medium	high
Number of conditions	1800	10800	10800	10800
Percentage of conditions when lowest rf TE bias < lowest lr TE bias	42.7%	37.3%	43.5%	**64.1%**
	Number of nonlinear terms			
	0	1	3	
Number of conditions	1800	16200	16200	
Percentage of conditions when lowest rf TE bias < lowest lr TE bias	42.7%	47.5%	49.1%	

Note. Rf represents random forests; lr represents logistic regression. Percentages greater than 50% are highlighted in bold.

Relative bias by within-cell condition. For each of the 171 between-cell conditions, 27 sets of hyper-parameters (3 levels for minimum terminal node size, 3 levels of mtry which is the number of covariates from which to select at each split, and 3 sampling methods) were examined for random forests. For each between-cell condition with each set of hyper-parameter settings we had 200 replications. Thus, in total, there were 923,400 TE estimates using random forests methods. Among these 923,400 TE estimates, only 132,213 of them had comparatively smaller relative bias than any logistic regression model. This indicated that, without appropriate selection of the hyper-parameters, it was not very likely for random forests to produce less biased TE estimate than logistic regression models. To identify under what hyper-parameter settings random forests was more likely to produce less biased TE estimates than logistic regression models, I checked the percentage of conditions when random forests produced a smaller relative bias than any logistic regression model for each level of within-cell factors. Unfortunately, results indicated that, among the hyper-parameters that I examined, there was not a single setting that led to an over 50% chance for random forests to perform better than logistic regression models (Table 8). However, the results did tell us that some hyper-parameters were not very likely to yield less biased TE estimates, for example, the minimum terminal node size of 1 and mtry of p. Using terminal node size of 1, there was only a 10.8% chance for random forests to produce less biased TE estimates than logistic regression models. The chance was much smaller than using other values as minimum terminal node size, such as 0.05*n and 0.1*n. Similarly, using mtry of p, there was only a 12.4% chance for random forests to produce less biased TE estimates than logistic regression models. The chance was much smaller than using sqrt(p) or log2(p)+1 as the number of covariates per split (Table 8).

Table 8

Percentage of conditions when the TE bias by random forests < the lowest TE bias by logistic regression models, by manipulated level of within-cell condition

	Minimum terminal node size		
	1	0.05*n	0.1*n
Percentage of conditions	10.8%	16.1%	16.1%
	Number of covariates per split		
	sqrt(p)	log2(p)+1	P
Percentage of conditions	15.7%	14.9%	12.4%
	Sampling method		
	with replace, n	w/o replace, 0.5*n	w/o replace, 0.632*n
Percentage of conditions	14.3%	14.7%	14.0%

Note. The total number of conditions for each cell is 307,800.

4.2.2 Hypothesis Testing of Relative Bias

The above summary statistics of the overall relative TE bias, and relative TE bias by between-cell and within-cell conditions indicated the potential impacts of some factors on the comparative performance of random forests over logistic regression models, to further examine whether the impacts were statistically significant and whether there were some interactions between these factors, I conducted two ANOVA tests: one for the impacts of these factors without any hyper-parameter tuning, and the other one for the impacts of these factors with hyper-parameter tuning, as well as some post-hoc analysis for each test. For the ANOVA test without any hyper-parameter tuning, the outcome was the difference in the relative TE bias by each random forests model with each set of hyper-parameters as compared to the best performing logistic regression model (absolute relative TE bias by each random forests model – lowest absolute relative TE bias by any logistic regression model), while for the ANOVA test with hyper-parameter tuning, the outcome was the difference in the relative TE bias by the best

performing random forests model as compared to the best performing logistic regression model (lowest absolute relative TE bias by random forests– lowest absolute relative TE bias by any logistic regression model). Results indicated that without hyper-parameter tuning, in general, the TE estimates by random forests were not comparable to that of the best performing logistic regression model, though the difference in the outcome differed by some between-cell factors, within-cell factors, and across between-cell and within-cell factors. However, with hyper-parameter tuning, for some conditions, such as high-level of nonlinearity degree and 20 subjects per covariate, random forests produced significantly less biased TE estimates than any logistic regression model. More details about the results are discussed in the following section.

4.2.2.1 ANOVA Test: Without Hyper-parameter Tuning

For the ANOVA test without hyper-parameter tuning, for each of the 171 between-cell conditions and each of the 200 replications, I calculated the relative TE bias by random forests with each of the 27 hyper-parameter settings, and then subtracted the relative TE bias by the corresponding (for the same between-cell condition and the same replication) best performing logistic regression model and used the difference as the outcome. Thus, in total, the analysis is based on 923,400 data points (171 between-cell conditions * 200 replications * 27 hyper-parameter settings).

Main effects of between-cell factors. Results show that all of the examined between-cell factors, including number of covariates (num_covariate), number of subjects per covariate (sub_cov_ratio), nonlinearity type (nonlinearity_type), nonlinearity degree (nonlinearity_degree) and number of nonlinear terms (nonlinearity_term), were significant predictors of the comparative performance of random forests over logistic regression (Table 9). Among these influential between-cell factors, sub_cov_ratio was most impactful with the largest effect size

(η^2: 0.099). Post-hoc analysis results showed that, sub_cov_ratio of 20 had the least biased TE estimates, which was followed by sub_cov_ratio of 200 and then by sub_cov_ratio of 500; the mean outcome for each condition was significantly different from the others (Table 10). This indicated that results from using a random forests model were more comparable to logistic regression under conditions with a smaller number of subjects per covariate. A possible reason for the less desirable performance at a larger number of subjects per covariate was that, for this condition, it was more likely to have complicated tree structures, and the model became more likely to overfit and led to extreme PS estimates, which might lead to biased TE estimates. After sub_cov_ratio, the 2nd most influential between-cell factor is num_covariate (η^2: 0.016). Results showed that, compared with num_covariate of 10 and 40, the random forests model produced significantly less biased TE estimates for num_covariate of 20 (Table 10). A possible reason for the difference was that, for num_covariate of 10, the tree structure might not be sophisticated enough and the model might be underfit while for num_covariate of 40 the tree structure might be too complicated and the model might be overfit which might lead to extreme PS estimates and biased TE estimates. The 3rd influential between-cell factor was nonlinearity_degree (η^2: 0.012). Aligned with the expectation, results showed that it was more likely for random forests to have more comparable performance as logistic regression at a higher level of nonlinearity degree. This was indicated by the statistically significant less biased TE estimates at the high level of nonlinearity degree when compared with no, low and medium levels of nonlinearity degree (Table 10). For the remaining two factors, nonlinearity_type and nonlinearity_term, although they were also significant predictors of the comparative performance of random forests over logistic regression, their impacts were very small which was indicated by the very small effect

sizes (nonlinearity_type, η^2: 0.002; nonlinearity_term, η^2: < 0.001). Thus, I do not dive deep into the post-hoc pairwise t-test results for these two factors.

Table 9

Mixed ANOVA test result of factors influencing the comparative performance of random forests over logistic regression

Type of Effect	Effect	F	p	Effect Size (η^2)
Main Effects: Between-cell Factors	sub_cov_ratio	50948.38	<0.001	0.099
	num_covariate	7469.528	<0.001	0.016
	nonlinearity_degree	5477.987	<0.001	0.012
	nonlinearity_type	549.818	<0.001	0.002
	nonlinearity_term	20.68	<0.001	<0.001
Main Effects: Within-cell Factors	mtry	15477.801	<0.001	0.032
	node_size	14079.712	<0.001	0.030
	sampling_method	118.616	<0.001	<0.001
Interactions: Between-cell Factors	num_covariate * nonlinearity_degree	1433.689	<0.001	0.009
	sub_cov_ratio * num_covariate	1731.946	<0.001	0.007
	nonlinearity_degree * nonlinearity_type	1567.459	<0.001	0.007
	num_covariate * nonlinearity_type	897.923	<0.001	0.004
	sub_cov_ratio * nonlinearity_degree	452.158	<0.001	0.003
	sub_cov_ratio * nonlinearity_type	389.205	<0.001	0.002
	nonlinearity_degree * nonlinearity_term	106.214	<0.001	<0.001
	sub_cov_ratio * nonlinearity_term	12.15	<0.001	<0.001
	num_covariate * nonlinearity_term	6.128	0.002	<0.001
Interactions: Within-cell Factors	mtry * node_size	4705.406	<0.001	0.020
	node_size * sampling_method	53.21	<0.001	<0.001
	mtry * sampling_method	16.721	<0.001	<0.001
	mtry * node_size * sampling_method	7.267	<0.001	<0.001
Interactions: Between-cell & Within-cell Factors	sub_cov_ratio * node_size	10330.601	<0.001	0.043
	num_covariate * node_size	5401.314	<0.001	0.023
	sub_cov_ratio * mtry	3979.87	<0.001	0.017
	num_covariate * mtry	2099.953	<0.001	0.009
	nonlinearity_type * mtry	422.959	<0.001	0.002
	nonlinearity_degree * node_size	220.179	<0.001	0.001
	nonlinearity_type * node_size	162.319	<0.001	<0.001
	sub_cov_ratio * sampling_method	119.086	<0.001	<0.001
	nonlinearity_degree * mtry	67.37	<0.001	<0.001
	num_covariate * sampling_method	79.483	<0.001	<0.001
	nonlinearity_degree * sampling_method	11.071	<0.001	<0.001
	nonlinearity_term * mtry	24.372	<0.001	<0.001
	nonlinearity_type * sampling_method	4.293	0.002	<0.001

Table 10

Means and standard deviations of the differences in relative TE bias between random forests and logistic regression, by between-cell factor

sub_cov_ratio			
20	200	500	
0.164_a	0.453_b	0.595_c	
(0.308)	(0.659)	(0.749)	
num_covariate			
10	20	40	
0.398_a	0.324_b	0.492_c	
(0.593)	(0.656)	(0.625)	
nonlinearity_degree			
no	low	medium	high
0.448_a	0.444_a	0.422_b	0.318_c
(0.587)	(0.628)	(0.632)	(0.624)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

Main effects of within-cell factors. Regarding the impact of hyper-settings on TE estimates, results showed that all three factors yielded significantly different outcomes across conditions, yet the impact of sampling_method was very small ($\eta^2 < 0.001$) (Table 9). The most influential within-cell factor was mtry which was the number of covariates from which to select at each split. Results showed that, on average, random forests produced relatively less biased TE estimates when using mtry of sqrt(p) as compared with using mtry of log2p+1 or p (Table 11). This indicated that, in general, the default setting of sqrt(p) was a good choice for mtry. A possible reason for the less desirable performance when using mtry of p was that, for this condition, at each split the covariate was selected from the pool of all covariates which might lead to the tree being dominated by some highly influential factors and making some other

influential factors have small chances if no chance to be taken into account. This might lead to the tree and the corresponding random forests model to be underfit and have poor performance in PS estimates, thus influencing the TE estimates. The comparatively less desirable performance when using mtry of log2p+1 was possibly due to the same reason. Specifically, log2p+1 was greater than sqrt(p) which enabled the splitting covariate to be selected among a larger pool of covariates, thus reducing the chances of some non-dominating but also influential factors to be selected. Consequently, the random forests model might be underfit and had poor performance in PS estimates and TE estimates. For the 2nd influential factor, node_size, it seemed that, in general, node_size of 1 was a good choice as, on average, random forests had relatively less bias on TE estimates when compared with using node_size of 0.05*n and 0.1n (Table 11). However, the trend differed by some other between and within-cell factors which indicated the existence of some interaction effects. The section below discusses these as well as other interactions in detail.

Table 11

Means and standard deviations of the differences in relative TE bias between random forests and logistic regression, by within-cell factor

	mtry	
sqrt(p)	log2p+1	p
0.312_a	0.359_b	0.542_c
(0.489)	(0.572)	(0.769)
	node_size	
1	0.05*n	0.1*n
0.272_a	0.458_b	0.484_c
(0.273)	(0.741)	(0.732)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

Interactions. Results showed that there were multiple significant interaction effects (Table 9). In the following discussion, I primarily focus on those with at least a small effect size ($\eta^2 > 0.01$) as these interaction effects were more influential than others. The most influential interaction effect was the interaction of sub_cov_ratio and node_size (η^2: 0.043). Post-hoc analysis results showed that the optimal setting of node_size differed for different values of sub_cov_ratio. For sub_cov_ratio of 20, node_sizes of 0.05*n and 0.1*n seemed to be better choices than node_size of 1 as they enabled random forests to have more comparable TE estimates to logistic regression. However, for sub_cov_ratio of 200 and 500, it was the other way around. For these conditions, node_size of 1 seemed to be a better choice as it allowed random forests to have more comparable TE estimates to logistic regression (Table 12). A possible reason for the poor performance of using node_size of 0.05*n or 0.1*n at sub_cov_ratio of 200 and 500 was that, for these conditions, with more subjects per covariate the sample size

increased, setting node_size as 0.05*n or 0.1*n led to larger terminal node sizes, which might make the random forests model underfit and have poor performance in PS estimates and TE estimates. Besides sub_cov_ratio, the impact of node_size on the comparative performance of random forests over logistic regression also differed by num_covariate (η^2: 0.023). At num_covariate of 10 and 20, node_size of 1 seemed to be a better setting than node_size of 0.05*n or 0.1*n as it enabled random forests have more comparable TE estimates to logistic regression. However, for num_covariate of 40, using node_size of 1 did not lead to much improvement on TE estimates as compared with using node_size of 0.05*n or 0.1*n (Table 13). A possible reason for the diminishment of the advantages of using node_size of 1 at num_covariate of 40 was that, for this condition, using node_size of 1 might lead to complicated tree structure, and the random forests model might overfit, producing extreme PS estimates and thus biased TE estimates. In addition, there was also an interaction between node_size and mtry (η^2: 0.020). In other words, the optimal number of covariates from which to select at each split differed by different values of the minimum terminal node size. Specifically, when node_size of 0.05*n or node_size of 0.1*n was used, setting mtry as sqrt(p) or log2p+1 made random forests have relatively less biased TE estimated as compared with using mtry of p. However, when node_size was set as 1, the difference was very small for different settings of mtry (Table 14). This was possibly because when node_size was set as 1 the tree structure usually had more levels and more splits, creating more chances to take account the effect of each covariate. As such, the choice of mtry did not matter substantially. Besides the above interaction effects, there was also an interaction between sub_cov_ratio and mtry. Specifically, for sub_cov_ratio of 200 and sub_cov_ratio of 500, setting mtry as sqrt(p) or log2p+1 led to less biased TE estimated as compared with using mtry of p. However, for sub_cov_ratio of 10, the difference between

different settings of mtry became much smaller (Table 15). A possible reason was that when sub_cov_ratio was 10, the sample size was small, and the tree structure was comparatively simple, thus regardless of what values we used for mtry the tree structure would not be very different. As a consequence, the PS estimates would not differ substantially and the TE estimates would be similar.

Table 12

*Means and standard deviations of the differences in relative TE bias between random forests and logistic regression, by sub_cov_ratio * node_size*

sub_cov_ratio	node_size		
	1	0.05*n	0.1*n
20	0.235_a	0.120_b	0.138_c
	(0.328)	(0.282)	(0.300)
200	0.285_a	0.469_b	0.606_c
	(0.245)	(0.713)	(0.827)
500	0.295_a	0.784_b	0.706_c
	(0.231)	(0.916)	(0.807)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

Table 13

*Means and standard deviations of the differences in relative TE bias between random forests and logistic regression, by num_covariate * node_size*

num_covariate	node_size		
	1	0.05*n	0.1*n
10	0.128_a	0.500_b	0.565_c
	(0.144)	(0.654)	(0.704)
20	0.191_a	0.418_b	0.363_c
	(0.177)	(0.819)	(0.749)
40	0.497_a	0.456_b	0.523_c
	(0.306)	(0.738)	(0.728)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

Table 14

*Means and standard deviations of the differences in relative TE bias between random forests and logistic regression, by node_size * mtry*

node_size	mtry		
	sqrt(p)	log2p+1	p
1	0.280_a	0.277_a	0.258_b
	(0.280)	(0.275)	(0.263)
0.05*n	0.327_a	0.388_b	0.658_c
	(0.575)	(0.661)	(0.904)
0.1*n	0.331_a	0.412_b	0.708_c
	(0.553)	(0.677)	(0.875)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

Table 15

*Means and standard deviations of the differences in relative TE bias between random forest sand logistic regression, by sub_cov_ratio * mtry*

sub_cov_ratio	mtry		
	sqrt(p)	log2p+1	p
20	0.160_a	0.160_a	0.174_b
	(0.302)	(0.302)	(0.320)
200	0.303_a	0.357_b	0.701_c
	(0.371)	(0.467)	(0.925)
500	0.475_a	0.561_b	0.750_c
	(0.662)	(0.770)	(0.782)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference.

From the above results, we can see that without hyper-parameter tuning, in general, the performance of the random forests model was not comparable to that of logistic regression, which was indicated by the positive difference in the relative bias of TE estimates between random forests and logistic regression across conditions (Table 10-14). However, this did not mean that random forest approaches with appropriate settings of the hyper-parameters did not have a chance to beat logistic regression. The section below discusses the comparative performance of random forests with appropriate hyper-parameters against logistic regression.

4.2.2.2 ANOVA Test: With Hyper-parameter Tuning

As discussed above, to examine the comparative performance of random forests with appropriate hyper-parameters against logistic regression, I conducted another ANOVA test. Different from the 1st ANOVA test in which I compared the performance of each random forests model with each set of hyper-parameters with the best performing logistic regression model (with the lowest absolute relative TE bias), in this ANOVA test, for each between-cell condition

and each replication, I selected the lowest absolute relative TE bias by random forests models (rf_best) and compared it with the lowest absolute relative TE bias by logistic regression models (lr_best). In this way, we made a fairer comparison of the best performance between random forests and logistic regression. For this ANOVA test, in total, we had 32,400 data points (171 between-cell conditions * 200 replications).

Similar to the results of the ANOVA test without hyper-parameter tuning, results of this ANOVA test indicated that, all the between-cell factors were significant predictors of the comparative performance of random forests over logistic regression, yet the impact of nonlinearity_term was very small ($\eta^2 < 0.001$)) (Table 16). However, unlike the previous ANOVA test and the corresponding pairwise t-tests in which, on average, logistic regression consistently performed better than random forest, from this ANOVA test and the corresponding pairwise t-tests we found that for some conditions, such as 20 covariates, high-level nonlinearity degree, and 20 subjects per covariate, on average, the relative TE bias from the best performing random forests model was actually smaller than that of the best performing logistic regression model (Table 17). This indicated that if appropriate hyper-parameters were being used it was likely for random forests to produce less biased TEs than logistic regression when there was a moderate number of covariates (e.g., 20 covariates), or when a high-level of nonlinearity existed, or when there were a smaller number of subjects per covariate. Interaction effects were also found between these factors. For example, the impact of nonlinearity degree level on the comparative performance of random forests over logistic regression differed for different numbers of covariates. Specifically, for the conditions of 20 covariates, the lowest absolute relative TE bias by random forests was consistently smaller than that of logistic regression for all levels of nonlinearity degree (Table 18). However, for the conditions of 10 covariates and 40

covariates, the best performance of random forests only beat the best performance of logistic regression when a high-level of nonlinearity existed. This was possibly because the examined hyper-parameter settings worked better for the conditions of 20 covariates as compared with the conditions of 10 covariates and 40 covariates. In addition, there was also an interaction between num_covariate and sub_cov_ratio. For the conditions of 20 covariates, rf_best beat lr_best for sub_cov_ratio of 20 and sub_cov_ratio of 200 while for the conditions of 10 covariates rf_best was better than lr_best only when sub_cov_ratio was 20. And when there were 40 covariates, lr_best always beat rf_best regardless of sub_cov_ratio (Table 19). The comparative better performance of rf_best over lr_best for these specific num_covariate and sub_cov_ratio conditions was possibly because the examined hyper-parameter settings worked better for such conditions. What's more, there was also an interaction between num_covariate and nonlinearity_type. Specifically, for the conditions of 20 covariates, regardless of nonlinearity type, the lowest absolute relative TE bias by random forests was consistently lower than that of logistic regression while for the conditions of 10 covariates rf_best beat lr_best only for the nonlinearity type of exponential. And for the conditions of 40 covariates, rf_best did not beat lr_best for any of the nonlinearity types (Table 20). The comparative better performance of rf_best over lr_best for the nonlinearity type of exponential at conditions of 10 covariates and 20 covariates indicated that with appropriate hyper-parameter settings random forests could provide more accurate TE estimates than logistic regression for other types of nonlinearity (such as exponential) besides nonlinearity types of interaction and quadratic which was not covered by previous studies (Lee et al., 2010).

Table 16

Mixed ANOVA test result of factors influencing the comparative performance of the best performing random forests over the best performing logistic regression

Type of Effect	Effect	F	p	Effect Size (η^2)
Main Effects	num_covariate	1619.501	<0.001	0.087
	nonlinearity_degree	1478.169	<0.001	0.080
	sub_cov_ratio	520.56	<0.001	0.030
	nonlinearity_type	115.756	<0.001	0.010
	nonlinearity_term	0.958	0.328	<0.001
Interactions	num_covariate * nonlinearity_degree	229.933	<0.001	0.021
	sub_cov_ratio * num_covariate	166.602	<0.001	0.019
	num_covariate * nonlinearity_type	122.332	<0.001	0.014
	nonlinearity_degree * nonlinearity_type	70.27	<0.001	0.008
	sub_cov_ratio * nonlinearity_degree	42.903	<0.001	0.007
	sub_cov_ratio * nonlinearity_type	9.795	<0.001	0.001
	num_covariate * nonlinearity_term	16.134	<0.001	<0.001
	nonlinearity_degree * nonlinearity_term	13.759	<0.001	<0.001
	sub_cov_ratio * nonlinearity_term	3.787	0.023	<0.001

Table 17

Means and standard deviations of the differences in relative TE bias between the best performing random forests and the best performing logistic regression, by between-cell factor

num_covariate			
10	20	40	
0.007	**-0.021**	0.106	
(0.125)	(0.155)	(0.287)	
nonlinearity_degree			
no	low	medium	high
0.088	0.083	0.045	**-0.045**
(0.257)	(0.232)	(0.204)	(0.151)
sub_cov_ratio			
20	200	500	
-0.013	0.047	0.058	
(0.273)	(0.180)	(0.147)	
nonlinearity_type			
main effects only	quadratic	interaction	exponential
0.088	0.045	0.018	0.02
(0.257)	(0.187)	(0.214)	(0.214)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference. Bold values represent the best performing random forests have smaller relative TE bias than the best performing logistic regression.

Table 18

*Means and standard deviations of the differences in relative TE bias between the best performing random forests and the best performing logistic regression, by num_covariate * nonlinearity_degree*

num_covariate	nonlinearity_degree			
	no	low	medium	high
10	0.028_{a}	0.018_{b}	0.004_{b}	**-0.024_{c}**
	(0.106)	(0.123)	(0.140)	(0.107)
20	**-0.001_{a}**	**-0.022_{b}**	**-0.036_{c}**	**-0.046_{c}**
	(0.157)	(0.151)	(0.163)	(0.147)
40	0.305_{a}	0.244_{b}	0.117_{c}	**-0.075_{d}**
	(0.300)	(0.289)	(0.265)	(0.185)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference. Bold values represent the conditions where the best performing random forests have smaller relative TE bias than the best performing logistic regression.

Table 19

*Means and standard deviations of the differences in relative TE bias between the best performing random forests and the best performing logistic regression, by num_covariate * sub_cov_ratio*

num_covariate	sub_cov_ratio		
	20	200	500
10	**-0.063$_a$**	0.034$_b$	0.049$_c$
	(0.141)	(0.109)	(0.090)
20	**-0.083$_a$**	**-0.010$_b$**	0.030$_c$
	(0.206)	(0.108)	(0.105)
40	0.107$_{ab}$	0.118$_a$	0.094$_b$
	(0.373)	(0.255)	(0.208)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference. Bold values represent the conditions where the best performing random forests have smaller relative TE bias than the best performing logistic regression.

Table 20

*Means and standard deviations of the differences in relative TE bias between the best performing random forests and the best performing logistic regression, by num_covariate * nonlinearity_type*

num_covariate	nonlinearity_type			
	main effects only	quadratic	interaction	exponential
10	0.028_a (0.106)	0.023_b (0.118)	0.016_b (0.157)	**-0.018_c** (0.090)
20	**-0.001_a** (0.157)	**-0.003_a** (0.143)	**-0.014_b** (0.163)	**-0.042_c** (0.154)
40	0.305_a (0.300)	0.121_b (0.248)	0.044_c (0.291)	0.120_b (0.300)

Note. Standard deviations are in parentheses. Means that do not share subscripts differ by $p < 0.05$ according to Tukey's Honestly Significant Difference. Bold values represent the conditions where the best performing random forests have smaller relative TE bias than the best performing logistic regression.

In summary, from the above analysis, we can see that, in general, logistic regression provided less biased TE estimates than random forests without hyper-parameter tuning. However, when appropriate hyper-parameters were used, for some conditions, such as when there was a moderate number of covariates, when high-level nonlinearity existed, or when there was a smaller number of subjects per covariate, on average, random forests produced less biased TE estimates than logistic regression. Regarding the impact of the hyper-parameters on the TE estimates, in general the choice of the sampling method did not substantially influence the accuracy of the TE estimates using random forests and the default setting of mtry, sqrt(p), was a good choice for most conditions. However, for the minimum terminal node size, the optimal setting differed for different conditions. For a larger number of subjects per covariate (e.g., 200

or 500), the default setting of 1 was a good choice: leading to less biased TE estimates. However, when there was a smaller number of subjects per covariate, other recommended settings of the minimum terminal node size, such as 0.05*n or 0.1*n, were better choices as they lead to less biased TE estimates.

4.3 Empirical variance of the estimated treatment effects

In addition to examining the accuracy of the TE estimates, I also checked the reliability of the TE estimates. To make the findings generalizable to other studies which have TEs on different scales, I used the empirical standard deviation of the relative bias of the TE estimates instead of the empirical standard deviation of the TE estimates to measure how reliable the TE estimate is. The section below presents the reliability results for each model, for random forests with different hyper-settings, for different between-cell conditions, and by between-cell and within-cell conditions.

Reliability of the TE estimates for each model. From the distribution of the relative TE bias for each model we can see that in general the TE estimates by the random forests model were spread out over a wider range, indicating that the TE estimates by the random forests model were less reliable than those of logistic regression models (Figure 14). The empirical standard deviation of the relative bias of the TE estimates by the random forests model was much larger than that of any of the logistic regression models, which also indicated the TE estimates by random forests were less reliable (Table 21). This finding was not surprising as the TE estimates by random forests included the results from all of the 27 hyper-settings and as discussed earlier the TE estimates from different hyper-settings could be very different.

Figure 14

Distribution of relative TE bias by model

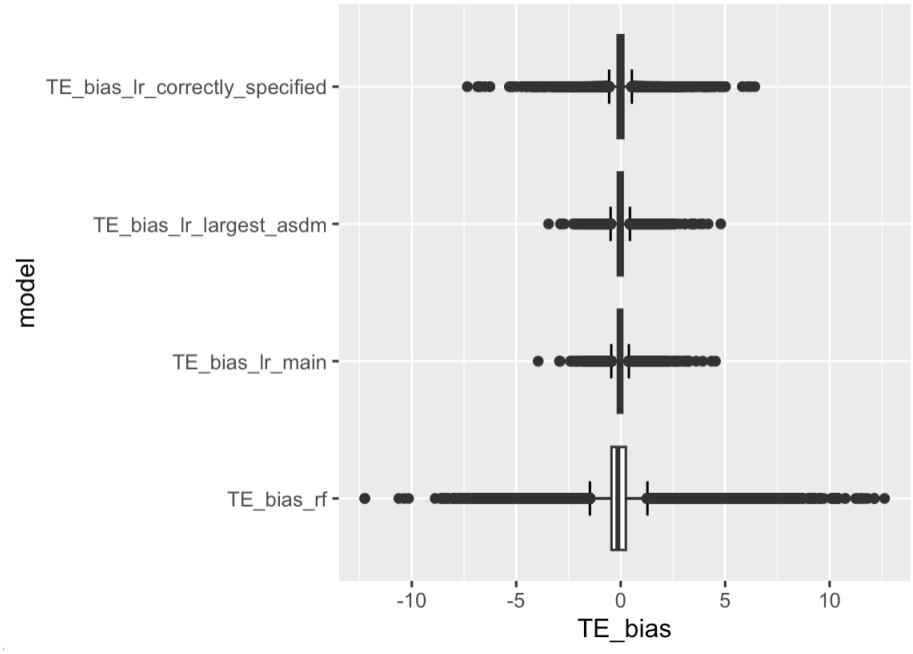


Table 21

Empirical standard deviation of the relative bias of the TE estimates by model

Model	Empirical Standard Deviation
lr_main	0.285
lr_largest_asmd	0.314
lr_correctly_specified	0.489
rf	0.802

Reliability of the TE estimates for random forests with different hyper-settings. To further examine the reliability of the TE estimates for random forests with different hyper-settings, I checked the empirical standard deviation of the relative bias of the TE estimates for each hyper-parameter. Results showed that node_size of 1 had a smaller empirical standard deviation than other settings of node_size, which indicated that in general node_size of 1 led to

more reliable TE estimates (Table 22). A possible reason for the more reliable TE estimates at node_size 1 was that with this hyper-setting (only one subject in each terminal node) the PS of each subject was less influenced by other subjects (in random forests PS is calculated as the percentage of treated subjects in the terminal node) thus the PS estimates were more stable from one replication to another which led to more reliable TE estimates. In addition, results also showed that for mtry, the number of covariates from which to select at each split, settings of sqrt(p) and log2p+1 had smaller empirical standard deviations than the setting of p. This was possibly because using p as mtry enabled the selection of the best covariate (has the largest amount of entropy reduction) among the whole set of covariates at each split which might lead the tree structure and the corresponding random forests model to overfit. And usually when a model overfits, the estimates have larger variance. For sampling_method, there was no salient difference between different sampling methods.

Table 22

Empirical standard deviation of the relative bias of the TE estimates for random forests, by hyper-setting

Node_size	Empirical Standard Deviation
1	0.338
0.05*n	0.904
0.1*n	0.918
Mtry	**Empirical Standard Deviation**
sqrt(p)	0.632
log2p+1	0.728
p	0.993
Sampling_method	**Empirical Standard Deviation**
with replace, n	0.797
w/o replace, 0.5*n	0.793
w/o replace, 0.632*n	0.815

Reliability of the TE estimates for each model by between-cell condition. To examine whether the reliability of the TE estimates differed for different between-cell conditions, I checked the empirical standard deviation of the relative bias of the TE estimates for each model and each between-cell condition. Results showed that for the between-cell factor of num_covariate, as the number of covariates increased the TE estimates by logistic regression models had larger empirical standard deviations while for random forests the TE estimates had comparable empirical standard deviations for different numbers of covariates (Table 23). This indicated that the reliability of the TE estimates by random forests was consistent for different numbers of covariates while for logistic regression models the reliability decreased with increasing number of covariates. For the between-cell factor of sub_cov_ratio, a very interesting finding was that for logistic regression models the empirical standard deviation of the relative bias of the TE estimates became smaller with increasing number of subjects per covariate while for random forests it was the other way around: the empirical standard deviation of the relative bias of the TE estimates got larger when we had more subjects per covariate (Table 24). This indicated that, for logistic regression models, the reliability of the TE estimates improved with more subjects per covariate while for random forests the reliability dropped with more subjects per covariate. This was because for logistic regression models when we had more subjects per covariate, we had a larger sample size, and with a larger sample size usually the model became less likely to overfit thus the variance of the TE estimates became smaller. For random forests, the larger variance of the TE estimates at larger numbers of subjects per covariate was mainly from the hyper-settings of node_size of 0.05*n and node_size of 0.1*n. More details will be presented later when we discuss the interaction between num_covariate and node_size. For the

between-cell factor of nonlinearity_type, results showed that for logistic regression models, especially for lr_correctly_specified and lr_largest_asmd, the empirical standard deviation of the relative bias of the TE estimates was larger for the nonlinearity type of interaction compared with other nonlinearity types (Table 25). A possible reason was that for the nonlinearity type of interaction we had more extreme nonlinearity degree: the high-level nonlinearity degree of interaction was 0.27, larger than the high-level nonlinearity degree of other nonlinearity types. With a higher level of nonlinearity degree, the model could be more complex. And usually when a model became more complex, the estimates had larger variance (Mehta et al., 2019). However, for random forests, we did not find a big difference in the empirical standard deviation of the relative bias of the TE estimates for different nonlinearity types, which indicated that the reliability of the TE estimates was comparable for different types of nonlinear relations. For the between-cell factor of nonlinearity_degree, results showed that, for random forests, the reliability was similar for different levels of nonlinearity degree which was indicated by the comparable empirical standard deviation of the relative bias of the TE estimates across different levels of nonlinearity degree (Table 26). Yet, for logistic regression models, especially lr_correctly_specified and lr_largest_asmd, the empirical standard deviation of the relative bias of the TE estimates was larger for the high level of nonlinearity degree. A possible reason was that for a higher level of nonlinearity degree the model became more complex which led to a larger variance in the estimates. Finally, for the between-cell factor of nonlinearity_term, results showed that for random forests the empirical standard deviation of the relative bias of the TE estimates was similar for different numbers of nonlinearity term, which indicated that the reliability of the TE estimates was comparable across different numbers of nonlinearity term (Table 27). For lr_main and lr_largest_asmd the empirical standard deviation of the relative bias

of the TE estimates was also comparable for different numbers of nonlinearity term while for lr_correctly_specified the empirical standard deviation of the relative bias of the TE estimates got larger with increasing number of nonlinearity term. A possible reason was that as the number of nonlinearity terms increased the total number of terms also increased in lr_correctly_specified and the model became more complex. As discussed earlier, when a model became more complex, it had larger variance.

Table 23

Empirical standard deviation of the relative bias of the TE estimates for each model, by num_covariate

Model	Num_covariate	Empirical Standard Deviation
lr_main	10	0.178
	20	0.248
	40	0.386
lr_largest_asmd	10	0.259
	20	0.299
	40	0.370
lr_correctly_specified	10	0.300
	20	0.424
	40	0.667
rf	10	0.708
	20	0.757
	40	0.784

Table 24

Empirical standard deviation of the relative bias of the TE estimates for each model, by sub_cov_ratio

Model	Sub_cov_ratio	Empirical Standard Deviation
lr_main	20	0.446
	200	0.170
	500	0.123
lr_largest_asmd	20	0.468
	200	0.210
	500	0.172
lr_correctly_specified	20	0.696
	200	0.376
	500	0.302
rf	20	0.399
	200	0.850
	500	0.979

Table 25

Empirical standard deviation of the relative bias of the TE estimates for each model, by nonlinearity_type

Model	Nonlinearity_type	Empirical Standard Deviation
lr_main	main	0.303
	quadratic	0.236
	interaction	0.310
	exponential	0.295
lr_largest_asmd	main	0.264
	quadratic	0.261
	interaction	0.390
	exponential	0.277
lr_correctly_specified	main	0.303
	quadratic	0.435
	interaction	0.664
	exponential	0.335
rf	main	0.770
	quadratic	0.825
	interaction	0.767
	exponential	0.815

Table 26

Empirical standard deviation of the relative bias of the TE estimates for each model, by nonlinearity_degree

Model	Nonlinearity_degree	Empirical Standard Deviation
lr_main	no	0.303
	low	0.284
	medium	0.288
	high	0.272
lr_largest_asmd	no	0.264
	low	0.275
	medium	0.298
	high	0.362
lr_correctly_specified	no	0.303
	low	0.332
	medium	0.418
	high	0.676
rf	no	0.770
	low	0.821
	medium	0.834
	high	0.735

Table 27

Empirical standard deviation of the relative bias of the TE estimates for each model, By nonlinearity_term

Model	Nonlinearity_term	Empirical Standard Deviation
lr_main	0	0.303
	1	0.275
	3	0.291
lr_largest_asmd	0	0.264
	1	0.313
	3	0.320
lr_correctly_specified	0	0.303
	1	0.463
	3	0.530
rf	0	0.770
	1	0.805
	3	0.800

Reliability of the TE estimate for random forests by both between-cell factor and hyper-setting. To examine whether the impact of the random forest hyper-settings on the reliability of the TE estimates differed for different between-cell conditions, I checked the empirical standard deviation of the relative bias of the TE estimates for the combination of each between-cell factor and each hyper-parameter (interaction effects of between-cell factors and hyper-settings). Except for the interaction effects of sub_cov_ratio * node_size and num_covariate * mtry, for all the other combinations of between-cell factors and hyper-parameters the impact of the hyper-setting on the reliability of the TE estimates was comparable for different levels of the between-cell factor. For sub_cov_ratio * node_size, we can see that for node_size of 1 the empirical standard deviation of the relative bias of the TE estimates was consistent for different numbers of sub_cov_ratio while for node_size settings of 0.05*n and 0.1*n the empirical standard deviation became larger with increasing number of sub_cov_ratio

(Table 28). This indicated that for node_size of 1 the TE estimates were quite stable regardless of the number of subjects per covariate while for node_sizes of 0.05*n and 0.1*n the TE estimates became less reliable when we had more subjects per covariate. As discussed earlier, a possible reason for the smaller variance of the TE estimates at node_size 1 was that, with this hyper-setting, the PS of each subject was less influenced by other subjects thus the PS estimates were more stable which led to more reliable TE estimates. The consistent small variance of the TE estimates for node_size 1 at different numbers of sub_cov_ratio indicated that the impact of node_size 1 on the reliability of the TE estimates was robust to different numbers of subjects per covariate. For node_sizes of 0.05*n and 0.1*n, the TE estimates became less reliable with increasing number of sub_cov_ratio because with these hyper-settings as the sub_cov_ratio increased the sample size also increased and with a larger sample size the terminal node size became larger which meant that the PS estimate of each subject would be impacted by more subjects. In other words, it meant that the PS estimates would become less reliable thus the TE estimates would become less reliable. For num_covariate * mtry, an interesting finding was that for mtry settings of sqrt(p) and log2p+1 the empirical standard deviation of the relative bias of the TE estimates became smaller with increasing number of covariates while for mtry of p it was the opposite: the empirical standard deviation got larger when we had more covariates (Table 29). This indicated that for mtry settings of sqrt(p) and log2p+1 the reliability of the TE estimates improved with increasing number of covariates while for mtry of p the reliability of the TE estimates dropped with increasing number of covariates. A possible reason was that for mtry of p when the number of covariates increased the model structure became more complex which led to a larger variance of the PS estimates and a larger variance of the TE estimates. For mtry settings of sqrt(p) and log2p+1, as the number of covariates increased at each split the candidate

covariate pool size did not increase proportionally. For example, when the number of covariates increased from 10 to 20, at each split the candidate covariate pool size just increased from 3 to 4 for mtry of sqrt(p), from 4 to 5 for mtry of log2p+1. Since the candidate covariate pool size did not increase proportionally, more randomness was brought into the model, making the model less likely to overfit, which led to a smaller variance of the PS estimates and a smaller variance of the TE estimates.

Table 28

*Empirical standard deviation of the relative bias of the TE estimates for random forest, by sub_cov_ratio * node_size*

Sub_cov_ratio	Node_size	Empirical Standard Deviation
20	1	0.393
200	1	0.310
500	1	0.299
20	0.05*n	0.358
200	0.05*n	0.881
500	0.05*n	1.140
20	0.1*n	0.412
200	0.1*n	1.054
500	0.1*n	1.059

Table 29

*Empirical standard deviation of the relative bias of the TE estimates for random forest, by num_covariate * mtry*

Num_covariate	Mtry	Empirical Standard Deviation
10	sqrt(p)	0.660
20	sqrt(p)	0.561
40	sqrt(p)	0.417
10	log2p+1	0.718
20	log2p+1	0.711
40	log2p+1	0.524
10	p	0.744
20	p	0.938
40	p	1.167

In summary, we can see that, in general, the TE estimate by random forests was not as reliable as that of logistic regression models. Yet, for some hyper-settings, such as node_size of 1, the reliability of the TE estimate by random forests could be comparable to that of logistic regression models. And the reliability of the TE estimate by random forests differed for different numbers of subjects per covariate with different hyper-settings. When node_size was 1, the reliability of the TE estimate was similar for different numbers of subjects per covariates. However, for other settings of node_size, such as 0.05*n and 0.1*n, the TE estimate was more reliable for a smaller number of subjects per covariate. For mtry settings of sqrt(p) and log2p+1 the TE estimate was more reliable with increasing number of covariates while for mtry of p the reliability of the TE estimate dropped with increasing number of covariates.

Chapter 5. Empirical Illustration

In addition to the simulation study described in Chapters 3 and 4, this dissertation also includes a real data illustration which is used as a tutorial about when to use which method, logistic regression or random forests, to estimate the PSs, and how to choose the hyperparameters in random forests. Data from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K:1998-1999) is used to assess the effects of kindergarten retention on students' reading and math abilities.

5.1 Data

The ECLS-K data were released by the U.S. National Center for Education Statistics (NCES). The data file contains repeated observations of a nationally representative sample of students, their parents, teachers and schools. The students were followed from kindergarten to 8th grade. In this study, I focused on kindergarten and the 1st grade, from Fall 1998 to Spring 2000. This analysis was based on a previous study by Hong and Raudenbush (2005) about the potential effects of kindergarten retention on students' reading and math scores. The treatment is whether a student was retained at kindergarten in the academic year of 1999-2000. The outcomes are the student's estimated reading and math abilities which were calibrated by IRT test scores in Spring 2000. Different from Hong and Raudenbush's study in which missing values were imputed, in this study, without knowledge about how the missing imputation was conducted in the study by Hong and Raudenbush and also for the sake of simplicity, the analysis was limited to students who have complete data on the treatment, the outcomes, and also the covariates. The analytic sample therefore contained 771 students. Among all the students, 28 were retained at kindergarten in 1999-2000 while 743 were promoted to first grade.

5.2 Methods for Empirical Illustration

5.2.1 Variable selection

Considering the covariates that influence the treatment assignment, I used the same set of covariates as was used in Hong and Raudenbush's study (2005). They identified 207 pretreatment covariates that were significantly associated with kindergarten retention, including students' demographic and psychological characteristics as well as family, classroom, and school characteristics. Through stepwise logistic regression, they finally included 38 covariates and the quadratic terms for 7 of the 38 covariates in the PS model. The descriptions of the covariates are presented in Table 30. In this study, I followed the practice of Hong and Raudenbush and included the 38 covariates in the PS models. In general, for any study, if there is no preliminary analysis about the influential covariates, we can leverage random forests to select the covariates. Basically, in random forests the importance of a covariate is measured by the difference in the model fit with and without this covariate (Breiman, 2001; Hothorn et al., 2006). This variable importance measure by random forests has been suggested for relevant covariate selection in multiple fields (Bureau et al., 2005; Chen et al., 2020; Lunetta et al., 2004).

Table 30

Selected Variables (ECLS-K) for the Data Analysis

	variable (recoded)	time	type	original variable in the data and description
treatment	REPEAT_K	spring 2000	categorical	retention/promotion status at the spring semester of 2000 (T4GLVL)
outcome	READ_WAVE4	spring 2000	continuous	reading IRT score at the spring semester of 2000 (C4R4RSCL)
	MATH_WAVE4	spring 2000	continuous	math IRT score at the spring semester of 2000 (C4R4MSCL)
predictors	AGE_MONTHS	fall 1998	continuous	student's age (in months) at kindergarten entry (R1_KAGE)
	GENDER	fall 1998	categorical	student's gender (GENDER); 2 levels, male, female
	READ_BOOK_FREQ	fall 1998	categorical	student frequently reading books outside school (P1CHREAD); 4 levels, never, 1-2 times/week, 3-6 times/week, everyday
	PERCENT_REPEAT_K	fall 1998	continuous	number of students in class repeating kindergarten (A1REPK) / total class enrollment (A1TOTAG)
	PERCENT_BOYS (and its quadratic term)	fall 1998	continuous	number of boys in class (A1BOYS) / total class enrollment (A1TOTAG)
	ORIENTATION	fall 1998	categorical	teacher giving parents orientation at the beginning of that year (B1PRNTOR); 2 levels, yes, no
	CLASS_HOURS_PER_WEEK	fall 1998	continuous	number of class hours per day (A1HRSDA) * number of instructional days per week (A1DYSWK)
	CLASS_ACTIVITY_LEVEL	fall 1998	categorical	time on teacher-directed whole class activity (B1WHLCLS); 4 levels, <= 0.5 h, about 1 h, about 2 h, >= 3 h
	TEACHER_EXP_YEARS	fall 1998	continuous	number of years teacher taught preschool (B1YRSPRE)
	READ_WAVE2 (and its quadratic term)	spring 1999	continuous	reading IRT score at the spring semester of 1999 (C2R4RSCL)
	MATH_RATING (and its quadratic term)	spring 1999	categorical	teacher rating of student's math skills (T2RTMTH); 5 levels, far below average, below average, average, above average, far above average
	GENERAL_IRT	spring 1999	continuous	general knowledge IRT score (C2RGSCAL)
	PROBLEM_BEHAVIOR_SCORE	spring 1999	continuous	teacher rating of student's externalizing problem behavior (1- 4) (T2EXTERN)
	MATH_ARS	spring 1999	continuous	teacher report of student's math ARS score (1 - 5) (T2RARSMA)

HEALTH_FALL_BEHIND	spring 1999	categorical	student falling behind due to health (T2FLBHND); 2 levels, yes, no
SCIENCE_SOCIAL_SKILLS	spring 1999	categorical	teacher rating of student's science/social skills (T2RTSCI); 5 levels, far below average, below average, average, above average, far above average
PARENT_MEETING_HOURS	spring 1999	continuous	school administrator hours/week meeting with parents (0-40) (S2TALKPT)
TEACHER_HOME_VISIT	spring 1999	categorical	teacher seeing homes before the school year begin (B2HMEVST); 2 levels, yes, no
NUM_READ_GROUPS	spring 1999	continuous	number of achievement groups in reading (0-5) (T2NORDGP)
LOWEST_READ_GROUP	spring 1999	categorical	student in the lowest reading group (T2CHRDGP = 1)
COMPUTER_ADEQUACY	spring 1999	categorical	class has adequate computer equipment (A2COMPEQ); 5 levels
MINORITY_LEVEL	spring 1999	categorical	percent of minority students (S2KMINOR); 5 levels, < 10, 10 - 25, 25 - 50, 50 - 75, >=75
HAS_GRADE_SEVEN	spring 1999	categorical	having grade level beyond the sixth grade (S27TH); 2 levels, yes, no
NUM_ESL_TEACHERS	spring 1999	continuous	school number of FTE (full-time equivalent) ESL-bilingual teachers (0-4) (S2ESLFTE)
REPORT_CARDS_FREQ (and its quadratic term)	spring 1999	categorical	school's frequency of report cards (S2RPRTCD); 5 levels, never, 1/year, 2-3 times/year, 4-6 times/year, >=7 times/ year
PARENT_PERFORM_FREQ (and its quadratic term)	spring 1999	categorical	school's frequency of performance to which parents are invited (S2INVITE); 5 levels, never, 1/year, 2-3 times/year, 4-6 times/year, >=7 times/ year
CLASSROOM_PROGRAM_FREQ (and its quadratic term)	spring 1999	categorical	school's frequency of classroom programs (S2CLASPR); 5 levels, never, 1/year, 2-3 times/year, 4-6 times/year, >=7 times/ year
PARENT_TEACHER_MEET_FREQ	spring 1999	categorical	school's frequency of PTA/PTO (parent teacher organization) meetings (S2PTAMT); 5 levels, never, 1/year, 2-3 times/year, 4-6 times/year, >=7 times/ year
PARENT_TEACHER_CONFERENCE_FREQ	spring 1999	categorical	school's frequency of parent-teacher conferences (S2PTCONF); 5 levels, never, 1/year, 2-3 times/year, 4-6 times/year, >=7 times/ year
TITLE1	spring 1999	categorical	school received funding for Title I professional development (S2TT1PD); 2 levels, yes, no
SCHOOL_HIGHEST_SALARY_LEVEL	spring 1999	categorical	school highest annual base salary (S2HISLRY); 5 levels, <25K, 25K-35K, 35K - 45K, 45K-60K, >=60K
NUM_PRINCIPAL_COURSES	spring 1999	continuous	number of courses in early childhood (S2CRSECE) and child development (S2CRSCDV) taken by the principal

PRINCIPAL_EVAL_LEVEL	spring 1999	categorical	influence of staff support evaluation on principal performance evaluation (S2STFSPP)
EMPHASIS_LANGUAGE_NUMBER_SKILLS (and its quadratic term)	spring 1999	categorical	school's emphasis on goals of language and number skills (S2GOAL1); 3 levels, no/minor emphasis, moderate, major
SCHOOL_HISPANIC_LEVEL	spring 1999	categorical	school percent of Hispanic students (S2PCTHSP); 5 levels, 0, 0-5, 5-10, 10-25, >=25
REQUIRE_ACADEMIC_RECORD	spring 1999	categorical	school required academic record (S2ACADRC); 2 levels, yes, no
SECURITY_GUARDS	spring 1999	categorical	school with security guards (S2SECURT); 2 levels, yes, no
ATTITUDE_ON_HALLWAYS	spring 1999	categorical	Attitude on school having decorated hallways (K2Q6_A); 4 levels, from strongly agree to strongly disagree

5.2.2 Data Preparation

Since results from the simulation study indicate that the existence of a high-level of nonlinearity degree influences the comparative performance of random forests over logistic regression, to determine whether there is a need to use random forests, before running the analysis I checked the existence of nonlinearity for this dataset. Note that, different from the simulation study in which the nonlinearity degree was measured as the average of the absolute difference between the estimated PS based on the best fitting main effects only logistic regression model and the true PSs, in this empirical study, the true PSs were unknown thus there was no way to measure the nonlinearity degree. Instead, I checked the existence of nonlinearity through comparing the model fit of the main effects only logistic regression model with other logistic regression models that contained different types and different numbers of nonlinear terms. Results indicated the model fit of the logistic regression models with nonlinear terms was not statistically better than that of the main effects only logistic regression model, thus I thought that this dataset did not contain a high-level of nonlinearity.

5.2.3 Analysis

The effect of kindergarten retention on students' reading and math scores was examined using the five-step PS procedure. After identifying the covariates, the PS scores were estimated. Regarding which model to use to estimate the PSs, here as a demonstration, to present the differences in the estimate of the treatment effect using different models in a more comprehensive way, I used almost all of the models that were explored in the simulation study (except the correctly specified logistic regression model as it was unknown and thus unavailable in the empirical study). Regarding what hyperparameters to use in random forests, as a demonstration, I used all the 27 hyper-settings that were examined in the simulation study.

Specifically, I used the minimum leaf node size of 1, 39 (5% of the sample size), and 77 (10% of the sample size). The number of covariates to select from at each split was set to 38 (all the covariates), 6 (sqrt(p), where p is the total number of covariates), and 7 (log2p+1, where p is the total number of covariates). For each tree in random forests, the subjects were selected using three sampling methods, sampling with replacement with the original sample size of 771, sampling without replacement with a size of 487 (0.632 of the original sample size), and sampling without replacement with a size of 386 (0.5 of the original sample size). In summary, in addition to the two logistic regression models (main effects only logistic regression, logistic regression with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates), 27 (3*3*3) random forests models were used to estimate the PSs. Then subjects were weighted based on the estimated PSs using Equation 11. Next, balance on the covariates was checked using ASMD on the set of 38 covariates. Finally, the treatment effect of kindergarten retention was estimated using the weighted group mean difference on reading and math test scores separately.

Unlike in a simulation study, in an empirical study the true treatment effect is unknown, thus there is no way to calculate the bias of the estimates of the treatment effect to tell which PS model performs better. Instead, I compared the estimates of the treatment effect using logistic regression models and random forests models with different hyperparameters to see if the estimates were very different from each other. In addition, I also checked the covariate balance using different models as theoretically better balance on the covariates usually leads to a more accurate estimate of the treatment effect.

5.3 Results

Results indicated that, before adjustment, the covariates were quite unbalanced between the treatment (students who were retained at kindergarten) group and the control group (students who were promoted to the 1st grade): the average ASMD was 0.53. And the differences between the two groups on reading and math scores were pretty large: the average reading score of the treatment group was 29.86 lower than that of the control group; the average math score of the treatment group was 22.25 lower than of the control group. Yet, after using PS weighting (via the main effects only logistic regression model) to adjust the differences in the covariates between the treatment and control groups, the covariate balance became much better: ASMD reached to 0.23, meeting the commonly used criterion of covariate balance. And the differences in the reading and math scores between the two groups became much smaller: the average weighted group mean difference in reading was reduced to -8.64 (subtracting the weighted average reading score of the control group from that of the treatment group) and the average weighted group mean difference in math was reduced to -4.02 (subtracting the weighted average math score of the control group from that of the treatment group) (Table 31). This indicated that after adjusting the differences in the covariates, on average, the reading score of the students who were retained at kindergarten was only 8.64 lower than the students who were promoted to the 1st grade and the average math score of the retained students was only 4.02 lower than their counterparts who were promoted to the 1st grade. The results using the logistic regression model with the quadratic term of the most unbalanced covariate and its interaction with all the other covariates were highly similar to the results using the main effects only logistic regression model with some minor differences from the 5th decimal place.

For random forests, the results varied substantially when different hyper-parameters were used. Specifically, the average ASMD ranged from 0.25 to 0.34, the average group mean difference in reading ranged from -13.46 to -4.82 (subtracting the weighted average reading score of the control group from that of the treatment group), and the average group mean difference in math ranged from -8.95 to -2.35 (subtracting the weighted average math score of the control group from that of the treatment group). With regard to what hyper-parameters to use, based on the results of the simulation study, I would recommend using this hyper-setting: node_size of 0.05*n, mtry of sqrt(p), sampling without replacement, sampling_size of 0.632*n because the dataset of this empirical study had a small number of subjects per covariate (about 20) and according to the results of the simulation study for datasets with a small number of subjects per covariate node_size of 0.05*n was more likely to provide a less biased TE estimate than the default node_size setting of 1 and also the node_size of 0.1*n. For mtry, I would recommend using sqrt(p) as, in general, it was a good choice based on the results of the simulation study. For the sampling method, the simulation study found no salient difference using different sampling methods. In this study, I would recommend using sampling without replacement with sampling_size of 0.632*n as it had the best covariate balance when compared with the other two sampling methods (Table 32).

With regard to whether we should use random forests or logistic regression models, for this dataset, I would prefer using logistic regression models, or more accurately the main effects only logistic regression model. This is because results of the simulation study indicated that for datasets with a large number of covariates and without high-level nonlinearity degree the main effects only logistic regression model was more likely to provide less biased TE estimates. In this study, the dataset perfectly matched the above conditions as it contained a larger number of

covariates (38 covariates) and high-level nonlinearity degree was not found. Thus, for this study, I would recommend using the main effects only logistic regression. As mentioned earlier, using the main effects only logistic regression, the ASMD was about 0.23. Some researchers may argue that the covariate balance is not good enough and the TE estimates may still be biased. To further reduce the bias in the TE estimates, we can use a doubly-robust method, which regresses the outcomes on the covariates and the treatment condition on the matched sample or uses weighted regression of the outcome on the covariates. But this is out of the scope of this study. More details can be found in the studies by Funk et al. (2011) and by Stuart and Rubin (2007).

Table 31

ASMD, TE_read and TE_math before PS weighting and after PS weighting by logistic regression models

Weighting	Model Type	ASMD	TE_read	TE_math
Before	NA	0.527	-29.864	-22.247
After	lr_main	0.225	-8.641	-4.021
	lr_largest_asmd	0.225	-8.641	-4.021

Note. The estimates by lr_main and lr_largest_asmd have some minor differences from the 5th decimal place.

Table 32

ASMD, TE_read and TE_math by random forests using different hyper-parameters

node_size	mtry	sampling_replace	sampling_size	ASMD	TE_read	TE_math
1	p	TRUE	n	0.295	-10.020	-6.275
1	p	FALSE	0.632n	0.297	-10.226	-6.370
1	p	FALSE	0.5*n	0.285	-10.598	-6.971
1	sqrt(p)	TRUE	n	0.267	-13.457	-8.719
1	sqrt(p)	FALSE	0.632n	0.260	-13.034	-8.952
1	sqrt(p)	FALSE	0.5*n	0.254	-12.279	-8.026
1	log2p+1	TRUE	n	0.263	-12.534	-8.232
1	log2p+1	FALSE	0.632n	0.254	-12.314	-8.336
1	log2p+1	FALSE	0.5*n	0.269	-12.248	-8.246
0.05*n	p	TRUE	n	0.332	-6.730	-2.850
0.05*n	p	FALSE	0.632n	0.331	-5.999	-3.083
0.05*n	p	FALSE	0.5*n	0.318	-8.624	-4.511
0.05*n	sqrt(p)	TRUE	n	0.277	-8.814	-4.735
0.05*n	sqrt(p)	FALSE	0.632n	0.266	-9.692	-5.183
0.05*n	sqrt(p)	FALSE	0.5*n	0.274	-9.867	-5.246
0.05*n	log2p+1	TRUE	n	0.284	-10.022	-5.745
0.05*n	log2p+1	FALSE	0.632n	0.268	-8.725	-4.930
0.05*n	log2p+1	FALSE	0.5*n	0.282	-9.842	-5.597
0.1*n	p	TRUE	n	0.334	-7.161	-3.085
0.1*n	p	FALSE	0.632n	0.344	-4.818	-2.345
0.1*n	p	FALSE	0.5*n	0.330	-7.945	-3.852
0.1*n	sqrt(p)	TRUE	n	0.287	-8.822	-4.608
0.1*n	sqrt(p)	FALSE	0.632n	0.263	-9.492	-4.973
0.1*n	sqrt(p)	FALSE	0.5*n	0.296	-8.066	-3.591
0.1*n	log2p+1	TRUE	n	0.297	-7.327	-3.199
0.1*n	log2p+1	FALSE	0.632n	0.282	-9.230	-4.858
0.1*n	log2p+1	FALSE	0.5*n	0.287	-9.549	-4.704

5.4 Discussion

In this Chapter, I used an empirical data analysis to demonstrate the use of logistic regression and random forests with different hyperparameters in estimating the PSs and the treatment effect. The results of the simulation study provide some guidance about when to use

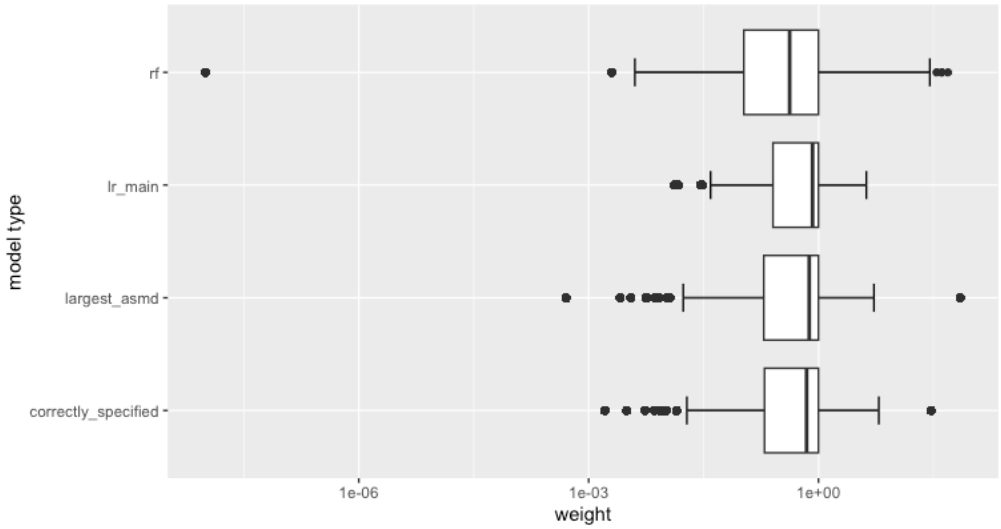
random forests over logistic regression and how to choose the hyperparameters in random forests. This empirical analysis can be used as a tutorial and illustration about how to use the guidance to choose the appropriate PS model and to set the values of the hyperparameters in random forests in the context of PS methods in practice. In this study, I used the R package randomForest. All the R code is provided in Appendix C. Researchers can easily adapt the R code based on the characteristics of their data to implement a random forests model with appropriate hyperparameters.

Chapter 6 Discussion

This dissertation, via a simulation study, provides information to answer the research questions of 1) in what way do the values of the hyperparameters in random forests influence the estimate of the treatment effect using PS weighting, and 2) under what conditions does the method of random forests perform better than logistic regression in estimating the treatment effect. Results from this study show that, in general, without hyper-parameter tuning, the performance of random forests is not comparable to that of logistic regression, which is indicated by the larger average ASMD across covariates, the larger relative bias of the TE estimates, and the greater variance of the TE estimates by random forests. One possible reason for the less comparable performance of random forests is the existence of extreme weights using random forests. To verify this hypothesis, I checked the weight distribution for each model on a randomly selected condition (condition A: 10 covariates, 20 subjects per covariate, nonlinearity type of interaction, medium level nonlinearity degree, 3 nonlinear terms). Results show that compared with logistic regression models, especially the main effects only logistic regression model, the random forests model has a larger variance in weights with more extreme weights (Figure 15). The extreme weights will bias the TE estimates as the TE is estimated as the difference in the weighted means of the outcome between the treatment and control groups (Hirano & Imbens, 2001; Li & Thomas, 2019). Further study needs to be done to trim the extreme weights and recheck the performance. Results from this study also show that when appropriate hyper-parameters are used, the performance of random forests improves substantially in balancing covariates though it still is not as good as logistic regression.

Figure 15

Weight distribution for subjects the control group by model, for the randomly picked condition A



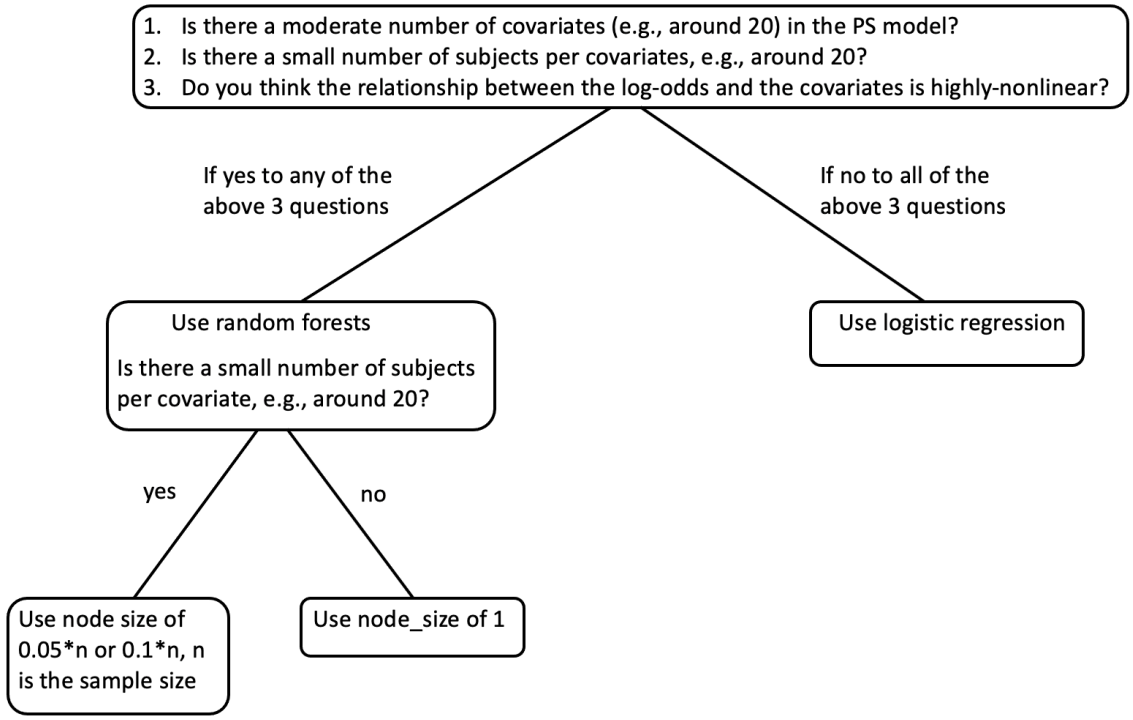
Note. This figure only shows the weights for subjects in the control group as the weight is 1 for subjects in the treatment group. For better visualization, the x-axis represents the weight on log scale.

Despite the less comparable performance of random forests in balancing covariates, with appropriate hyper-parameters, on average, random forests can provide more accurate TE estimates than logistic regression, especially for conditions with a moderate number of covariates, with a high-level nonlinearity, or with a smaller number of subjects per covariate. Regarding the impact of the hyper-parameters on the performance of random forests, results show that the sampling method does not matter a lot, and the default setting of mtry (the number of covariates from which to select at each split) sqrt(p) is in general a good setting as it enables better covariate balance, smaller bias and variance in the TE estimates. However, for the hyper-parameter of node_size (minimum terminal node size), the default setting of 1 is not always a good choice. Though, in general, node_size of 1 leads to better covariate balance and smaller

variance of the TE estimates, for conditions with a smaller number of subjects per covariate other settings of node_size such as 0.05*n and 0.1*n lead to less biased TE estimates than using node_size of 1. Considering that the number of subjects per covariate is the most influential factor of the comparative performance of random forests over logistic regression on accurately estimating the TEs and random forests is more likely to provide less biased TE estimates than logistic regression for conditions with a smaller number of subjects per covariate, choosing the right hyper-settings, especially right node_size, at such conditions is of great importance. Based on these findings, I made a flow chart to guide our decision about when to use which model and the choice of the hyper-parameters in random forests in practical settings (Figure 16). Specifically, given the conditions from my simulation study, I would recommend using random forests if there is a moderate number of covariates (e.g., around 20) in the PS model, or there is a small number of subjects per covariate (e.g., around 20), or we think the relationship between the log-odds of being assigned into the treatment group and the covariates is highly nonlinear. Otherwise, I would recommend using logistic regression as the PS model. When we use random forests, I would recommend using node_size of 5% of the sample size or 10% of the sample size if there is a small number of subjects per covariate. Otherwise, I would recommend using node_size of 1. This guidance is not guaranteed to achieve the best performance for all conditions. Yet, given the conditions I examined in the simulation study, it is very likely to help us improve the quality of the TE estimates.

Figure 16

Flow chart about when to use which model and the choice of the hyper-parameters in random forests



One interesting observation is that logistic regression models consistently perform better than random forests in balancing covariates (indicated by the smaller average ASMD using logistic regression models) even with hyper-parameter tuning. However, in terms of accuracy of the TE estimates (measured by the relative bias of the TE estimates), random forests can perform better than logistic regression models when appropriate hyper-parameters are used. This inconsistency of the comparative performance of random forests over logistic regression on the ASMD and the relative bias of the TE estimates indicates that maybe inspecting the ASMD on linear terms alone (the common practice to assess covariate balance) is not sufficient to measure the balance of the covariates. Considering that nonlinear terms may also have an impact on the

treatment assignment, it may be advisable to also check the ASMD on nonlinear terms. To verify this hypothesis, I checked the ASMD on both linear and nonlinear terms (all possible two-way interactions) on a randomly selected condition (condition A: 10 covariates, 20 subjects per covariate, nonlinearity type of interaction, medium level nonlinearity degree, 3 nonlinear terms). Results indicate that, using this measure of covariate balance, the random forests model performs better than logistic regression models on balancing the covariates when appropriate hyper-parameters are used, which is more consistent with the finding on the relative bias of the TE estimates (Table 33). Although additional research is required, these findings imply that it may be crucial to consider covariate balance on nonlinear terms when developing propensity score models and conducting balance checks, especially when the true propensity score model contains nonlinear terms. In addition to ASMD, variance ratios may also be used to assess the covariate balance. Basically, for each covariate, we can check the ratio of the variance of this covariate in the treatment group to that in the control group. And variance ratios close to 1 indicate good balance (Austin, 2009; Zhang et al., 2019). It would be beneficial to use variance ratios in combination of ASMD as ASMD primarily assesses the covariate balance in terms of means while variance ratios allow us to examine the covariate balance in terms of the distributions.

Table 33

Average ASMD with and without nonlinear terms for each model, for the randomly picked condition A

Model	Avg ASMD (only linear terms)	Avg ASMD (with nonlinear terms)
lr_main	0.14	0.173
lr_largest_asmd	0.164	0.176
lr_correctly_specified	0.165	0.18
rf	0.207	0.185
rf_best	0.146	0.145

Another interesting finding is that the result of this study is not quite consistent with the findings from the study of Lee et al. (2010). In that study, Lee and his colleagues found that random forests (using the default hyper-parameters) consistently provided more accurate TE estimates than logistic regression. However, in this study, I find that random forests only performs better than logistic regression models for some specific conditions when appropriate hyper-parameters are used. I tried to figure out the possible reasons via comparing the differences in the simulation study design between these two studies. Of course, my study covers more conditions than the study by Lee and his colleagues. But for the overlap conditions between their study and mine, namely conditions of no nonlinearity, conditions with one interaction term, and conditions with three interaction terms, the simulation study design of this study is quite similar to that of Lee et al. (2010). A minor difference is that in this study I used different coefficients for the intercept and nonlinear terms in the true PS model (to achieve specified nonlinearity degree and treat control ratio) when generating the data. But I do not think this difference should lead to the difference seen in the result (the comparative performance of random forests over logistic regression in estimating the TEs), as I also replicated the study of Lee et al. (2010) using the same coefficients as what Lee and his colleagues used but the results

still differed from what they reported (Table 34). Basically, results from the replication study indicate that the random forests model performs better (more accurate TE estimates) than logistic regression only for the conditions with one interaction term and also for the conditions with three interactions. Yet, for the conditions without any nonlinear terms, logistic regression performs better (more accurate TE estimates), which is more aligned with the results from my study.

Table 34

Means and standard errors of the relative TE bias for each model, from the study by Lee et al. (2010) and my replication of their study

Study	Model	Data from main effects only model	Data from model with one quadratic term	Data from model with three quadratic terms
Lee et al. (2010)	lr	11.2 (0.094)	13.8 (0.087)	12.1 (0.084)
	rf	9.2 (0.083)	11.4 (0.085)	8.6 (0.066)
My replication of the study by Lee et al. (2010)	lr	9.5 (0.074)	10.0 (0.076)	12.1 (0.084)
	rf	13.3 (0.087)	9.3 (0.068)	8.6 (0.066)

However, this simulation study is not without limitations. One potential limitation is the home advantage for logistic regression models due to the data generation method. Specifically, since the data are generated using logistic regression models, logistic regression models have some advantage over random forests in accurately estimating the PS and the corresponding TEs though the advantage is not salient through my examination. Basically, I randomly chose one condition (condition B: 10 covariates, 20 subjects per covariate, nonlinearity type of quadratic, low level nonlinearity degree, 1 nonlinear term) and examined the home advantage through checking the average PS error (the difference between the estimated PS and the true PS) and the average relative bias of the TE estimate for each model, the correlation between PS error and

relative bias of the TE estimate, and identifying whether PS error is a significant predictor of relative bias of the TE estimate. Results supported the existence of the home advantage as on average logistic regression models had smaller PS error than random forests and PS error was a significant predictor of the relative bias of the TE estimate ($p < 0.05$). Yet, the advantage was not very big as smaller PS error did not necessarily mean smaller bias in the TE estimates and the correlation between PS error and TE estimate bias was weak ($\rho = 0.03$) (Table 35). To avoid the issue of unfair comparison due to the potential home advantage, I thought about using some other data generation method, such as using other link function instead of logit. However, such kind of methods cannot well address the issue. Even if I use another method to generate the data, the home advantage may still exist as it is possible that logistic regression models may be more alike the true data generation method than random forests, or the other way around. And under such conditions, it becomes unclear which type of method has the home advantage. Considering these, finally I decided to use logistic regression models as the true data generation method as it was a common practice for data generation in previous studies and using this data generation method at least we know that logistic regression models have some home advantage. Specifically, if the random forests model performs better than logistic regression models, we can make the claim that the random forests model does perform better even with the home advantage for logistic regression models. Another limitation is that this study only covered limited conditions: limited types of PS methods, limited data generation conditions, and limited sets of hyperparameters in random forests. First of all, this study only explored one type of PS method, PS weighting, which may influence the generalizability of the results. Further study needs to be done to see if the result is consistent for other PS methods like PS matching. For PS matching, the conditioning of subjects in the treatment and control groups relies less on accurate estimates

of the PS, thus the difference between random forests and logistic regression may not be as salient as it is for PS weighting. Besides that, in this study I only covered limited data generation conditions. For example, I only manipulated limited between-cell factors and only used the treatment to control ratio of 1:1, which might influence the generalizability of the results. Additional research could be conducted to expand the analysis to cover more conditions and investigate the effects of different treatment to control ratios. In addition, in this study, for random forests I only used limited levels for the manipulated hyperparameters. The values I chose are either the default values in R packages or supported by previous studies. Further study can be made to expand the comparison to more levels or more hyperparameters. Besides the limitation in covering limited conditions, this simulation study may be better designed on some aspects. For example, in this study I quantified the nonlinearity degree via calculating the average of the absolute differences between the true PSs and the estimated PSs based on the best fitting main effects only logistic regression model. Yet, it may be more direct to measure the nonlinearity degree on logit scale as the true PS models are on logit scale. In this study, I used the average of the absolute differences between the true PSs and the estimated PSs as a measure of the nonlinearity degree because in this way the value of the nonlinearity degree is more interpretable, representing on average how different the estimated PS is from the true PS if the PS model is linear. Another example is that in this study, to make the simulation study design consistent with those of some previous studies (Lee et al., 2010; Setoguchi et al., 2008), I did not split the whole sample into training and validation subsets and choose the best performing random forests model based on the performance on the validation dataset, which is a typical practice in the conventional application of random forests: classification (Huang & Boutros,

2016; Verikas et al., 2011). Such design might make random forests models overfit and influence the generalizability of the results.

Table 35

Average ps error and average relative TE bias by model, for the randomly picked condition B

Model	Avg PS error	Avg relative TE bias
lr_main	0.088	0.206
lr_largest_asmd	0.107	0.228
lr_correctly_specified	0.077	0.246
rf	0.177	0.235

Even with these limitations, this study adds to the literature about the effects of the hyperparameters in random forest approaches on the estimate of the treatment effect in the context of PS methods, and provides some guidance about when to use random forests over logistic regression and how to choose the values of the hyperparameters in random forests using PS weighting. In addition to the simulation study, an empirical study is also included as a tutorial about when to use which method and how to choose the hyperparameters in random forests in practical settings. With the increasing interest in machine learning methods, such as random forests, it would be of great significance to know its theoretical framework and get some guidance about how to implement it in practice before it is more widely used in PS methods.

For future studies, it would be useful to trim the extreme weights and recheck of the performance of random forests and logistic regression models on estimating the PSs and the TEs. For random forests, it might be helpful to split the data into training and validation subsets to avoid overfitting and improve the generalizability of the results. In terms of checking the balance of the covariates, more studies could be made to examine the ASMD on both linear and nonlinear terms and use some other methods like variance ratios to check the covariate balance.

Furthermore, additional research could be conducted to expand the analysis to cover some other PS methods like PS matching, cover more conditions with more hyper-parameters in random forests, and examine the effects of different treatment to control ratios. For hyper-parameters in random forests, in this study I explored the top three most influential hyper-parameters, yet there are some other hyper-parameters that might also have an impact on the performance of random forests, such as the number of trees and the criterion to select the splitting covariate. Further studies could be made to examine the impacts of these hyper-parameters. About the treatment to control ratio, in practical settings, beside the ratio of 1:1 some other commonly seen treatment to control ratios include 4:6, 3:7, and 1:9 (Dehejia & Wahba, 1999; Hirano & Imbens, 2003; Ozminkowski et al., 2007; Zanutto, 2021). More studies could be made to explore the effects of different treatment to control ratios. In addition, future studies could examine the use of some other ML methods in the context of PS methods, such as boosting, neural network etc. In some recent studies such methods have been found to perform pretty well (sometimes even better than random forests) in accurately and reliably estimating the TEs (Ferri-García & Rueda, 2020; Kern et al., 2020; Westreich et al., 2010).

Appendix A

Pilot Study about Number of Nonlinear Terms

To determine whether there is a need to include number of nonlinear terms as a factor in the simulation study, I conducted a pilot study to check the impact of different numbers of nonlinear terms on the comparative performance of random forests over logistic regression. To simplify the analysis, I chose a subset of the conditions. Specifically, for between-cell conditions I chose the conditions of 10 covariates and a high-level of nonlinearity degree for each of the three nonlinearity types (quadratic, interaction, and exponential) while for within-cell conditions I compared the main effects only logistic regression model with the random forests model with the default hyper-parameters (node_size, 1; mtry: sqrt(p); sampling with replacement with the original sample size). The analysis was conducted in the following steps:

1. For each nonlinearity type, I generated the data based on two PS models: one with a small number of nonlinear terms and one with a comparatively large number of nonlinear terms. In this pilot study, I set the small number of nonlinear terms to be 1 and the comparatively large number of nonlinear terms to be 3. This choice was to represent the ratio of linear versus nonlinear terms in some practical studies and also ensured sufficient differences between levels (Hong, 2005). Coefficients were adjusted to ensure that, for each nonlinearity type, the datasets with different numbers of nonlinear terms have the same nonlinearity degree.
2. For datasets with each number of nonlinear terms, I estimated the TEs using both the main effects only logistic regression model and the random forests model with the different hyper-parameters.
3. I calculated the difference in the estimated TEs between random forests and logistic regression for each nonlinearity type with each number of nonlinear terms. For example, for the nonlinearity type of quadratic, I obtained the differences in the estimated TEs between random forests and logistic regression for one quadratic term (diff_quad_one) and also for three quadratic terms (diff_quad_three).
4. I repeated the above Steps 1-3 for 100 replications.
5. For each nonlinearity type, I conducted a t-test to check whether the difference in the estimated TEs between random forests and logistic regression differs significantly for different numbers of nonlinear terms.

Results indicated that, for the nonlinearity types of quadratic and interaction, the difference in the estimated TEs between random forests and logistic regression differed significantly for different numbers of nonlinear terms ($p < 0.05$). For the nonlinearity type of exponential, the difference is not statistically significant yet its p-value is very close to the cut-off value of 0.05 ($p = 0.07$). Based on this, I would say that the number of nonlinear terms has an impact on the comparative performance of random forests over logistic regression in correctly estimating the TE, thus I included it as a factor in the simulation study.

Appendix B

Coefficients in the Simulation Study

Table 1

Coefficients of the nonlinear terms in the conditions of 10 covariates

Nonlinear	Quadratic			Nonlinear	Interaction			Nonlinear	Exponential		
Term	0.05	0.1	0.18	Term	0.05	0.1	0.27	Term	0.01	0.03	0.05
X_2^2	0.2721	0.5892	1.4997	$X_1 * X_2$	-0.7453	-1.9668	21.8272	e^{X_2}	0.0538	0.2042	0.4107
X_4^2	0.0195	0.0682	0.1577	$X_3 * X_4$	-0.0459	0.1949	-2.7801	e^{X_4}	0.0097	0.0101	-0.0459
X_7^2	0.0060	0.0210	-0.1617	$X_6 * X_7$	0.1005	0.0599	8.9499	e^{X_7}	0.0030	-0.0246	0.1005

Table 2

Coefficients of the nonlinear terms in the conditions of 20 covariates

Nonlinear	Quadratic			Nonlinear	Interaction			Nonlinear	Exponential		
Term	0.05	0.1	0.18	Term	0.05	0.1	0.27	Term	0.01	0.03	0.05
X_2^2	0.2466	0.7175	-0.6085	$X_1 * X_2$	-0.2083	1.1628	9.6205	e^{X_2}	0.0358	0.2054	0.4824
X_4^2	0.0091	0.0137	0.0553	$X_3 * X_4$	0.0914	0.2766	-7.9193	e^{X_4}	0.0046	0.0277	0.0277
X_7^2	-0.0248	-0.0371	0.0440	$X_6 * X_7$	-0.2476	0.2202	8.2457	e^{X_7}	-0.0124	0.0220	0.0220
X_{12}^2	0.1721	-0.0390	1.3046	$X_{11} * X_{12}$	0.7068	-1.2244	3.2579	$e^{X_{12}}$	0.0141	0.0757	-0.1439
X_{14}^2	-0.0154	-0.0231	-0.0296	$X_{13} * X_{14}$	-0.1540	-0.1478	-1.3170	$e^{X_{14}}$	-0.0077	-0.0148	-0.0148
X_{17}^2	0.0138	0.0207	0.0553	$X_{16} * X_{17}$	0.1382	0.2763	0.9087	$e^{X_{17}}$	0.0069	0.0276	0.0276

Table 3

Coefficients of the nonlinear terms in the conditions of 40 covariates

Nonlinear	Quadratic			Nonlinear	Interaction			Nonlinear	Exponential		
Term	0.05	0.1	0.18	Term	0.05	0.1	0.27	Term	0.01	0.03	0.05
X_2^2	0.0271	0.3671	-0.5780	$X_1 * X_2$	0.1268	-1.1990	9.0060	e^{X_2}	0.0277	0.1684	0.4995
X_4^2	0.0434	-0.1289	-0.0322	$X_3 * X_4$	-0.0610	0.1828	1.9214	e^{X_4}	-0.0114	-0.0228	-0.0456
X_7^2	0.0144	0.0291	0.0073	$X_6 * X_7$	0.0605	0.4404	6.6940	e^{X_7}	0.0032	0.0064	0.0128
X_{12}^2	0.0495	-0.1993	0.8520	$X_{11} * X_{12}$	0.5445	-1.2273	3.4361	$e^{X_{12}}$	0.0220	0.1245	-0.1183
X_{14}^2	0.0004	0.0439	0.0110	$X_{13} * X_{14}$	-0.1302	-0.3209	-6.6063	$e^{X_{14}}$	-0.0024	-0.0048	-0.0097
X_{17}^2	-0.0294	-0.1222	-0.0306	$X_{16} * X_{17}$	-0.0859	0.1976	3.8147	$e^{X_{17}}$	-0.0107	-0.0214	-0.0428
X_{22}^2	-0.0243	0.5566	0.7375	$X_{21} * X_{22}$	0.4888	-1.3695	-8.4435	$e^{X_{22}}$	-0.0148	-0.2983	0.4060
X_{24}^2	-0.0454	-0.0680	-0.0170	$X_{23} * X_{24}$	0.0101	-0.3161	1.0524	$e^{X_{24}}$	0.0038	0.0076	0.0151
X_{27}^2	0.0212	0.1033	0.0258	$X_{26} * X_{27}$	0.0594	0.2159	-3.8618	$e^{X_{27}}$	-0.0435	-0.0869	-0.1739
X_{32}^2	0.3721	-0.1432	-0.8129	$X_{31} * X_{32}$	0.4137	0.1583	-8.6523	$e^{X_{32}}$	0.0276	-0.0684	-0.3438
X_{34}^2	0.0038	0.0364	0.0091	$X_{33} * X_{34}$	-0.1290	0.1242	6.5419	$e^{X_{34}}$	-0.0450	-0.0901	-0.1801
X_{37}^2	-0.0433	-0.0750	-0.0188	$X_{36} * X_{37}$	0.1917	0.1013	6.3484	$e^{X_{37}}$	0.0349	0.0697	0.1395

Table 4

Coefficients of the linear terms across conditions

Coefficient	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Value	0.80	-0.25	0.60	-0.40	-0.80	-0.50	0.70

Note. β_0 is 0 to make it consistent with the study of Lee et al. (2010).

Appendix C

Code of the Empirical Study

```
### ---------- Empirical Study Analysis ---------- ###
library(randomForest)

# read data
# setwd("~/Desktop/umd-dissertation/empirical_study")
child <- read.csv('data/child_engineered.csv')
data <- child[, !names(child) %in% c('READ_WAVE4', 'MATH_WAVE4')]

# check avg outcomes by group
aggregate(child$READ_WAVE4, by=list(child$REPEAT_K), mean)
aggregate(child$MATH_WAVE4, by=list(child$REPEAT_K), mean)

# unweighted asmd
repeat_k <- data$REPEAT_K
asmd_values <- c()
ws <- c(data[, 2:39])
for (w in ws) {
  avg_w_treat <- mean(w[repeat_k == 1])
  sd_w_treat <- sd(w[repeat_k == 1])
  avg_w_control <- mean(w[repeat_k == 0])
  w_asmd <- abs(avg_w_treat - avg_w_control) / sd_w_treat
  asmd_values <- c(asmd_values, w_asmd)
}
asmd_values
mean(asmd_values)

## ---------- Fit Logistic Regression Models ---------- ##
## lr, main effects only
lr_main <- glm(REPEAT_K ~ ., family = binomial(), data)
pred <- predict(lr_main, data)
est_ps <- exp(pred) / (1 + exp(pred))
weight <- ifelse(data['REPEAT_K'] == 0, est_ps / (1 - est_ps), 1)

TE_read <- weighted.mean(child[data['REPEAT_K'] == 1, 'READ_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
  weighted.mean(child[data['REPEAT_K'] == 0, 'READ_WAVE4'], weight[data['REPEAT_K']
== 0]) # -8.640991
TE_math <- weighted.mean(child[data['REPEAT_K'] == 1, 'MATH_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
  weighted.mean(child[data['REPEAT_K'] == 0, 'MATH_WAVE4'], weight[data['REPEAT_K']
== 0]) # -4.020549
TE_read
TE_math
```

```
# asmd by lr_main
asmd_values <- c()
for (w in ws) {
  avg_w_treat <- mean(w[repeat_k == 1])
  sd_w_treat <- sd(w[repeat_k == 1])
  weighted_avg_w_control <- weighted.mean(w[repeat_k == 0], weight[repeat_k == 0])
  w_asmd <- abs(avg_w_treat - weighted_avg_w_control) / sd_w_treat
  asmd_values <- append(asmd_values, w_asmd)
}
asmd_values
mean(asmd_values)

## lr_largest_asmd
# find the most unbalanced covariate
asmd_values
which.max(asmd_values)
ws[which.max(asmd_values)]

data_largest_asmd <- data
for (i in 1:length(ws)) {
  data_largest_asmd[, 39+i] <- ws[[i]] * ws[[20]]
}

lr_largest_asmd <- glm(REPEAT_K ~ ., family = binomial(), data_largest_asmd)
pred <- predict(lr_largest_asmd, data_largest_asmd)
est_ps <- exp(pred) / (1 + exp(pred))
weight <- ifelse(data['REPEAT_K'] == 0, est_ps / (1 - est_ps), 1)
TE_read <- weighted.mean(child[data['REPEAT_K'] == 1, 'READ_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
  weighted.mean(child[data['REPEAT_K'] == 0, 'READ_WAVE4'], weight[data['REPEAT_K']
== 0]) # -8.64; -8.640991
TE_math <- weighted.mean(child[data['REPEAT_K'] == 1, 'MATH_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
  weighted.mean(child[data['REPEAT_K'] == 0, 'MATH_WAVE4'], weight[data['REPEAT_K']
== 0]) # - 4.02; -4.020549
TE_read
TE_math

# asmd by lr_largest_asmd
asmd_values <- c()
for (w in ws) {
  avg_w_treat <- mean(w[repeat_k == 1])
  sd_w_treat <- sd(w[repeat_k == 1])
  weighted_avg_w_control <- weighted.mean(w[repeat_k == 0], weight[repeat_k == 0])
  w_asmd <- abs(avg_w_treat - weighted_avg_w_control) / sd_w_treat
```

```
  asmd_values <- append(asmd_values, w_asmd)
}
asmd_values
mean(asmd_values)

## ---------- Fit Random Forests Models ---------- ##
# write functions to calcuate asmd, te
compute_weighted_asmd <- function(data, weight) {
  repeat_k <- data$REPEAT_K
  ws <- c(data[, 2:39])
  asmd_values <- c()
  for (w in ws) {
    avg_w_treat <- mean(w[repeat_k == 1])
    sd_w_treat <- sd(w[repeat_k == 1])
    weighted_avg_w_control <- weighted.mean(w[repeat_k == 0], weight[repeat_k == 0])
    w_asmd <- abs(avg_w_treat - weighted_avg_w_control) / sd_w_treat
    asmd_values <- append(asmd_values, w_asmd)
  }
  return(mean(asmd_values))
}

compute_te <- function(data, child, weight) {
  TE_read <- weighted.mean(child[data['REPEAT_K'] == 1, 'READ_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
    weighted.mean(child[data['REPEAT_K'] == 0, 'READ_WAVE4'], weight[data['REPEAT_K']
== 0])
  TE_math <- weighted.mean(child[data['REPEAT_K'] == 1, 'MATH_WAVE4'],
weight[data['REPEAT_K'] == 1]) -
    weighted.mean(child[data['REPEAT_K'] == 0, 'MATH_WAVE4'], weight[data['REPEAT_K']
== 0])
  return(list(TE_read = TE_read, TE_math = TE_math))
}

# get the results for diff rf hypers
n <- 771
p <- 38
node_sizes <- list(list(name = '1', value = 1),
                   list(name = '0.05*n', value = round(0.05*n)),
                   list(name = '0.1*n', value = round(0.1*n)))
mtrys <- list(list(name = 'p', value = round(p)),
              list(name = 'sqrt(p)', value = round(sqrt(p))),
              list(name = 'log2p+1', value = ceiling(log2(p) + 1)))
sampling_replaces <- c(TRUE, FALSE, FALSE)
sampling_sizes <- list(list(name = 'n', value = n),
                       list(name = '0.632n', value = round(0.632*n)),
                       list(name = '0.5*n', value = round(0.5*n)))
```

```
result <- data.frame(matrix(ncol = 0, nrow = 27))
node_size_col <- c()
mtry_col <- c()
sampling_replace_col <- c()
sampling_size_col <- c()
asmd_col <- c()
TE_read_col <- c()
TE_math_col <- c()

rf_x <- data[, 2:39]
rf_y <- data$REPEAT_K

set.seed(412)
for (node_size in node_sizes) {
  node_size_name <- node_size$name
  node_size_value <- node_size$value
  for (mtry in mtrys) {
    mtry_name <- mtry$name
    mtry_value <- mtry$value
    for (i in 1:length(sampling_replaces)) {
      rf <- randomForest(x = rf_x, y = as.factor(rf_y),
                         nodesize = node_size_value, mtry = mtry_value,
                         replace = sampling_replaces[i], sampsize = sampling_sizes[[i]]$value)
      est_ps <- predict(rf, data, type='prob')[, '1']
      weight <- ifelse(data['REPEAT_K'] == 0, est_ps / (1 - est_ps), 1)

      asmd <- compute_weighted_asmd(data, weight)
      te_results <- compute_te(data, child, weight)
      TE_read <- te_results$TE_read
      TE_math <- te_results$TE_math

      node_size_col <- c(node_size_col, node_size_name)
      mtry_col <- c(mtry_col, mtry_name)
      sampling_replace_col <- c(sampling_replace_col, sampling_replaces[i])
      sampling_size_col <- c(sampling_size_col, sampling_sizes[[i]]$name)
      asmd_col <- c(asmd_col, asmd)
      TE_read_col <- c(TE_read_col, TE_read)
      TE_math_col <- c(TE_math_col, TE_math)
    }
  }
}

result$node_size <- node_size_col
result$mtry <- mtry_col
result$sampling_replace <- sampling_replace_col
```

```
result$sampling_size <- sampling_size_col
result$asmd <- asmd_col
result$TE_read <- TE_read_col
result$TE_math <- TE_math_col

write.csv(result, 'data/results_rf.csv', row.names = F)
```

References

Amini, P., Maroufizadeh, S., O Samani, R., Hamidi, O., & Sepidarkish, M. (2022). Factors associated with macrosomia among singleton live-birth: A comparison between logistic regression, random forest and artificial neural network methods. *Epidemiology, Biostatistics, and Public Health, 13*(4). https://doi.org/10.2427/11985.

Arceneaux, K., Gerber, A., & Green, D. P. (2010). A cautionary note on the use of matching to estimate causal effects: an experimental example comparing matching estimates to an experimental benchmark. *Sociological Methods and Research, 39*(2), 256.

Arpino, B. & Cannas, M. (2015). Comparing different approaches for propensity score matching with clustered data: a simulation study (RECSM Working Paper No. 43, 1-26). Barcelona, Spain: University Pompeu Fabra.

Arpino, B. & Cannas, M. (2018). Machine learning for propensity score matching and weighting: comparing different estimation techniques and assessing different balance diagnostics. *Research and Expertise Centre for Survey Methodology*, working pager No. 54. https://repositori.upf.edu/bitstream/handle/10230/33768/RECSM%20working%20paper54.pdf?sequence=1&isAllowed=y

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*, 3083-3107.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399-424.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. *Multivariate Behavioral Research, 47*, 115-135.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*, 1057-1069.

Bai, H. (2011). A comparison of propensity score matching methods for reducing selection bias. *International Journal of Research & Method in Education, 34*(1), 81-107.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health, 9*(6), 377 – 385.

Bernald, S., Heutte, L., & Adam, S. (2007). Using random forests for handwritten digit recognition. *International Conference on Document Analysis and Recognition*, 1043-1047.

Bernard, S., Heutte, L., Adam, S. (2009). Influence of hyperparameters on random forest accuracy. *Lecture Notes in Computer Science*, 171-180.

Breiman, L., Friedman, Olshen, R., Stone, C. (1984). Classification and regression trees. *Wadsworth Books*, 358.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Barman, R., Saha, S., & Das, S. (2014). Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLOS ONE, 9*(11), 1–11.

Bauer, E., & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning, 36*(1), 105–139. https://doi.org/10.1023/A:1007515423169.

Bostrom, H. (2007). Estimating class probabilities in random forests. *IEEE International Conference on Machine Learning and Applications.*

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., & Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology, 28*(2), 171–182. https://doi.org/10.1002/gepi.20041.

Cavicchioli, M., Papana, A., Dagiasis, A. P., & Pistoresi, B. (2019). A Random Forests Approach to Assess Determinants of Central Bank Independence. *Journal of Modern Applied Statistical Methods, 17*(2). https://doi.org/10.22237/jmasm/1553610953.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–7t2.

Cham, H. N. (2013). Propensity score estimation with random forests (Doctoral dissertation). Retrieved from Arizona State University.

Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data, 7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313.

Cochran, W. G. & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *The Indian Journal of Statistics, Series A*, 417-446.

Cochran, W.G. & Chambers, S. P. (1965). The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society, Series A*(128), 134–155.

Couronne, R., Probst, P., & Boulesteix, A. L., (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), https://doi.org/10.1186/s12859-018-2264-5.

Cox, D. R. (1958). *Planning of Experiments*. New York: John Wiley & Sons.

Craycroft, J. (2016). Propensity score methods: a simulation and case study involving breast cancer patients. (Doctoral Dissertation), Retrieved from University of Louisville.

Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics, 10*, 117-131.

Dagostino, R. B. (1998). Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

Dankowski, T. & Ziegler, A. (2016). Calibrating random forests for probability estimation. *Statistics in Medicine, 35*, 3949-3960.

Dehejia, R. H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*, 1053–1062.

Dehejia, R. H. & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1) 151–161.

Delgado, M. F., Cernadas, E., Barro, S., Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133-3181.

Dettling, M. & Buhlmann, P. (2003). How to use boosting for tumor classification with gene expression data. Preprint. http://stat.ethz.ch/dettling/boosting.

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics, 95*(3), 932–945.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning, 40*, 139-157.

Dietterich, T. D. (2002). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.). Lecture notes in computer science: Multipl classifier systems (pp. 1-15). New York: Springer.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*, 1231–1236.

Emancipator, K., & Kroll, M. H. (1993). A quantitative measure of nonlinearity. *Clin Chem, 39*(5), 766-772.

Ferri-García, R., & Rueda, M. del M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLOS ONE, 15*(4), e0231500. https://doi.org/10.1371/journal.pone.0231500.

Frangakis, C. & Rubin, D. B. (2002). Principal stratification in casual inference. *Biometrics, 58*(1), 21–29.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *The Annuals of Statistics, 28* 337-407.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annuals of Statistics, 29*, 1189-12.

Friedman, J. H., & Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference, 137*(3), 669-683.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*, 761-767.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning.* New York: Springer. p. 315. ISBN 978-1-4614-7137-0.

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (1998). Dual and Simultaneous Sensitivity Analysis for Matched Pairs. *Biometrika, 85*(4), 907–920.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 36*(1), 3-42.

Goldstein, B., A., Polley, E. C., & Briggs, F. (2011). Random forests for genetic n association studies. *Statistical Applications in Genetics and Molecular Biology, 10*, 32.

Goldstein, B., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal, 38*(23), 1805-1814.

Grace, J.B. (2008). Structural equation modeling for observational studies. *Journal of Wildlife Management, 72*, 14–22.

Greenland, S. (1996). Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology, 25*(6), 1107–1116. https://doi.org/10.1093/ije/25.6.1107.

Gu, X.S. & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405-420.

Guo, S. & Fraser, M. (2010). Propensity score analysis: Statistical methods and applications. *Thousand Oaks, Calif.*: Sage Publications.

Hahn, J. 1998. On the role of the propensity score in efficient estimation of average treatment effects. *Econometrica, 66*(2), 315-331.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*, 234.

Harding, D. J. (2003). Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy. *American Journal of Sociology, 109*(3), 676–719. https://doi.org/10.1086/379217.

Harrell, F. E. (2001). Multivariable Modeling Strategies. In F. E. Harrell (Ed.), *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis* (pp. 53–85). Springer. https://doi.org/10.1007/978-1-4757-3462-1_4.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer: New York.

Hausman, J., Newey, W. K., Woutersen, T., Chao, J., & Swanson, N. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics, 3*(2), 211–55.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Review of Economic Studies, 64*, 605–654.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. *Applied Bayesian and causal inference from an incomplete data perspective* (pp. 49-60). New York: Wiley.

Hirano, K., Imbens, G. W. (2003). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services & Outcomes Research Methodology 2*, 259–278.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236.

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

Hong, G. & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901–910.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205-224.

Hong, G. (2010). Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics, 35*(5), 499-531.

Hong, G. (2012). Marginal mean weighting through stratification: a generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological Methods, 17*(1), 44-60.

Hopfenberg, R., & Pimentel, D. (2001). Human population numbers as a function of food supply. *Environment, Development and Sustainability, 3*, 1-15.
Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics, 3*, 651-674.
Huang, B. F. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC Bioinformatics, 17*(1), 331. https://doi.org/10.1186/s12859-016-1228-x.
Hutter, F., Hoos, H., Brown, K. L. (2014). *International Conference on Machine Learning.*
Im, M. H., Hughes, J. N., Kwok, O., Puckett, S., & Cerda, C. A. (2013). Effect of retention in elementary grades on transition to middle school. *Journal of School Psychology, 51*(3), 349–365. https://doi.org/10.1016/j.jsp.2013.01.004.
Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*(171), 481-502.
Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, 76*(1), 243-263.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*, 4-29.
Jiang, R., Tang, W., Wu, X., Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics, 10*(Suppl 1), S65.
Kern, C., Li, Y., & Wang, L. (2021). Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples. *Journal of Survey Statistics and Methodology, 9*(5), 1088–1113. https://doi.org/10.1093/jssam/smaa028.
Kotthoff, L., Thornton, C., Hoos, H., Hutter, F., Brown, K. L. (2017). *Journal of Machine Learning Research, 18*, 1-5.
Kruppa, J., Schawarz, A., Arminger, G., Ziegler, A. (2013). Consumer credit rish: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*, 5125 – 5131.
Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology, 163*, 262- 270.
Latinne, P., Debeir, O., & Decaestecker, C. (2001) Limiting the number of trees in random forests. Paper presented at 2nd International Workshop on Multiple Classifier System.
Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.
Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337–346.
Lee, H. & Munk, T. (2008). Using regression discontinuity design for program evaluation, section on survey research methods, joint statistical meetings. Denver CO: *American Statistical Association*.
Li, F., & Thomas, L. E. (2018). Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwy201
Li, J. (2016). Modeling approaches for cost and cost- effectiveness estimation using observational data. (Doctoral Dissertation), Retrieved from University of Pennsylvania.
Li, X. R. (2012). Measure of nonlinearity for stochastic systems. *Proceedings of the 15th International Conference on Information Fusion*, 1073-1080.

Liaw, A., Wiener, M., (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.
Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies. *Biometrics, 54*(3), 948. https://doi.org/10.2307/2533848.
Lin, Y., Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association, 101*, 578–590.
Lingle, J. A. (2009). Evaluating the performance of propensity scores to address selection bias in a multilevel context: A Monte Carlo simulation study and application using a national data set. Unpublished Dissertation.
Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science*, *14*(6), 570–580. https://doi.org/10.1007/s11121-012-0339-5.
Loh, W. & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association, 83*, 715–725.
Loh, W. (2014). Fifty years of classification and regression trees. *International Statistical Review, 82*(3), 329-348.
Lunetta, K. L., Hayward, L. B., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics, 5*(1), 32. https://doi.org/10.1186/1471-2156-5-32
MacCallum, R. C. & Austin, J. T. (2000) Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201-226.
Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods Inf Med, 51*(1), 74-81.
Martinez-Munoz, G., & Suaresz, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition, 43*, 143-152.
Mehta, P., Bukov, M., Wang, C.-H., Day, A. G. R., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports, 810*, 1–124. https://doi.org/10.1016/j.physrep.2019.03.001.
McCaffrey, D., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403-425.
McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramch Strobl and, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models, 32: 3388–414. doi:10.1002/ sim.5753.
Moisen, G. G. (2008). Classification and regression trees. *Ecological Informatics*, 582-588.
Morgan, S. L. & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*, 231-281.
Morgan, S.L. & Winship, C. (2007). *Counterfactuals and causal inference: methods and princirples for social research*. New York: Cambridge University Press.
Mullahy, J. (1997). Instrumental-variable estimation of count data models: applications to models of cigarette smoking behaviour. *Review of Economics and Statistics, 11*, 586–593.
Nichols, A. & McBride, L. (2017). Propensity scores and causal inference using machine learning methods. 2017 Stata Conference 13, Stata Users Group.

Orlandi, I., Oneto, L., Anguita, D. (2016). Random forests modern selection. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Oshiro, T. M., Perez, P. S., Baranauskas, J. A. (2012). How many trees in a random forest? In Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg.

Ozminkowski, R. J., Wang, S., & Walsh, J. K. (2007). The direct and indirect costs of untreated insomnia in adults in the United States. *Sleep, 30*(3), 263–273. https://doi.org/10.1093/sleep/30.3.263.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Pan, W. & Bai, H. (2015). *Propensity score analysis: Fundamentals and developments.* New York, NY: The Guilford Press.

Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825 – 2830.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373 - 1379.

Peters, A., Hothorn, T., Berthold, L. (2002). Ipred: Improved predictors. *R News, 2*(2), 33-36.

Pham, T. T. (2017). Large scale causal inference with machine learning. (Doctoral Dissertation), Retrieved from Stanford University.

Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 14*(1), 137. https://doi.org/10.1186/1471-2288-14-137.

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research, 18*, 1-18.

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research, 20*(1), 1934–1965.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301. Retrieved from https://doi.org/10.1002/widm.1301

Raghu, A., Devarsetty, P., David, P., Lionel, T., & Gari, C. (2015). Implications of cardiovascular disease risk assessment using the who/ish risk prediction charts in rural india. Https://doi: 10.1371/journal.pone. 0133618.

Rao, J. S. & Potts, W. J. E. (1997). Visualizing bagged decision trees. Proceedings of the *Third International Conference on Knowledge Discovery and Data Mining* (pp. 243–246). Newport Beach, CA: AAAI Press.

Ridgeway, G. (1999). *The State of Boosting ∗*. https://www.semanticscholar.org/paper/The-State-of-Boosting-%E2%88%97-Ridgeway/1aac6453fbb8333ee638b6d8b2bb2aff06c3654b.

Rijin, J., & Hutter, F., (2018). Hyperparameter importance across datasets. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

Rosenbaum, P. R. (1987). Model-based direct adjustment. *The Journal of the American Statistician 82*, 387-394.
Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of American Statistician Association, 84*(408), 1024-32.
Rosenbaum, P. R. (1996). Observational studies and nonrandomized experiments. *Handbook of Statistics 13*, 181–197.
Rosenbaum, P. R. (2002). *Observational Studies.* Springer. https://doi.org/10.1007/978-1-4757-3692-2, 105-170.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.
Rosenbaum, P. R. & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41*, 103-116.
Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*, 159-184.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non- randomized studies. *Journal of Educational Psychology, 66*, 688-701.
Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annuals of Statistics 7*, 34–58.
Rubin, D. B. & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics, 52*, 249-264.
Segal, M. R., Barbour, J. D., Grant, R. M. (2004), Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology, 3*, article 2.
Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546–555.
Shadish W. R. (2002). Revisiting field experimentation: field notes for the future. *Psychologocial Methods, 7*, 3-18.
Shadish W. R., Clark, M. H., Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.
Shadish, W. R. & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews, 10*(1), 19-26.
Shen, C., Li, X., Li, L., & Were, M. C. (2011). Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical Journal, 53*(5), 822–837. https://doi.org/10.1002/bimj.201100042.
Smith, J. & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics, 125*(1–2), 305–353.
Sommer, C., Straehle, C., Koethe, U., & Hamprecht, F. (2010). Interactive Learning and Segmentation Tool Kit. In *Systems Biology of Human Disease*, 230–33.
Steiner, P. M. & Cook, T. D. (2013). Matching and propensity scores. In *T.D. Little (Ed), The Oxford Handbook of Quantitative Methods, Volume 1(Foundations),* 237- 259. New York, NY: Oxford University Press.

Steiner, P.M., Cook, T.D., & Shadish W.R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*, 213-236.

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. Academic Emergency *Medicine, 18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

Stone R. A., Obrosky D. S., Singer D. E., Kapoor W.N., Fine M.J. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. Pneumonia Patient Outcomes Research Team (PORT) Investigators. *Medical Care, 33*(4), 56-66.

Strobl, C., Boulesteix, A. L., Augustin, T. (2007). Unbiased split selection for classification trees based on Gini index. *Computational Statistics and Data Analysis, 52*, 483-501.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for Random Forests. *BMC Bioinformatics, 9*(307).

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: sources and a solution. *BMC Bioinformatics*, 8-25.

Strobl, C., Malley, J., Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods 14*(4), 323-348.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: A review journal of the Institute of Mathematical Statistics, 25*(1), 1-21.

Stuart, E., & Rubin, D. (2008). Best practices in quasi–experimental designs: matching methods for causal inference. In *Best Practices in Quantitative Methods* (pp. 155-176). SAGE Publications, Inc., https://doi.org/10.4135/9781412995627.

Thoemmes, F. & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118.

Thoemmes, F. & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*, 514-543.

VanderWeele, T. J., & Arah, O. A. (2011). Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology, 22*(1), 42–52. https://doi.org/10.1097/EDE.0b013e3181f74493.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330–349. https://doi.org/10.1016/j.patcog.2010.08.011.

Watkins, S., Jonsson-funk, M., Brookhart, M., Rosenberg S., O'shea T., & Daniels J. (2013) An empirical comparison of tree-based methods for propensity score estimation. *Health Service Research, 48*(5), 1798–817.

Westreich, D., Lessler, J., & Funk, M.J. (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*, 826– 833.

Wolfe, F., & Michaud, K. (2004). Lymphoma in rheumatoid arthritis: The effect of methotrexate and anti-tumor necrosis factor therapy in 18,572 patients. *Arthritis Rheum, 50*, 1740–1751.

Wu, Q., Burges, C., & Svore, K., Gao, J., (2009). Adapting boosting for information retrieval measures. *Learning to Rank for Information Retrieval*, DOI 10.1007/s10791-009-9112-1.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics, 72*(4), 1055–1065. https://doi.org/10.1111/biom.12505.

Zanatto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4*, 67-91.

Zanutto, E. L. (2021). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science, 4*(1), 67–91. https://doi.org/10.6339/JDS.2006.04(1).233

Zanutto, E., Lu, B., & Hornik, R. (2005). Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Antidrug Media Campaign. *Journal of Educational and Behavioral Statistics, 30*(1), 59–73. https://doi.org/10.3102/10769986030001059.

Zhang, Y. & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C 58*, 308–324.

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine, 7*(1), 16. https://doi.org/10.21037/atm.2018.12.10.

Zhao, P., Su, X., Ge, T., & Fan, J. (2016). Propensity Score and Proximity Matching Using Random Forest. *Contemporary Clinical Trials, 47*, 85–92. https://doi.org/10.1016/j.cct.2015.12.012.

Zhu, Y., (2013). Data-adaptive approaches to modeling propensity scores in causal inference problems. (Doctoral Dissertation), Retrieved from Pennsylvania State University.