Propensity Score Estimation with Random Forests

by

Hei Ning Cham

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2013 by the
Graduate Supervisory Committee:

Jenn-Yun Tein, Co-Chair
Stephen West, Co-Chair
Craig Enders
David MacKinnon

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

Random Forests is a statistical learning method which has been proposed for propensity score estimation models that involve complex interactions, nonlinear relationships, or both of the covariates. In this dissertation I conducted a simulation study to examine the effects of three Random Forests model specifications in propensity score analysis. The results suggested that, depending on the nature of data, optimal specification of (1) decision rules to select the covariate and its split value in a Classification Tree, (2) the number of covariates randomly sampled for selection, and (3) methods of estimating Random Forests propensity scores could potentially produce an unbiased average treatment effect estimate after propensity scores weighting by the odds adjustment. Compared to the logistic regression estimation model using the true propensity score model, Random Forests had an additional advantage in producing unbiased estimated standard error and correct statistical inference of the average treatment effect. The relationship between the balance on the covariates' means and the bias of average treatment effect estimate was examined both within and between conditions of the simulation. Within conditions, across repeated samples there was no noticeable correlation between the covariates' mean differences and the magnitude of bias of average treatment effect estimate for the covariates that were imbalanced before adjustment. Between conditions, small mean differences of covariates after propensity score adjustment were *not* sensitive enough to identify the optimal Random Forests model specification for propensity score analysis.

ACKNOWLEDGEMENTS

I thank Jan Hughes, Department of Educational Psychology, Texas A&M University, College Station, TX 77843, and the grant from the National Institute of Child Health and Human Development (grant number HD039367), for generously providing the data for the empirical example.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

.

.

Chapter 1

INTRODUCTION

In a non-randomized study (a.k.a. observational study), the assignment rule of participants to treatment conditions ($Z = 1$ for treatment; $Z = 0$ for control) is unknown. Participants in different treatment conditions can differ in their distributions of pre-treatment covariates (**X**) which are also related to the outcome (Hill, 2008). In order to provide causal inference of the average treatment effect for the outcome in a non-randomized study, Rosenbaum and Rubin (1983) defined the strong ignorability assumption based on the potential outcomes framework (e.g., Rubin, 2005; West, Cham, & Liu, in press; West & Thoemmes, 2010). The assumption involves two propositions: (1) The potential outcomes of a participant *i* in the treatment and control groups are conditionally independent of the treatment assignment *Z*, given all the *measured* pre-treatment covariates **X**. (2) The participant *i* has non-zero probabilities of being assigned to the treatment and control groups, given his / her covariates **X** values.

In order to reduce the high dimensionality of many covariates **X**, Rosenbaum and Rubin (1983) defined the propensity score $e(\mathbf{X})$, which is the probability of the participant *i* being assigned to the treatment group ($Z = 1$) given his / her covariates **X** values:

$$e_i(\mathbf{X}) = Pr(Z_i = 1 \mid \mathbf{X}_i) \qquad (1)$$

As long as the strong ignorability assumption (defined above) holds for the covariates **X**, the assumption also holds for the propensity score (Rosenbaum & Rubin, 1983, theorem 3). Given that the strong ignorability assumption holds, balance on participants'

propensity scores $e(\mathbf{X})$ between treatment conditions provides an unbiased average treatment effect estimate in the non-randomized study.

In practice, participants' propensity scores need to be estimated based on the covariates **X** values. Drake (1993) showed that a misspecified propensity score model produces a biased average treatment effect estimate. Linear logistic regression is typically used to estimate propensity scores; however, a variety of statistical learning methods have also been proposed, including Classification and Regression Trees (Luellen, Shadish, & Clark, 2005), Random Forests (Lee, Lessler, & Stuart, 2010), Generalized Boosted Modeling (McCaffrey, Ridgeway, & Morral, 2004), Neural Networks (Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008), and Support Vector Machines (Westreich, Lessler, & Funk, 2010). These statistical learning methods may be preferable to linear logistic regression when the true population propensity score model involves complex interactions, nonlinear relationships, or both of the covariates on treatment assignment.

This dissertation investigates the performance of the Random Forests method of estimating propensity scores. Two studies to date have studied the performance of this method in the propensity score analysis context (Austin, 2012; Lee et al., 2010). There is a lack of understanding of the effects of some of the important Random Forests model specifications in propensity score estimation. In this dissertation, I first provide an overview of the Random Forests method and explicate the key model specifications in propensity score analysis. The effects of the key model specifications are then studied through a simulation study.

Overview of Random Forests

Breiman (2001) developed the Random Forests method as an automatic, nonparametric procedure for addressing regression problems with complex interactions, nonlinear relationships, or both of many covariates on the outcome. Comprehensive introductions are provided by Berk (2008), Hastie, Tibshirani, and Friedman (2009), and Strobl, Malley, and Tutz (2009). This section introduces Random Forests in its propensity score estimation context using an empirical example from Im, Hughes, Kwok, Puckett, and Cerda (2013). Using logistic regression, they conducted propensity score analysis to equate students who were retained (held back) once during grades 1 to 5 (treatment group) with students who were never retained (control group). A total of 67 covariates were identified and measured before any retention occurred.

The Random Forests method involves several steps. First, multiple random samples are drawn from the data. Breiman (2001, theorem 2.1) proved that increasing the number of random samples does not overfit the Random Forests model. Second, a *Classification Tree* model (Breiman, Friedman, Stone, & Olshen, 1984; Morgan & Sonquist, 1963) of the covariates **X** on treatment assignment Z is estimated using the data from each random sample. Third, the participants' propensity scores are estimated from each Classification Tree model. Fourth, these propensity scores are averaged across all models to obtain the Random Forests propensity score for each participant.

Figure 1 illustrates a *Classification Tree* model based on the Im et al. (2013) data. At each step of the Classification Tree procedure, a *node* of a covariate (indicated by the ellipse with the covariate's name) is selected with a split value. Participants are then

classified into two nodes at the next level[1]. Consider the leftmost branch of the Classification Tree in Figure 1. First, participants who score ≤ -0.69 on the covariate *AVGZF1* (district literacy average *z* score) are classified into the node on the left at the next level. Then, participants who score ≤ 3.67 on the covariate *TACH1* (teacher perception of child achievement) in this node are further classified into the *terminal node* A at the bottom of the model. The bar chart of the terminal node A (n = 108 participants) represents the proportion of the treatment group participants (proportion = 0.565; marked as 1, dark bar) and control group participants (proportion = 0.435; marked as 0, light bar) in this terminal node. The participants who are classified into this terminal node are given propensity scores equal to the proportion of the treatment group participants in this terminal node (proportion = 0.565).

[1] Although splitting into more than two nodes is possible, it is not recommended because this exhausts the search too quickly (Hastie et al., 2009, p. 311).

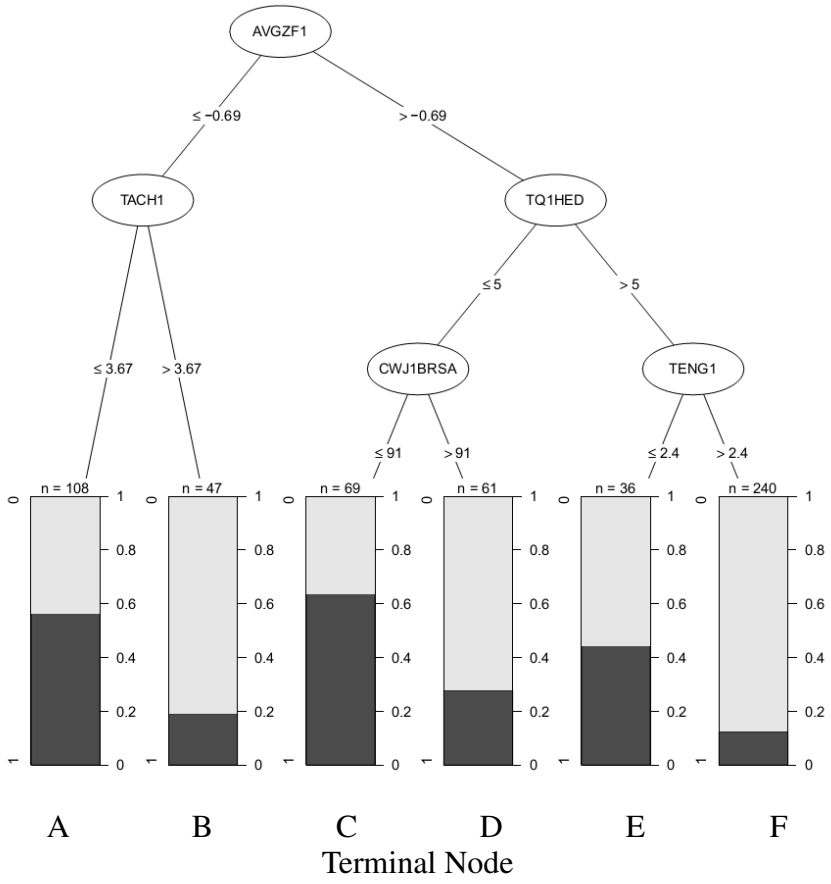


Figure 1. An illustration of Classification Tree model. The ellipses are the nodes of covariates which classify the participants into two nodes at the next level. In each terminal node (A to F) at the bottom of the model, a bar chart is displayed. The number of participants in each terminal node is shown at the top of the bar. The bar chart represents the proportion of treatment group participants (= 1) using a dark bar and the proportion of control group participants (= 0) using a light bar in each terminal node. *AVGZF1* is district literacy average *z* score, *TACH1* is teacher perception of child achievement, *TQ1HED* is teacher expected students' highest education level, *CWJ1BRSA* is Woodcock-Johnson III broad reading standard score (Woodcock, McGrew, & Mather, 2001), *TENG1* is teacher perception of child school engagement.

Key Random Forests Model Specifications

The section explicates some key Random Forests model specifications applicable in propensity score estimation.

Covariate and split value selection. There are two types of decision rules (criteria) for selecting the covariate and its split value. The first type is to search for a covariate and its split that maximizes the reduction of the *impurity measure* after splitting. The Gini index is a common choice of the impurity measure (Berk, 2008, equation 3.6; Hastie et al., 2009, equation 9.17).

$$\text{Gini Index} = 2p\,(1 - p) \qquad (2)$$

where p is the proportion of the treatment group participants in the node. One property of the Gini index is that it has a tendency to produce one node that has higher proportion of treatment group participants than control group participants, and one node that has relatively balanced proportions of treatment and control group participants (Berk, 2008, p. 114-115; Breiman et al., 1984, p. 111; Hastie et al., 2010, p. 309). An important disadvantage of the Gini index is that it is biased towards selecting covariates with many categories and continuous covariates rather than binary covariates in which the data contain different types of covariates (Berk, 2008, pp. 137, 150; Hastie et al., 2009, p. 310; Hothorn, Hornik, & Zeileis, 2006; Strobl et al., 2009, p. 342; see Breiman et al., 1984; Kim & Loh, 2001; Strobl, Boulesteix, & Augustin, 2007; White & Liu, 1994). In other words, the binary covariates in the mixture of different types of covariates are more likely to be ignored when building the trees. One reason underlying this disadvantage is that the Gini index is used to select the covariate and its split value simultaneously (Hothorn, Hornik, et al., 2006, p. 4).

The second type of decision rule is the conditional significance test (Hothorn, Hornik, et al., 2006). At the node that is being considered for splitting, a permutation test is performed separately for each covariate. The null hypothesis is that the specific covariate is *not* associated with the treatment groups Z given all other covariates. The covariate that rejects the null hypothesis with the smallest p-value of the significance test is selected. Two test statistics are available. The first test statistic is the maximum of a z-score statistic (in absolute value) across the selected covariate values, which is asymptotically normally distributed. The second test statistic has a quadratic form which is asymptotically χ^2 distributed. The computation of the second test statistic is less time-consuming than the first test statistic. To adjust the Type I error rate for multiple comparisons, Bonferroni or Monte Carlo resampling adjustments may be performed[2]. If the smallest p value is greater than the adjusted specified α level, no splitting occurs. After selecting a covariate, a second permutation test of the null hypothesis that the proportions of treatment group participants are equal between the two nodes is conducted across the selected covariate values. The covariate value with the largest test statistic is selected. Because separate conditional significance tests are used for selecting the covariate and for choosing the split value, this decision rule reduces the covariate selection bias relative to the Gini index (Hothorn, Hornik, et al., 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007).

I hypothesized that the conditional significance test decision rule would be preferred to the Gini index in the propensity score estimation context for two reasons. (1)

[2] In addition to the Bonferroni and Monte Carlo resampling adjustments, Hothorn, Hornik, et al. (2006) developed a global permutation test of the null hypothesis that all covariates are not associated with the treatment conditions. If this test is non-significant, there is no splitting. This test is "less attractive for learning samples [data used to estimate the Classification Tree] with missing values" (Hothorn, Hornik, et al., 2006, pp. 5-6).

The conditional significance test selects the covariate that is most associated with the treatment conditions, which is intuitively reasonable for propensity score estimation. (2) The conditional significance test reduces the bias towards selecting particular types of covariates. These features may reduce the bias of the average treatment effect estimate.

Random sampling of covariates for selection. When deciding whether to split a node, the analyst can choose a random subset of the covariates instead of all covariates. This strategy has the following properties. First, it reduces the sampling variability of the Random Forests propensity scores (Berk, 2008, pp. 194-198; Breiman, 2001; Hastie et al., 2009, p. 588; Strobl et al., 2009, p. 333). Second, it increases the chance of selecting covariates that are (1) more associated with other covariates but less associated with the treatment conditions (Berk, 2008, pp. 194-198; Strobl et al., 2009, p. 333), and (2) less associated with the treatment conditions but more associated with the outcome in propensity score analysis relative to other covariates. Based on these two properties, I hypothesized that a random subset of the covariates would produce a less biased average treatment effect estimate.

Once the decision has been made to randomly sample the covariates, the next step is to determine the number of covariates to be randomly sampled. Hastie et al. (2009, p. 592), Liaw and Wiener (2002), and Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008) suggested that the number of covariates in the random subset for covariate and split selection should be equal to the square root of the total number of covariates. Strobl et al.'s (2008) simulation study found that increasing the number of covariates in the subset can reduce covariate selection bias when the set includes some highly correlated covariates (see also Hastie et al., 2009, p. 601). Preliminary exploration suggested that

the performance of the Random Forests propensity scores is sensitive to the number of covariates to be sampled, depending heavily on the specific research question and the nature and quality of the data. Given the lack of prior theoretical or empirical work, I offered no hypothesis here.

Methods for estimating Random Forests propensity scores. There are two methods of using the Classification Tree model from each random sample of the data to estimate participants' overall propensity scores (Berk, 2008, pp. 189, 193; Hastie et al., 2009, pp. 592-593; Strobl et al., 2009, p. 335). In the first method, *all* participants in the original data are entered into *all* the Classification Tree models to estimate their propensity scores. In the second method, only the participants who are *not* in the random sample are entered into the Classification Tree model that is estimated by this random sample to obtain their propensity scores. This group of participants that are *not* in the random sample is also known as the *out-of-bag* sample (a.k.a. hold out sample). The second method distinguishes those participants who are used to estimate the Classification Tree model from those participants whose propensity scores are estimated by that Classification Tree model (i.e., cross-validation sample). The second method reduces the tendency for propensity scores to be estimated that are biased toward the extreme values of 0.0 and 1.0 (Berk, 2008, p. 189; Strobl et al., 2009, p. 335). In propensity score analysis, this is favorable because this procedure typically produces propensity scores with a larger *common support region* (overlap) between the two treatment groups. A larger common support region increases the generalizability of the propensity score analysis. I hypothesized that the out-of-bag sample would produce a less biased average treatment effect estimate.

Other Random Forests Model Specifications

This section discusses other Random Forests model specifications that are not examined (i.e., held constant) in the simulation studies proposed in this dissertation. In my dissertation I focus on more generally applicable specifications rather than specifications that heavily depend on the specific research question and the nature and quality of the data.

Methods of drawing random samples. Random sampling of the data can be conducted (1) with or without replacement, and with (2) a designated proportion of the random sample size to original data sample size. Strobl, Boulesteix, Zeileis, et al.'s (2007) simulation study shows that (1) sampling without replacement and (2) obtaining random samples which are 0.632 times the sample size of the original data reduces the covariate selection bias towards covariates with many categories and continuous covariates.

Maximum depth of Classification Tree. The Classification Tree model in figure 1 has a depth of three. This depth allows a maximum of three-way interactions of the covariates. Strobl et al. (2009) discussed the meaning of interactions in Classification Tree in detail. Lee et al. (2010) and Strobl, Boulesteix, Zeileis, et al.'s (2007) simulation studies suggested no restrictions on the maximum depth of the tree model.

Minimum terminal node size. Hastie et al. (2009, p. 592) and Liaw and Wiener (2002) suggested that the minimum terminal node size should equal 1.0. This outcome is *not* favorable in propensity score estimation, because this specification eventually produces propensity scores that approach either 0.0 or 1.0. In other words, the propensity score common support region (overlapping) is greatly reduced.

Propensity Score Matching and Weighting

Matching, weighting, subclassification, and analysis of covariance methods can utilize participants' estimated propensity scores to estimate the average treatment effect (see Stuart, 2010 for a review). In this dissertation, the nearest neighbor matching (Rubin, 1973) and weighting by the odds (Hirano, Imbens, & Ridder, 2003) adjustment methods were used in the simulation study[3]. This section briefly describes these two methods.

Nearest neighbor matching method matches a treatment group participant with the control group participant(s) with the smallest difference between their estimated propensity scores. There are several options associated with this method. First, the analyst needs to decide whether the control group participants can be matched with or without replacement. For the without replacement option, the analyst needs to choose the order of the treatment group participants to be matched (starting from the participant with the largest, smallest, or proceeding in a random order of the estimated propensity scores). Second, the analyst needs to decide the number of control group participants to be matched with one treatment group participant. Third, the analyst may specify a matching caliper to prohibit matches if the difference is too large. These options are described in Dehijia and Wahba (2002) and Stuart (2010).

The weighting by the odds method provides the weighting adjustment of the treatment and control group participants for the average treatment effect estimation. All treatment group participants have a weight equal 1.0, whereas the control group participants have a weight equal $\hat{e}(\mathbf{X})/(1-\hat{e}(\mathbf{X}))$, where $\hat{e}(\mathbf{X})$ are the participants'

[3] Technically speaking, the nearest neighbor matching and weighting by the odds methods estimate the average treatment effect on the treated, which is the average treatment effect for the participants who are actually assigned to the treatment group. For the nearest neighbor matching method, when there are any treatment group participants who are not successfully matched with any control group participants, the effect obtained is not exactly the average treatment effect on the treated.

estimated propensity scores. To estimate the average treatment effect, participants' weights need to be accounted for. One approach is to use the survey sampling technique to compute the parameter estimates (e.g., weighted least squares estimation), standard errors (e.g., Taylor series linearization estimator), and significance tests that accounts for these weights (Lohr, 2010).

Primary Research Questions

This dissertation investigates the three key Random Forests model specifications described above in the context of propensity score analysis through a simulation study. In addition to the Random Forests model specifications that were not investigated by Austin (2012) or Lee et al. (2010), this dissertation investigates the performance of both the nearest neighborhood matching and weighting by the odds methods of estimating the average treatment effect. Based on the presentation above, the following findings concerning the average treatment effect estimation and the balance on covariates between treatment conditions were expected.

Average treatment effect estimation. Given that strong ignorability is met in the population, I hypothesized that there would be one or more optimal combination(s) of the Random Forests model specifications that would produce a minimally biased average treatment effect estimate. Based on the previous discussion, I hypothesized that the least biased average treatment effect estimates would be found under the conditional significance test decision rule, random sampling of covariates, and for out-of-bag samples This finding was expected to hold both for the nearest neighbor matching and weighting by the odds methods. I also investigated the standard error and statistical inference of the average treatment effect under different Random Forests model

specifications. The goal was to compare the average treatment effect estimate, standard error, and statistical inference across different Random Forests model specifications.

Balance on covariates' distributions between treatment conditions. The performance of estimated propensity scores is often evaluated by assessing the balance on the distribution of the set of covariates between treatment conditions after propensity score adjustment (e.g., Rubin, 2001; Stuart, 2010; West, Cham, Thoemmes, Renneberg, Weiler, & Schulze, 2013). It has been assumed that improved balance between the groups on the distributions of the covariates is associated with decreased bias in the average treatment effect estimate. However, a simulation study Lee et al. (2010) showed that, when comparing *between* different propensity score estimation methods, better balance on the covariates' marginal means does *not* necessarily relate to decreased bias of the average treatment effect estimate. Based on their findings, I investigated two questions:

(A) *Within* a certain Random Forests model specification, does the balance on the covariates' marginal means relate to the magnitude of bias of the average treatment effect estimate across repeated samples? If the answer is "yes", researchers can compare the balance on the covariates between independent data sets to decide which sample data set will lead to the least biased average treatment effect estimate.

(B) *Between* different Random Forests model specifications, does the balance on the covariates' marginal means relate to the magnitude of bias of the average treatment effect estimate? If the answer is "yes", researchers can compare the balance of the covariates between different Random Forests model specifications to decide which specification will lead to the least biased average treatment effect estimate.

Chapter 2

SIMULATION STUDY DESIGN

This chapter describes the design of a simulation study to address the primary research questions.

Population Propensity Score Analysis Models

The population propensity score analysis models for the simulation study were constructed based on an empirical example (Im et al., 2013), as well as simulation studies by Austin (2012), Lee et al. (2010) and Setoguchi et al. (2008).

Empirical example. As described in chapter 1, Im et al. (2013) equated students who were retained (held back) once during grades 1 to 5 (treatment group) with students who were never retained (control group). In the example, there were 67 pre-treatment covariates (19 binary, 45 continuous, and 3 ordered-categorical). In order to use their data as a basis for constructing the population propensity score analysis models, the empirical example data were preprocessed using the following procedures. First, missing values of the covariates were filled in by single imputation using the chained equations imputation procedure through IVEware 0.2 (Survey Research Center, University of Michigan, 2011). Second, the binary covariates were dummy coded 0.0 and 1.0. Third, the continuous covariates were standardized. Fourth, the ordered-categorical covariates were rescaled so that 0 indicated the lowest category. Fifth, participants' Woodcock-Johnson III broad reading W scores (Woodcock et al., 2001) in grade 6 were regarded as the outcome.

Covariates in population models. In the population models, a total of 64 covariates (16 binary, 40 continuous, 8 ordered-categorical) were randomly generated using the R 2.15.3 (R Core Team, 2013) package BinNor 2.0 (Amatya & Demirtas, 2012;

Demirtas & Doganay, 2012). The numbers of each type of covariate were chosen to roughly approximate those in the empirical example (19 binary, 45 continuous, and 3 ordered-categorical), and provided a balanced incomplete factorial design of the covariates in the population models (see below). The binary covariates (b_1 to b_{16}) were dummy coded 0.0 and 1.0 with a mean of 0.245, which was the average of the binary covariates' means in the empirical example. The continuous covariates (c_1 to c_{40}) were generated to be standard normally distributed. The ordered-categorical covariates (o_1 to o_8) were categorized from randomly generated standard normally distributed variables. The ordered-categorical covariates had 7 categories (0 to 6) and were generated to be approximately symmetrically distributed. Within Random Forests, the specification of the random sampling of covariates affects the chances that covariates with different correlations with other covariates would be selected to estimate the Classification Tree model. Consequently, the covariates were manipulated to have two levels of correlation (low versus high). The correlations between two low level covariates were set equal to the 10^{th} percentile of the magnitude of the covariates' correlations (in absolute values) in the empirical example. The correlations between two high level covariates were set equal to the 90^{th} percentile of the magnitude of the covariates' correlations in the empirical example. The correlations between the low and high level covariates were set equal to the 50^{th} percentile of the covariates' correlations in the empirical example. Note that the correlations of the ordered-categorical covariates did not exactly match the above values because of the categorization procedure. Table 1 shows the population parameters and the bottom panel shows the manipulations of the 64 covariates.

Propensity score models. Based on simulation study designs by Austin (2012), Lee et al. (2010), and Setoguchi et al. (2008), there were two population propensity score models. The first model was termed the *linear propensity score model*, in which all the covariates are linearly related to the propensity scores in the form of logistic regression below.

$$\ln\left(\frac{e(\mathbf{X})}{1-e(\mathbf{X})}\right) = \gamma_0 + \gamma_1 b_1 + \ldots + \gamma_{17} c_1 + \ldots + \gamma_{64} o_8 \quad (3)$$

$\ln(\cdot)$ is the natural logarithmic function, and $e(\mathbf{X})$ is the propensity score. The regression coefficients γ of the linear propensity score model in equation 3 were assigned based on the estimated linear propensity score model using the empirical example. The regression coefficients of the covariates (γ_1 to γ_{64}) were manipulated to have two levels (low versus high) separately for each type of covariate (binary, continuous, and ordered-categorical).

To determine the value of the regression coefficients γ, two criteria were applied in a large simulated dataset (n = 100,000) generated by the propensity score model of the γ values. First, the propensity scores distribution should not be extremely peaked at 0.0 and 1.0, with too few observations in between. Second, the propensity score model should be estimable with correct γ estimates and without error warnings using R glm procedure. The low level regression coefficients were set to the 1st, 5th, and 5th percentile of the estimated regression coefficients in the empirical example for the binary, continuous, and ordered-categorical covariates, respectively. The high level regression coefficients were set to the 5th, 40th, and 70th percentile of the estimated regression coefficients in the empirical example for the binary, continuous, and ordered-categorical covariates. The intercept γ_0 in equation 3 was chosen so that the average propensity score was about 0.38, which was approximately the proportion of the treatment group

participants in the empirical example (proportion = 0.32). (For the population parameters, see Table 1).

The second propensity score model was termed the *nonlinear propensity score model*, which added terms to the linear propensity score model above. Three types of nonlinear relationships of the covariates were added.

(A) *Two-way interaction only*. One pair of covariates of each covariate type combination (binary, ordered-categorical, continuous) was randomly selected to produce a two-way interaction.

(B) *Quadratic effect only*. Three *continuous* covariates were randomly selected to produce quadratic effects.

(C) *Two-way interaction and quadratic effects*. One pair of covariates of each combination of covariate type (binary, ordered-categorical, continuous) was randomly selected to produce the two-way interaction and quadratic effects of the two covariates. No quadratic effects were produced for the binary covariates for (C). No covariates were randomly selected more than once to produce (A) to (C). The regression coefficients γ of these two-way interactions and quadratic effects were random numbers that follow the uniform distribution between -0.1 and 0.1.

In the two propensity score models, the participants' treatment assignment Z was chosen based on their propensity scores $e(\mathbf{X})$. Each participant was assigned a random number that follows a uniform distribution between 0.0 and 1.0. Participants were assigned to the treatment group ($Z = 1$) if their random numbers are smaller than their propensity scores. Otherwise, the participants were assigned to the control group ($Z = 0$).

Based on the two propensity scores models, the covariates differed primarily between the treatment conditions in their marginal means (Lee et al., 2010).

Treatment-outcome model. The outcome *Y* was a continuous variable. Based on the simulation study designs by Austin (2012), Lee et al. (2010) and Setoguchi et al. (2008), the treatment assignment *Z* and covariates were linearly related to the outcome as follows.

$$Y = \beta_0 + \beta_1 b_1 + \ldots + \beta_{17} c_1 + \ldots + \beta_{64} o_8 + (ATE)\, Z + \varepsilon \qquad (4)$$

Z is the treatment assignment (0 for control and 1 for treatment) and ε is the residual which was randomly generated to be normally distributed with mean of 0.0 and variance of σ^2. The intercept β_0 and residual variance σ^2 were chosen based on the estimated treatment-outcome model using the empirical example. The regression coefficients of the covariates (β_1 to β_{64}) were separately manipulated to have two levels (low versus high) for each type of covariate (binary, continuous, and ordered-categorical). In a similar fashion, a large simulated dataset (n = 100,000) based on the empirical example was generated to determine the regression coefficients β_1 to β_{64} using the R lm procedure. The low and high level regression coefficients were set to the 10th and 90th percentile of the estimated regression coefficients in the empirical example for the binary, continuous, and ordered-categorical covariates, respectively. The *unstandardized* average treatment effect (*ATE*) was manipulated to have two levels: a null average treatment effect (*ATE* = 0.0) and a non-zero average treatment effect (*ATE* = 0.73). The non-zero average treatment effect approximated a moderate effect size in the treatment-outcome model according to Cohen's (1988) guidelines.

Table 1

Summary of Population Propensity Score Analysis Models

(1) Correlations of the covariate with other covariates

A. Correlations between two low level covariates — $r = 0.0142$

B. Correlations between two high level covariates — $r = 0.2748$

C. Correlations between the low level and high level covariates — $r = 0.0801$

(2) Regression coefficients of the propensity score models

A. Intercept of the propensity score models — $\gamma_0 = -3.2813$

B. Regression coefficients of the covariates (γ_1 to γ_{64}) of the linear propensity score model (equation 3)

Low		High	
Binary (b_1-b_{16})	$\gamma = 0.0427$	Binary (b_1-b_{16})	$\gamma = 0.1581$
Continuous (c_1-c_{40})	$\gamma = 0.0206$	Continuous (c_1-c_{40})	$\gamma = 0.1717$
Ordered-categorical (o_1-o_8)	$\gamma = 0.0388$	Ordered-categorical (o_1-o_8)	$\gamma = 0.1359$

C. Regression coefficients of the two-way interactions and quadratic effects of the covariates in the nonlinear propensity score model

Two-way interaction only		Quadratic effect only		Two-way interaction and quadratic effects			
	γ		γ	Interaction	γ	Quadratic	γ
$b_1 \times b_7$	0.0794	c_{22}^2	-0.0018	$b_9 \times b_6$	-0.0830		
$b_5 \times o_6$	0.0471	c_{18}^2	-0.0323	$b_{14} \times o_8$	-0.0538	o_8^2	-0.0462
$b_{12} \times c_{10}$	0.0387	c_1^2	-0.0917	$b_{10} \times c_{39}$	-0.0790	c_{39}^2	-0.0388
$o_7 \times o_2$	-0.0429			$o_5 \times o_1$	0.0672	o_5^2	0.0385
$o_4 \times c_9$	0.0378					o_1^2	0.0184
$c_6 \times c_{19}$	0.0896			$o_3 \times c_{21}$	0.0600	o_3^2	0.0355
						c_{21}^2	0.0709
				$c_5 \times c_{16}$	0.0110	c_5^2	-0.0897
						c_{16}^2	-0.0360

(3) Regression coefficients of the treatment-outcome model

A. Intercept of the treatment-outcome model — $\beta_0 = -0.2743$

B. Residual variance — $\sigma^2 = 0.5781$

C. Regression coefficients of the covariates (β_1 to β_{64}) of the treatment-outcome model (equation 4)

Low		High	
Binary (b_1-b_{16})	$\beta = 0.0153$	Binary (b_1-b_{16})	$\beta = 0.1673$
Continuous (c_1-c_{40})	$\beta = 0.0067$	Continuous (c_1-c_{40})	$\beta = 0.1098$
Ordered-categorical (o_1-o_8)	$\beta = 0.0293$	Ordered-categorical (o_1-o_8)	$\beta = 0.0890$

D. Zero average treatment effect (*ATE*) — $ATE = 0.0$

Non-zero average treatment effect (*ATE*) — $ATE = 0.73$

Summary: Resulting manipulations of the covariates on the factors (1) to (3) above

	(2) Low		(2) High	
	(3) Low	(3) High	(3) Low	(3) High
(1) Low	b_1-b_2; o_1; c_1-c_5	b_3-b_4; o_2; c_6-c_{10}	b_5-b_6; o_3; c_{11}-c_{15}	b_7-b_8; o_4; c_{16}-c_{20}
(1) High	b_9-b_{10}; o_5; c_{21}-c_{25}	b_{11}-b_{12}; o_6; c_{26}-c_{30}	b_{13}-b_{14}; o_7; c_{31}-c_{35}	b_{15}-b_{16}; o_8; c_{36}-c_{40}

Note. Manipulated factors on the covariates are (1) correlations of the covariate with other covariates, (2) regression coefficients of the propensity score models, and (3) regression coefficients of the treatment-outcome model.

Study Design

This section summarizes the design and implementation of the simulation study.

(1) Propensity score models. This between-subject factor had two levels, linear and nonlinear propensity score models.

(2) Magnitude of the average treatment effect. This between-subject factor had two levels, a null (*ATE* = 0) and non-zero (*ATE* = 0.73) *unstandardized* average treatment effect. When standardized, these average treatment effects represented a null and an approximately moderate effect size according to Cohen's (1988) normative values.

(3) Sample size. This between-subject factor had two levels, 600 and 2000. *N* = 600 approximated the sample size of the empirical example (*n* = 561). *N* = 2000 was used in Lee et al. (2010) and Setoguchi et al. (2008).

(4) Benchmark methods. Three benchmark methods were used to evaluate the different Random Forests model specifications described below: (a) no adjustment for covariates (uncorrected), (b) propensity score analysis using the estimated logistic regression propensity scores based on the true population propensity scores models (linear or nonlinear), (c) analysis of covariance (ANCOVA) using the true population treatment-outcome model. The benchmarks using the logistic regression propensity scores and ANCOVA represent the ideal performance if the true population model were known, not the actual performance that can be achieved by these methods in practice. These benchmark methods provide "within-subjects" (replication sample) comparisons.

The Random Forests model specifications used to estimate the propensity scores are described below. These factors were within-subjects.

(5) Covariate and split value selection criterion. Two criteria for the covariate and split value selection were used: (a) the Gini index decision rule was implemented through the R package ipred 0.9-1 (Peters & Hothorn, 2012); (b) the conditional significance test decision rule was implemented through R package party 1.0-6 (Hothorn, Bühlmann, Dudoit, Molinaro, & Van Der Laan, 2006; Strobl et al., 2008; Strobl, Boulesteix, Zeileis, et al., 2007). For the conditional significance test, the specification suggested by Strobl, Boulesteix, Zeileis, et al. (2007) was adopted – the quadratic form test statistic (asymptotically χ^2 distributed) without multiple testing adjustment.

(6) Random sampling of covariates for selection. This factor had three levels, (a) no random sampling (i.e., all covariates), (b) random sampling eight covariates, and (c) random sampling four covariates. The value of eight is suggested by Hastie et al. (2009, p. 592), Liaw and Wiener (2002), and Strobl et al. (2008), and equals the square root of the number of covariates (= $\sqrt{64}$). The value of four is based on a preliminary analysis of pilot data from the simulation study.

(7) Methods for estimating propensity scores. This factor had two levels, the use of (a) the full sample versus (b) the *out-of-bag* sample to estimate participants' propensity scores using each Classification Tree model.

Two different methods were also used to equate the two groups (treatment and control). Again, this factor was within-subjects.

(8) Methods of equating groups on propensity scores. This factor had two levels, (a) nearest neighbor matching and (b) weighting by the odds.

The nearest neighbor matching method was implemented through the R package MatchIt 2.4-20 (Ho, Imai, King, & Stuart, 2011). Matching without replacement used the

participants' propensity scores in a random order. One treatment participant was matched with one control participant. A caliper of 0.25 times the standard deviation of the estimated propensity scores was used to prohibit poor matches (Rosenbaum & Rubin, 1985; Stuart, 2010). To estimate the average treatment effect after matching, the outcome *Y* was linearly regressed on the treatment assignment *Z* using the R lm procedure.

To implement the weighting by the odds method, the outcome *Y* was linearly regressed on treatment assignment *Z* using the R package survey 3.29 (Lumley, 2004; 2012). This package incorporates survey sampling techniques to account for participants' weights (Lohr, 2010).

All other Random Forests model specifications were held constant in all conditions. These specifications included: (a) each Random Forests model had 500 Classification Tree models; (b) random sampling of the data without replacement was conducted using 0.632 times the sample size of the original data, the proportion suggested by Strobl, Boulesteix, Zeileis, et al. (2007); (c) the maximum depth of Classification Tree was not controlled, suggested by the results of Lee et al. (2010) and Strobl, Boulesteix, Zeileis, et al. (2007)[4]; and (d) the minimum terminal node size was set to five based on preliminary analysis of pilot data from the simulation study.

Table 2 summarizes the manipulated factors and factor levels of the simulation study design and presents the notation that will be used in later chapters of the dissertation. The resulting simulation study had a balanced, incomplete factorial design. The R package ipred does not offer random sampling of covariates for covariate and split

[4] Nevertheless, the R package ipred is limited to a maximum depth of 30, whereas the R package party does not have this limitation.

value selection[5]. As a result, the factors (5) covariate and split value selection and (6) random sampling of covariates for selection could not be fully crossed. Each condition was examined with 1,000 replications.

[5] There is another R package randomForest (Liaw & Wiener, 2002) that uses the Gini index decision rule and allows random sampling of covariates. However, it uses a different method of calculating the Random Forests propensity score. This method classifies the participants into treatment and control groups in each Classification Tree. The Random Forests propensity score is the proportion of the number of times being classified into the treatment group across all Classification Trees (A. Liaw, personal communication, July 24, 2012). This method of calculating the propensity score is *not* recommended (Boström, 2007; Hastie et al., 2009, p. 283). Criticisms of this method center on the rounding errors that occur from the estimated propensity score to treatment assignment decision for each participant within each Classification Tree.

Table 2

Manipulated Factors and Factor Levels of Simulation Study Design

Between-subject (Replication) Factors	Levels and Notations
1. Propensity Score Models	Linear, Nonlinear
2. Magnitudes of Average Treatment Effect	0, 0.73
3. Sample Size *N*	600, 2000
Within-subject (Replication) Factors	
4. Benchmark Methods	No adjustment (Uncorrected), Logistic Regression using the Population Propensity Score Model (Logistic), ANCOVA using the Population Treatment-outcome Model (ANCOVA)
5. Covariate and Split Value Selection	Gini Index (Gini), Conditional Significance Test (Sig)
6. Random Sampling of Covariates for Selection	No Random Sampling (NS), Randomly Sample 4 Covariates (S4), Randomly Sample 8 Covariates (S8)
7. Methods to Estimate Propensity Scores	Full Sample (F), Out-of-bag Sample (O)
8. Methods of Equating Groups	Matching, Weighting

Note. 0 and 0.73 represents the unstandardized effect sizes. When standardized, these correspond to a null effect and a moderate effect size in terms of Cohen's (1988) norms.

Chapter 3

RESULTS

To answer the primary research questions, the simulation study results are presented in three major sections: model convergence rates, average treatment effect estimation, and balance on covariates between treatment conditions. To facilitate the presentation, I will use the notation presented in Table 2.

Model Convergence Rates

A replication was defined as having converged when the estimated standard error of the average treatment effect estimate was positive, and the nearest neighbor matching was conducted without error messages. Only one replication in the entire simulation study failed to converge (Condition: $N = 600$, nonlinear model, when the Sig.O.NS random forest propensity scores and weighting were used). This non-converged replication was replaced by an independent, converged replication.

Average Treatment Effect Estimation

The goal of propensity score analysis is to produce an unbiased estimate of the average treatment effect (*ATE*), which permits causal inference in a non-randomized study, given the strong ignorability assumption. Figure 2 presents the results for the average treatment effect estimation. Figure 2 consists of four plots partitioned by sample size N (600, 2000) and the magnitude of population average treatment effect ($ATE = 0$ or 0.73). In each plot, there are 40 panels. The panels are arranged into 4 rows of propensity score models (linear, nonlinear) and methods of equating groups (matching and weighting) (see the rightmost axis labels). The panels are also arranged into 10 columns representing the two benchmark methods (ANCOVA, Logistic regression) and each of

the different Random Forests model specifications for estimating propensity scores (see the top axis labels). Recall that the ANCOVA and logistic regression propensity scores benchmark methods represent the *ideal* performance of the method under conditions in which the variables and the functional form of the estimated model correspond exactly to the population model; these results will *not* typically be obtained in practice. The uncorrected *ATE* estimation results were not displayed in Figure 1 but described in the text and presented in Appendix A.

In each panel there are two bar plots of the *ATE* estimate. The bold horizontal line in the middle of each bar plot is the mean *ATE* estimate across replications. The dashed horizontal line is the population *ATE* value (0 or 0.73). The numerical value *above* each bar plot is the mean *ATE* estimate across replications. In each column the bar plot on the *left* reflects ± 1 standard error of the *ATE* estimate (i.e., standard deviation of the *ATE* estimate across replications; labeled SD on the bottom axis). The bar plot on the *right* reflects ± 1 mean of the estimated standard error of the *ATE* estimate across replications (labeled SE on the bottom axis). The numerical value in parenthesis *under* each bar plot is the corresponding standard error value. Figure 2 presents two important aspects of the simulation results of *ATE* estimation: the bias of the *ATE* estimate and the standard error of the *ATE* estimate. The pattern of the results in Figure 2 was consistent between linear and nonlinear propensity score model, as well as between *ATE* = 0 and 0.73. However, note that the pattern of the results was different between sample size *N* = 600 and 2000. Below I focus on the results for the linear propensity score model, *ATE* = 0.73, and *N* = 600 and 2000.

Bias of average treatment effect estimate. If the bold horizontal line of the bar plot lies on the dashed horizontal line, the *ATE* estimate is unbiased. If the bold line is above the dashed line, the *ATE* is overestimated. If the bold line is below the dashed line, the *ATE* is underestimated. To supplement the graphical presentation, the percent bias of the *ATE* estimate was also computed. The percent bias of the *ATE* estimate is the mean difference between the *ATE* estimate and its population value across replications, with respect to the population value.

$$\text{Percent Bias of } ATE \text{ Estimate} = \frac{\sum_{i=1}^{Q}(\widehat{ATE} - ATE)/Q}{ATE} \quad (5)$$

$\widehat{ATE}$ is the average treatment effect estimate. $\Sigma(\cdot)$ is the summation function across Q replications. The percent bias is calculable only when *ATE* = 0.73. Percent bias of 0.0 indicates that the *ATE* estimate is unbiased. I loosely used Flora and Curran's (2004) suggestion of < 10% as a criterion for satisfactory percent bias. The complete results of the percent bias for the *ATE* estimate are presented in Appendix A.

For *N* = 600, for the linear model, the uncorrected *ATE* estimate seriously overestimated the population value (percent bias = 205.4%). As expected, ANCOVA using the true population treatment-outcome model produced an unbiased *ATE* estimate (percent bias = 0.38%). The logistic regression propensity scores using the true population propensity score model produced an unbiased *ATE* estimate when matching (percent bias = 10.3%) or weighting[6] (percent bias = 4.9%) adjustments were used. For the Random Forests propensity scores, the *ATE* estimate was influenced by the model specifications of (i) covariate and split value selection, (ii) random sampling of covariates for selection, (iii) methods for estimating Random Forests propensity scores, and (iv)

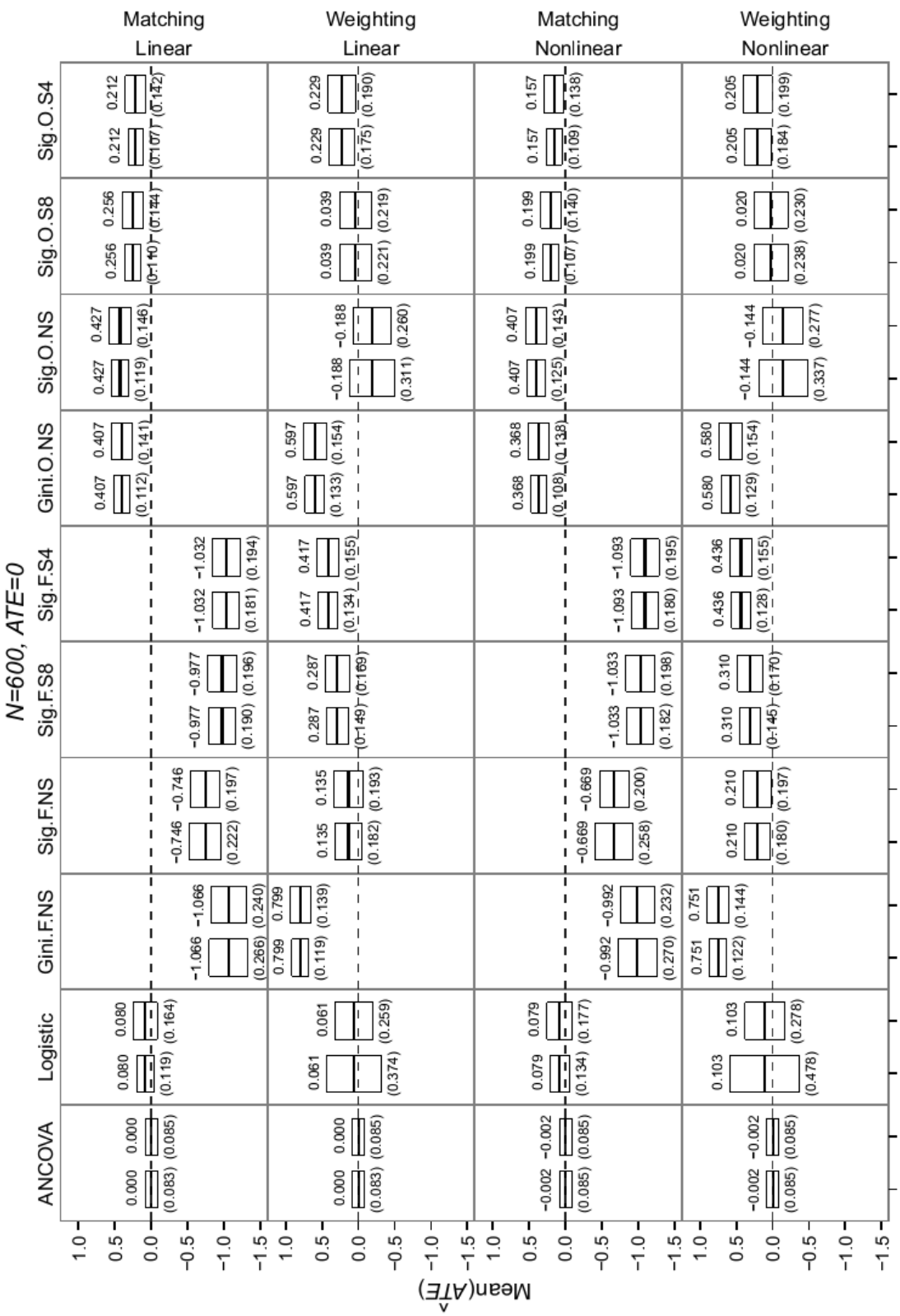
[6] In the nonlinear propensity score model, the logistic regression propensity scores using the weighting adjustment was overestimated (percent bias = 17.1%).

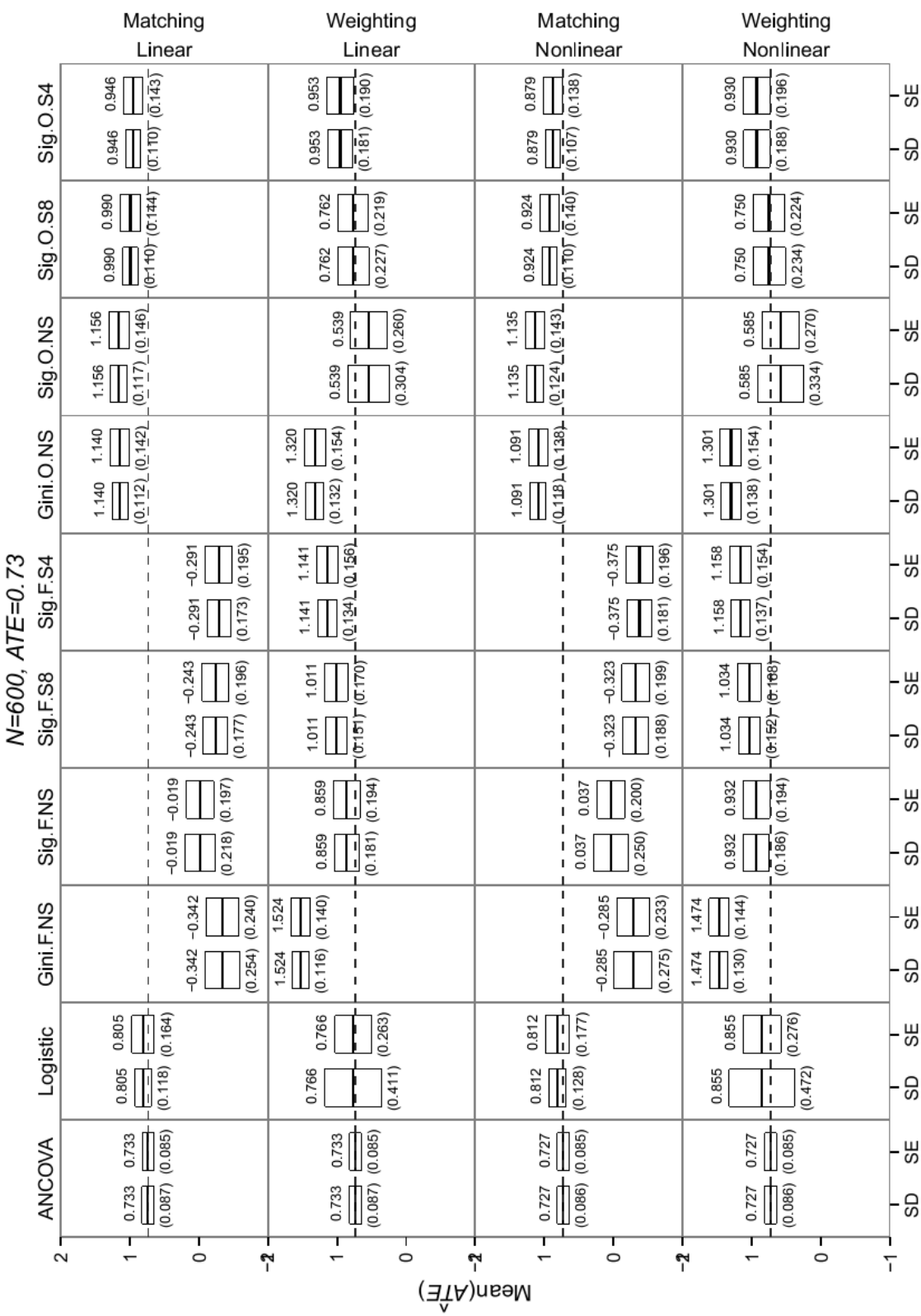
methods of equating groups on propensity scores. First, under the full sample specification, all Random Forests propensity scores underestimated the *ATE* when matching was used, whereas they overestimated the *ATE* when weighting was used. Under the full sample specification, the Gini index (Gini.F.NS) propensity scores estimated the *ATE* with a larger magnitude of bias than the conditional significance test (Sig.F.NS) propensity scores. Under the full sample and conditional significance test (Sig) propensity scores, no sampling of covariates (NS) produced the least biased but still unsatisfactory *ATE* estimate (Matching: percent bias = -102.6%; Weighting: percent bias = 17.7%), followed by sampling 8 (S8) and 4 (S4) covariates in ascending order. Second, under the out-of-bag sample specification and matching adjustment, all Random Forests propensity scores overestimated the *ATE*. In this situation, the Gini index (Gini.O.NS) and the conditional significance test (Sig.O.NS) propensity scores both overestimated *ATE* with similar magnitudes. Among the conditional significance test (Sig) propensity scores in this situation, sampling 4 covariates (S4) produced the least biased but still unsatisfactory *ATE* estimate (percent bias = 29.7%), followed by sampling 8 (S8) and no sampling (NS) in ascending order. Third, under the out-of-bag sample specification and weighting adjustment, the Gini index (Gini.O.NS) propensity scores estimated *ATE* with a larger magnitude of bias than the conditional significance test (Sig.O.NS) propensity scores. Among the conditional significance test (Sig) propensity scores in this situation, sampling 8 covariates (S8) produced an unbiased *ATE* estimate (percent bias = 4.4%). No sampling (NS) propensity scores underestimated *ATE* (percent bias = -26.1%), whereas sampling 4 (S4) propensity scores overestimated *ATE* (percent bias = 30.6%).

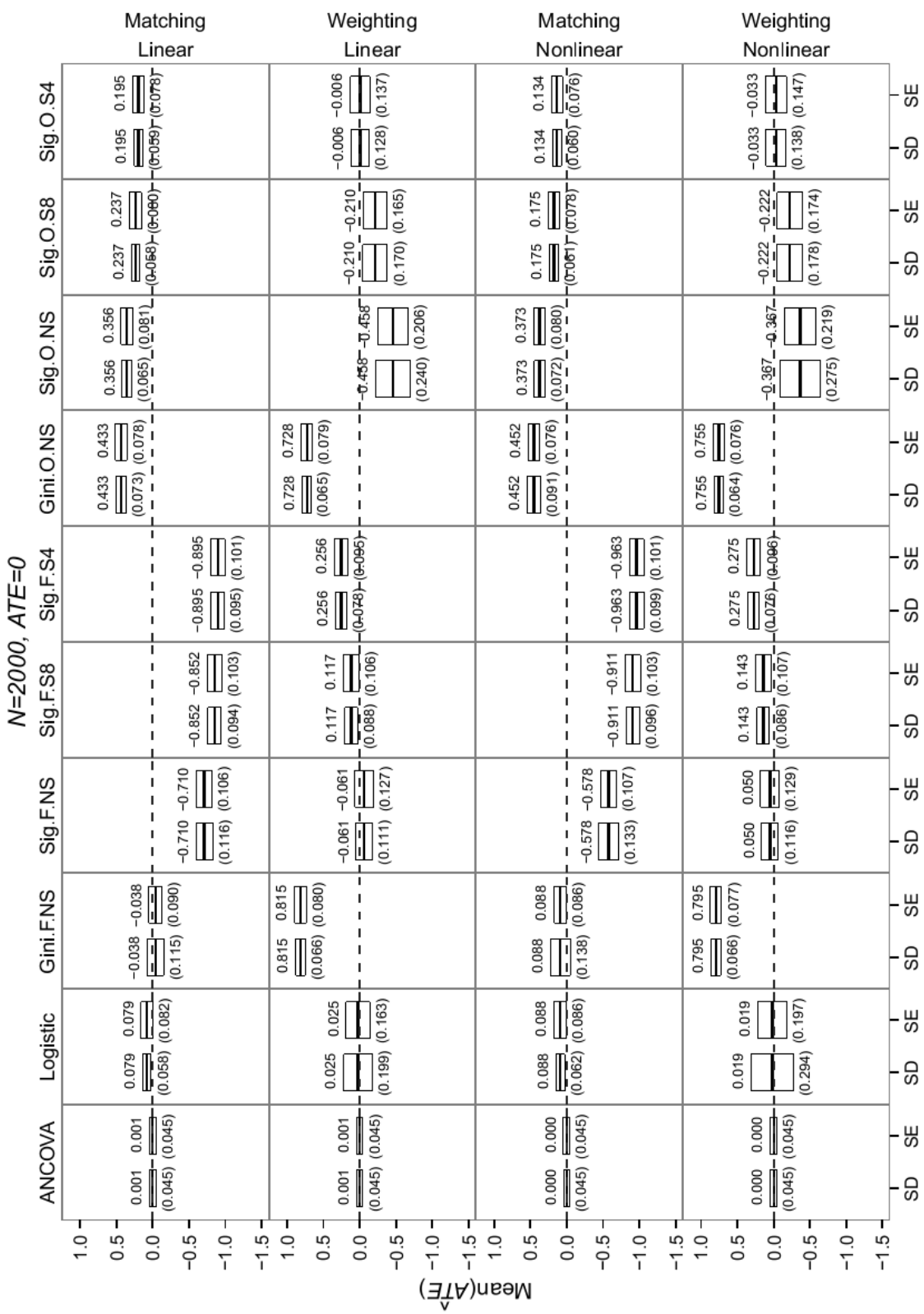
For $N = 2000$, the uncorrected *ATE* estimate again seriously overestimated the population value (percent bias = 206.1%). As expected, ANCOVA using the true population treatment-outcome model produced an unbiased *ATE* estimate (percent bias = -0.20%). The logistic regression propensity scores using the true population propensity score model produced an unbiased *ATE* estimate when matching (percent bias = 11.0%) or weighting (percent bias = 1.7%) adjustments were used. For the Random Forests propensity scores, the *ATE* estimate was influenced by the model specifications of (i) covariate and split value selection, (ii) random sampling of covariates for selection, (iii) methods for estimating Random Forests propensity scores, and (iv) methods of equating groups on propensity scores. First, under the full sample specification and matching adjustment, the Gini index (Gini.F.NS) propensity scores produced an unbiased *ATE* estimate (percent bias = -5.4%). Among the conditional significance test (Sig) propensity scores in this situation, no sampling covariates (NS) produced the least biased but unsatisfactory *ATE* estimate (percent bias = -97.7%), followed by sampling of 8 (S8) and 4 (S4) covariates in ascending order. Second, under the full sample specification and weighting adjustment, the Gini index (Gini.F.NS) propensity scores overestimated *ATE*. Among the conditional significance test (Sig) propensity scores in this situation, no sampling of covariates (NS) propensity scores produced an unbiased *ATE* estimate (percent bias = -8.9%). Sampling 8 (S8) and 4 (S4) propensity scores overestimated *ATE* in ascending order of bias. Third, under the out-of-bag sample specification and matching adjustment, all Random Forests propensity scores overestimated the *ATE*. In this situation, the Gini index (Gini.O.NS) and conditional significance test (Sig.O.NS) propensity scores both overestimated *ATE* with similar magnitudes. Among the

conditional significance test (Sig) propensity scores in this situation, sampling 4 covariates (S4) produced the least biased but unsatisfactory *ATE* estimate (percent bias = 26.8%), followed by sampling 8 (S8) and no sampling (NS) in ascending order. Fourth, under the out-of-bag sample specification and weighting adjustment, the Gini index (Gini.O.NS) propensity scores overestimated *ATE* (percent bias = 99.6%), whereas the conditional significance test (Sig.O.NS) propensity scores underestimated *ATE* (percent bias = -63.1%). Among the conditional significance test (Sig) propensity scores in this situation, sampling 4 covariates (S4) produced an unbiased *ATE* estimate (percent bias = -2.1%). Sampling 8 covariates (S8) and no sampling (NS) underestimated *ATE* in the ascending order of bias in magnitude.

To conclude the results, the two benchmarks methods ANCOVA and logistic regression propensity scores benchmark methods using the true population model produced unbiased *ATE* estimates across all sample sizes and propensity score models conditions. In $N = 600$, Sig.O.S8 using weighting produced an unbiased *ATE* estimate. In $N = 2000$, Gini.F.NS using matching, Sig.F.NS using weighting, and Sig.O.S4 using weighting produced unbiased *ATE* estimates.


N=600, ATE=0
Matching
Linear
Weighting
Linear
Matching
Nonlinear
Weighting
Nonlinear
ANCOVA
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
SD
SE
Mean(ATE)
1.0
0.5
0.0
−0.5
−1.0
−1.5


N=600, ATE=0.73
Matching
Linear
Weighting
Linear
Matching
Nonlinear
Weighting
Nonlinear
ANCOVA
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
SD
SE
Mean(ATE)


N=2000, ATE=0
Matching
Linear
Weighting
Linear
Matching
Nonlinear
Weighting
Nonlinear
ANCOVA
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
SD
SE
Mean(ATE)
1.0
0.5
0.0
−0.5
−1.0
−1.5

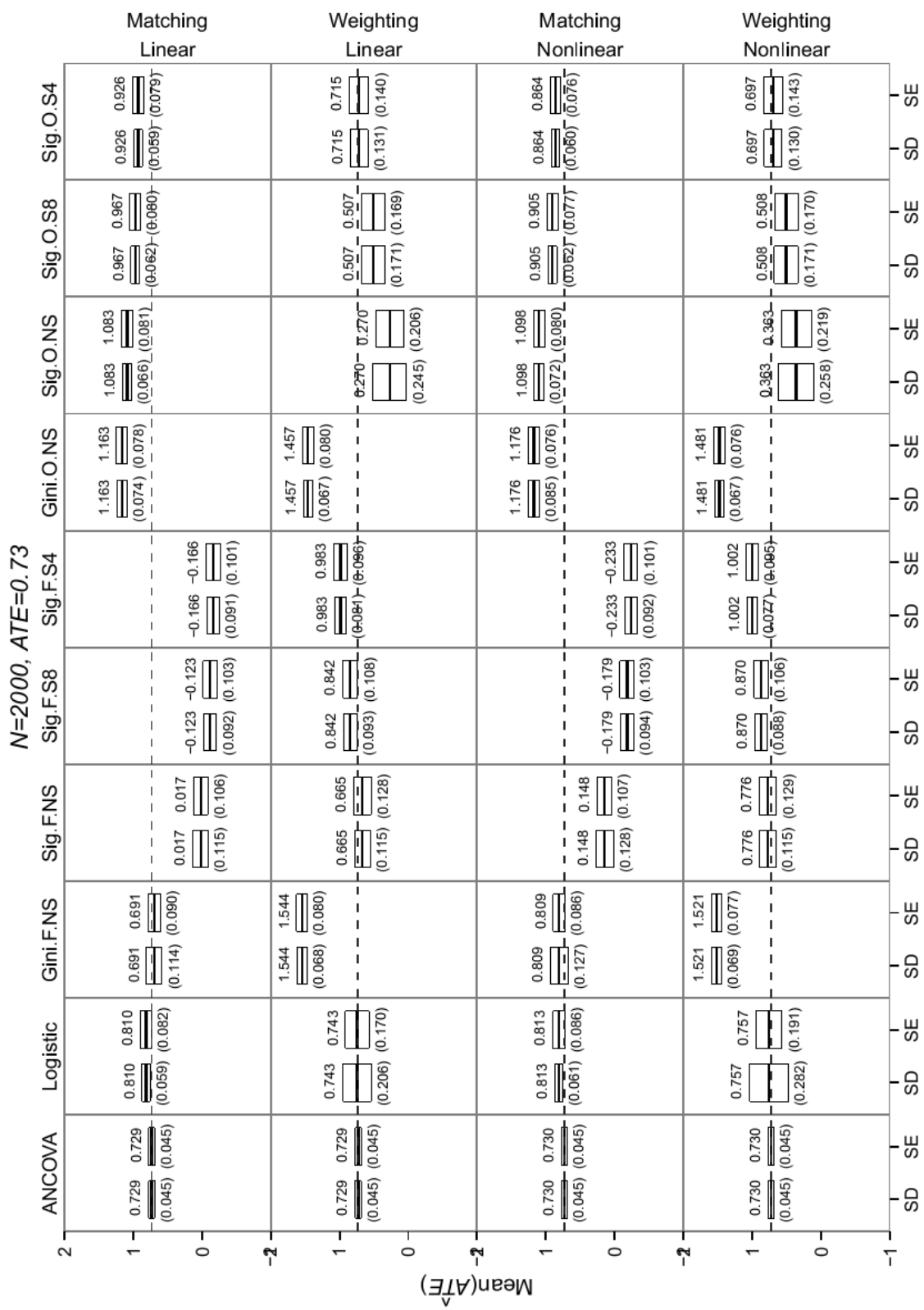

N=2000, ATE=0.73
Matching
Linear
Weighting
Linear
Matching
Nonlinear
Weighting
Nonlinear
ANCOVA
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
SD
SE
Mean(ATE)

Figure 2. Bar plots of average treatment effect estimate (*ATE*). In each panel there are two bar plots of the *ATE* estimate. The bold horizontal line in the middle of each bar plot is the mean *ATE* estimate across replications. The dashed horizontal line is the population *ATE* value (0 or 0.73). The bar plot on the *left* reflects ± 1 standard error of the *ATE* estimate (i.e., standard deviation of the *ATE* estimate across replications; labeled SD in the bottom axis). The bar plot on the *right* reflects ± 1 mean estimated standard error of the *ATE* estimate across replications (labeled SE in the bottom axis). The numerical value *above* each bar plot is the mean *ATE* estimate across replications. The numerical value in parenthesis *under* each bar plot is the corresponding standard error value.

Standard error of average treatment effect estimate. Figure 2 also presents the results concerning the standard error of the *ATE* estimate. As mentioned previously, in each panel of Figure 2 there are two bar plots of the *ATE* estimate. The bar plot on the *left* reflects ± 1 standard error of the *ATE* estimate (i.e., standard deviation of the *ATE* estimate across replications; labeled SD in the bottom axis). The bar plot on the *right* reflects ± 1 mean estimated standard error of the *ATE* estimate across replications (labeled SE in the bottom axis). The numerical value in parenthesis *under* each bar plot is the corresponding standard error value. Given the importance of producing an unbiased *ATE* estimate in propensity score analysis, the standard error results are described below only for the conditions where the *ATE* estimate was unbiased. Recall the Random Forests specifications and groups equating methods that produced unbiased *ATE* estimate: N = 600, Sig.O.S8 (weighting); N = 2000, Gini.F.NS (matching), Sig.F.NS (weighting), and Sig.O.S4 (weighting).

Sampling variability of ATE estimate. In each panel in Figure 2, the bar plot on the left (SD) shows the standard error of the *ATE* estimate. Given that an unbiased *ATE* estimate is achieved, the conditions that produce a smaller standard error of the *ATE* estimate are preferred. To supplement the graphical presentation, the mean square error of the *ATE* estimate was computed. The mean square error of the *ATE* estimate is the mean squared difference between the *ATE* estimate and its population value across replications.

$$\text{Mean Square Error of } ATE \text{ Estimate} = \sum_{i=1}^{Q}\left(\hat{ATE} - ATE\right)^2 / Q \qquad (6)$$

$\hat{ATE}$ is the average treatment effect estimate. $\Sigma(\cdot)$ is the summation function across Q replications. The mean square error also equals to (bias)^2 + (standard error)^2. Given an unbiased *ATE* estimate, the smaller its mean square error, the smaller its standard error. I

consider here only those conditions in which an unbiased *ATE* estimate was achieved. The complete results of the mean square error of the *ATE* estimate are presented in Appendix A.

For $N = 600$, ANCOVA had the smallest standard error (mean square error = 0.008), followed by logistic regression using matching (mean square error = 0.019), Sig.O.S8 using weighting (mean square error = 0.053) and logistic regression using weighting (mean square error = 0.170). For $N = 2000$, ANOVA had the smallest standard error (mean square error = 0.002), logistic regression using matching (mean square error = 0.010), followed by Gini.F.NS using matching (mean square error = 0.014), Sig.F.NS using weighting (mean square error = 0.017), Sig.O.S4 using weighting (mean square error = 0.017), and logistic regression using weighting (mean square error = 0.043)[7].

Bias of the estimated standard error of ATE estimate. If the bar plot on the left (SD) has the same vertical width as that on the right (SE), the estimated standard error of the *ATE* estimate is unbiased. If the vertical width of the bar plot on the left (SD) is smaller than that on the right (SE), the standard error is overestimated. If the vertical width of the bar plot on the left (SD) is greater than that on the right (SE), the standard error is underestimated. To supplement the graphical presentation, the percent bias of the estimated standard error of *ATE* estimate was computed. The percent bias of the estimated standard error is the mean difference between the estimated standard error of the *ATE* estimate and its population value across replications, with respect to the population value.

[7] In the nonlinear model, the ascending order of the mean square error was logistic regression using matching (mean square error = 0.011), Sig.F.NS using weighting (= 0.015), Sig.O.S4 using weighting (= 0.018), Gini.F.NS using matching (= 0.022), and logistic regression using weighting (mean square error = 0.080).

$$\text{Percent Bias of Estimated Standard Error of } \hat{ATE} = \frac{\sum_{i=1}^{Q}(\widehat{SE}(\hat{ATE}) - SD(\hat{ATE}))/Q}{SD(\hat{ATE})} \quad (7)$$

$\hat{ATE}$ is the average treatment effect estimate. $\widehat{SE}(\hat{ATE})$ is the estimated *SE* of *ATE* estimate. $SD(\hat{ATE})$ is the standard error of *ATE* estimate. $\Sigma(\cdot)$ is the summation function across Q replications. Percent bias of 0.0 indicates that the estimated standard error is unbiased. I again loosely used Flora and Curran's (2004) suggestion of < 10% as a criterion for satisfactory percent bias. The complete results of the percent bias of the estimated standard error of the *ATE* estimate are presented in Appendix A.

For *N* = 600, ANCOVA produced unbiased standard error estimate (percent bias = -1.9%). Logistic regression propensity scores using matching overestimated standard error (percent bias = 38.9%). Logistic regression using weighting underestimated standard error (percent bias = -36.0%). The Sig.O.S8 Random Forests propensity scores produced an unbiased standard error estimate when weighting was used (percent bias = -3.6%).

For *N* = 2000, ANCOVA produced unbiased standard error estimate (percent bias = -0.7%). Logistic regression propensity scores using matching overestimated the standard error (percent bias = 38.5%). Logistic regression using weighting underestimated the standard error (percent bias = -17.3%). For the Random Forests propensity scores, Gini.F.NS using matching underestimated the standard error (percent bias = -20.7%). Sig.F.NS using weighting produced an unbiased standard error estimate (percent bias = 11.0%). Sig.O.S4 using weighting also produced an unbiased standard error estimate, with a smaller bias compared to Sig.F.NS using weighting (percent bias = 6.8%).

Statistical inference of average treatment effect. Confidence intervals and hypothesis testing are used to provide statistical inferences about the *ATE*. Both methods require the analyst to specify an α level, typically 0.05.

The (1 - α)% confidence interval of *ATE* is $[\widehat{ATE} \pm t_{(1-\alpha/2,df)}\ \widehat{SE}(\widehat{ATE})]$. The degrees of freedom *df* of the critical $t_{(1-\alpha/2,df)}$ depends on the method of equating treatment groups (matching versus weighting). The coverage rate is the empirical proportion of replications in which the confidence interval includes the population *ATE* value. The coverage rate reflects both the bias of the *ATE* estimate and the bias of the estimated standard error of the *ATE* estimate. Given an unbiased *ATE* estimate, a coverage rate that equals (1 - α) indicates the confidence interval correctly represents the standard error of the *ATE*. In my analyses, α was set to the conventional level of 0.05. I loosely used the binomial confidence interval (Savalei, 2010) – a coverage rate that falls between $[(1 - \alpha) \pm z_{(1-\alpha/2)}\sqrt{\alpha(1-\alpha)/Q}]$ was deemed satisfactory (here, [0.936, 0.964]). Q is the number of replications.

In hypothesis testing, the null hypothesis *ATE* = 0 is tested by the test statistic $t^* = \widehat{ATE}/\widehat{SE}(\widehat{ATE})$. When the null hypothesis is true, t^* follows a *t* distribution with degrees of freedom *df*. The degrees of freedom *df* depends on the method of equating treatment groups (matching, weighting). For each condition, the empirical proportion of replications that reject the null hypothesis at the α level was computed. When the population *ATE* = 0, this proportion is the empirical Type I error rate of *ATE*. Given an unbiased *ATE* estimate, if the empirical Type I error rate equals the specified α, the hypothesis test correctly represents the Type I error rate. If the empirical Type I error rate is less than α, the Type I error rate is conservative. If the empirical Type I error rate is

greater than α, the Type-I error rate is inflated and the researcher rejects the null hypothesis more often than the specified α level. I again loosely used the binomial confidence interval (Savalei, 2010) – an empirical Type I error rate that falls between [$\alpha \pm z_{(1-\alpha/2)}\sqrt{\alpha(1-\alpha)/Q}$] was deemed satisfactory (here, [0.036, 0.064]). Q is the number of replications. When the population *ATE* = 0.73, the empirical proportion above is the empirical statistical power. Given a conservative or correct Type I error rate, the conditions that have higher empirical statistical power are preferred.

Given the importance of obtaining an unbiased *ATE* estimate in propensity score analysis, the statistical inference results are described below only for the conditions where the *ATE* estimate was unbiased. Besides the ANCOVA and logistic regression propensity scores benchmark methods, recall the Random Forests specifications and groups equating methods: N = 600, Sig.O.S8 (weighting); N = 2000, Gini.F.NS (matching), Sig.F.NS (weighting), and Sig.O.S4 (weighting). Table 3 presents the coverage rate when *ATE* = 0.73, empirical Type I error rate, and empirical statistical power for these conditions. The complete results of the coverage rate, empirical Type I error rate, and empirical statistical power results are presented in Appendix A.

For N = 600, ANCOVA produced confidence intervals that correctly represented the standard error (coverage rate = 0.942), and a correct Type I error rate (= 0.041). Logistic regression propensity scores using matching produced confidence intervals that over-represented the standard error (coverage rate = 0.983) and yielded a corresponding conservative Type I error rate (= 0.025). Logistic regression using weighting produced confidence intervals that under-represented the standard error (coverage rate = 0.803), and yielded a correspondingly inflated Type I error rate (= 0.187). The Sig.O.S8 Random

Forests propensity scores using weighting produced confidence intervals that slightly under-represented the standard error (coverage rate = 0.922), and correspondingly a slightly inflated Type I error rate (= 0.087). Logistic regression using matching had higher statistical power (= 1.0) than Sig.O.S8 using weighting (= 0.869).

For N = 2000, ANCOVA produced confidence intervals that correctly represented the standard error (coverage rate = 0.949), and a correct Type I error rate (= 0.053). Logistic regression using matching produced confidence intervals that under-represented the standard error (coverage rate = 0.905) and yielded a corresponding slightly inflated Type I error rate (= 0.073). Logistic regression using weighting also produced confidence intervals that under-represented the standard error (coverage rate = 0.900) and an inflated Type I error rate (= 0.115). For the Random Forests propensity scores, Gini.F.NS using matching produced confidence intervals that under-represented the standard error (coverage rate = 0.863) and yielded a correspondingly inflated Type I error rate (= 0.142). Sig.F.NS and Sig.O.S4 using weighting produced confidence intervals that correctly represented the standard error (coverage rate = 0.964 and 0.950, respectively), and a correct Type I error rates (= 0.035 and 0.040, respectively). Sig.F.NS had trivially higher statistical power (= 0.996) than the Sig.O.S4 (power = 0.962).

Table 3

Coverage Rate (*ATE* = 0.73), Empirical Type I Error Rate and Empirical Statistical Power

N = 600 (Linear Model)	Coverage Rate (*ATE* = 0.73)	Empirical Type I Error Rate	Empirical Statistical Power
ANCOVA	0.942	0.041	1.000
Logistic (Matching)	0.983	0.025	1.000
Logistic (Weighting)	0.803	0.187	0.783
Sig.O.S8 (Weighting)	0.922	0.087	0.869
N = 2000 (Linear Model)	Coverage Rate (*ATE* = 0.73)	Empirical Type I Error Rate	Empirical Statistical Power
ANCOVA	0.949	0.053	1.000
Logistic (Matching)	0.905	0.073	1.000
Logistic (Weighting)	0.900	0.115	0.902
Gini.F.NS (Matching)	0.863	0.142	1.000
Sig.F.NS (Weighting)	0.964	0.035	0.996
Sig.O.S4 (Weighting)	0.950	0.040	0.962

Balance on Covariates' Distributions between Treatment Conditions

This section presents the results that answer the primary research questions concerning the relationship between the balance on covariates' distributions and the bias of the average treatment effect estimate. In the simulation study design, the covariates and the propensity scores models were designed to primarily introduce mean differences on the covariates between the treatment conditions (Lee et al., 2010). In propensity score analysis, the standardized mean difference (*SMD*) is a commonly used measure that summarizes mean differences of covariates of different types and scales.

$$\text{Standardized Mean Difference } (SMD) = (\bar{X}_t - \bar{X}_c)/s_t^{un} \qquad (5)$$

$\bar{X}_t$ is the treatment group sample mean of covariate *X*, $\bar{X}_c$ is the control group' sample mean of covariate *X*, and s_t^{un} is the uncorrected treatment group sample standard deviation of covariate *X* (Ho et al., 2011; Stuart, 2010). A standardized mean difference (*SMD*) of 0.0 indicates perfect balance on the covariate's means between treatment conditions.

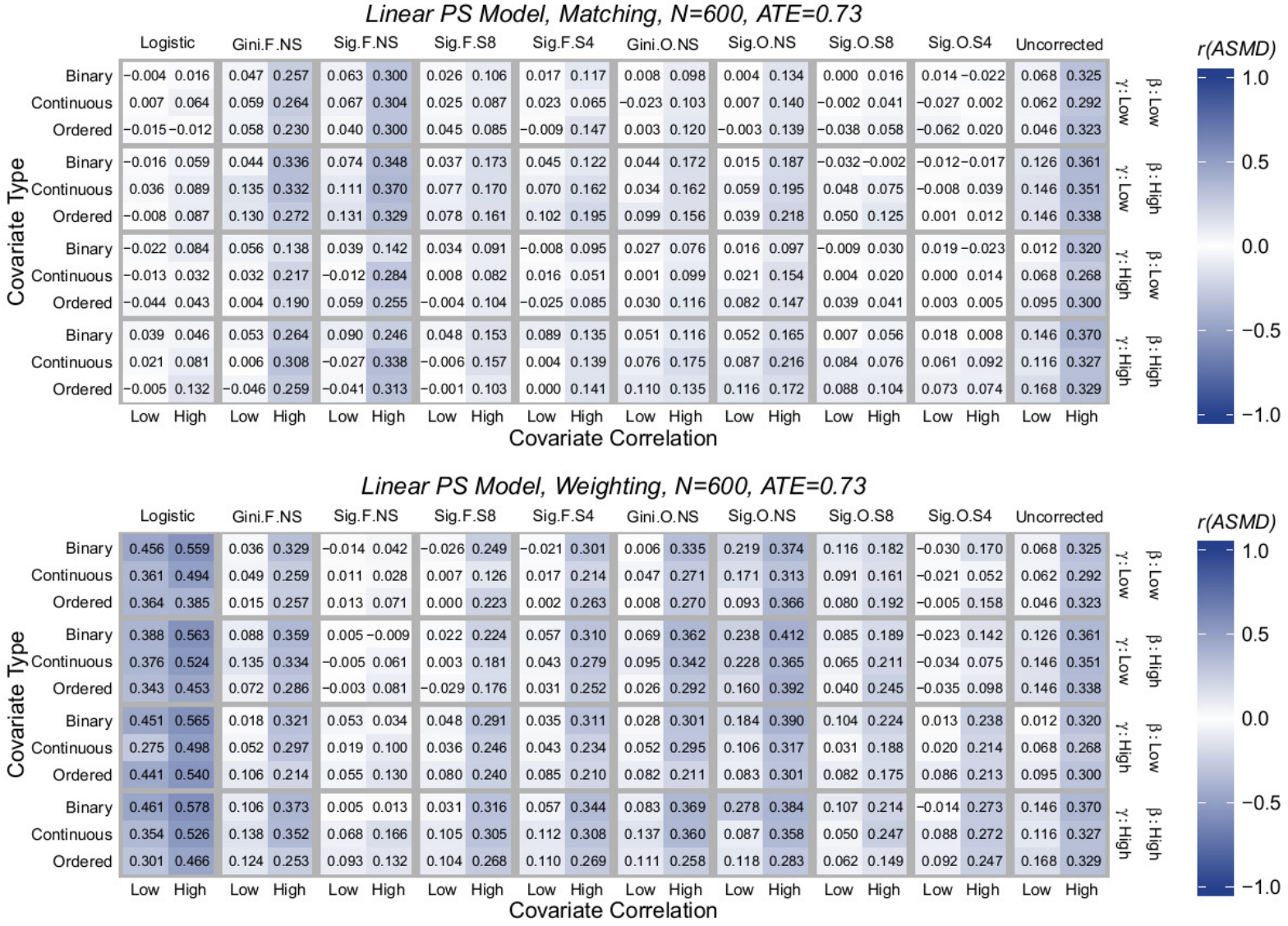
Correlation between standardized mean difference and magnitude of bias of *ATE* estimate. Figure 3 consists of tile plots of the correlation between the covariates' absolute standardized mean difference (*ASMD*, i.e., *SMD* ignoring direction) and the magnitude of the bias of *ATE* estimate across replications. The plots are partitioned by sample size *N* (600, 2000), population average treatment effect *ATE* (0, 0.73), propensity score model (linear, nonlinear), and methods of equating groups (matching, weighting). In each plot, the tiles are arranged into rows of covariate types (binary, ordered-categorical, continuous; see the leftmost axis labels), propensity score model regression coefficient levels (γ: low, high; see the rightmost axis labels), treatment-outcome model

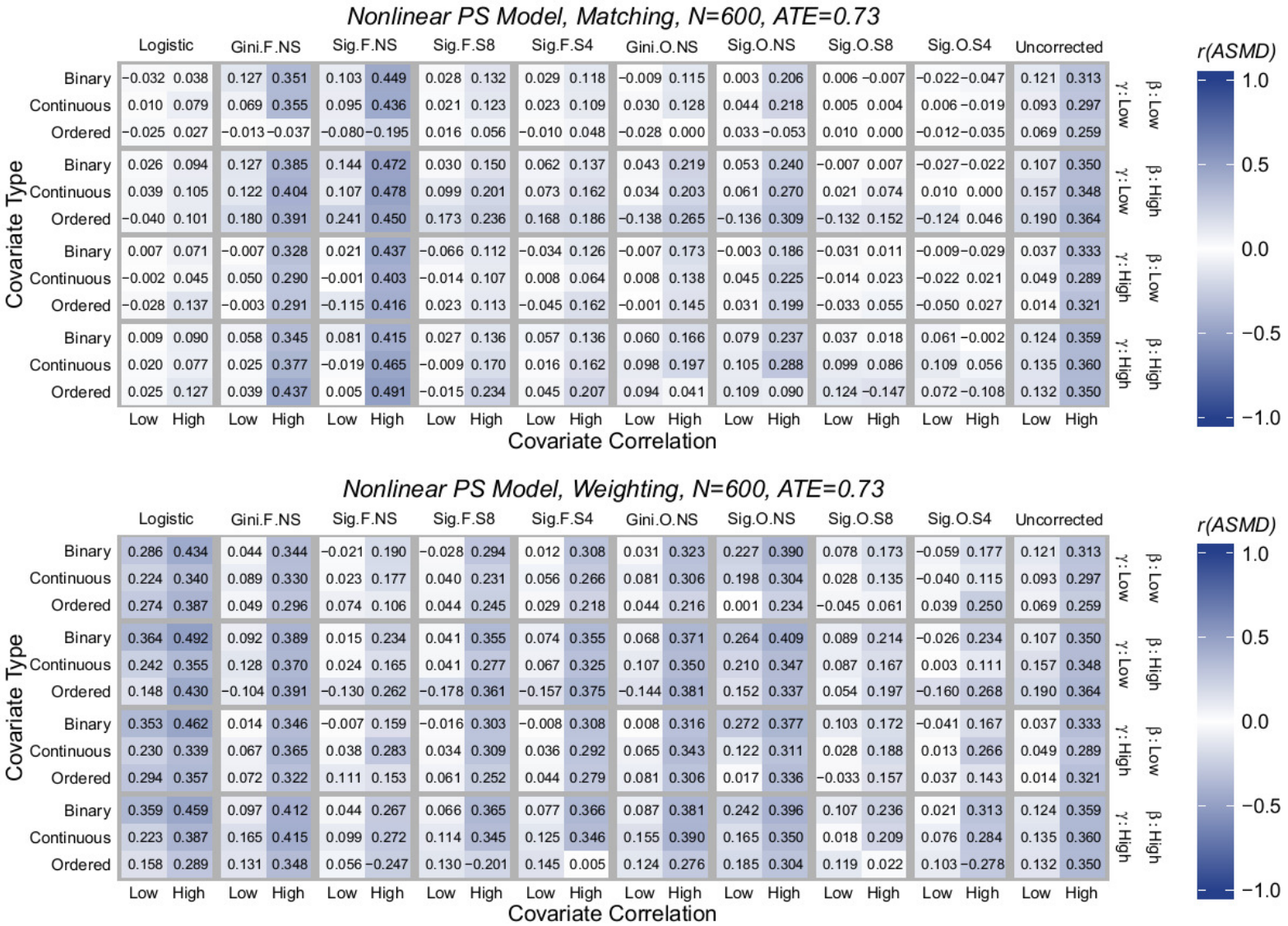
regression coefficient levels (β: low, high; see the rightmost axis labels). The tiles are also arranged into 10 columns representing the two benchmark methods for assessing covariate balance (logistic regression propensity scores; uncorrected) and the Random Forests model specifications for estimating propensity scores (see the top axis labels). The ANCOVA benchmark method does not allow direct examination of the balance of covariates distributions. The numerical value in each tile is the correlation *r(ASMD)* of the condition. When *r(ASMD)* = 0.0, the tile is white in color. As *r(ASMD)* approaches 1.0 or -1.0, the tile becomes increasingly dark. Based on Lee et al.'s (2010) simulation results that the correlations between *AMSD* and magnitude of bias of *ATE* estimate ranging from 0.38 to 0.66, I loosely used *r(ASMD)* > 0.3 as the criterion for a satisfactory relationship.

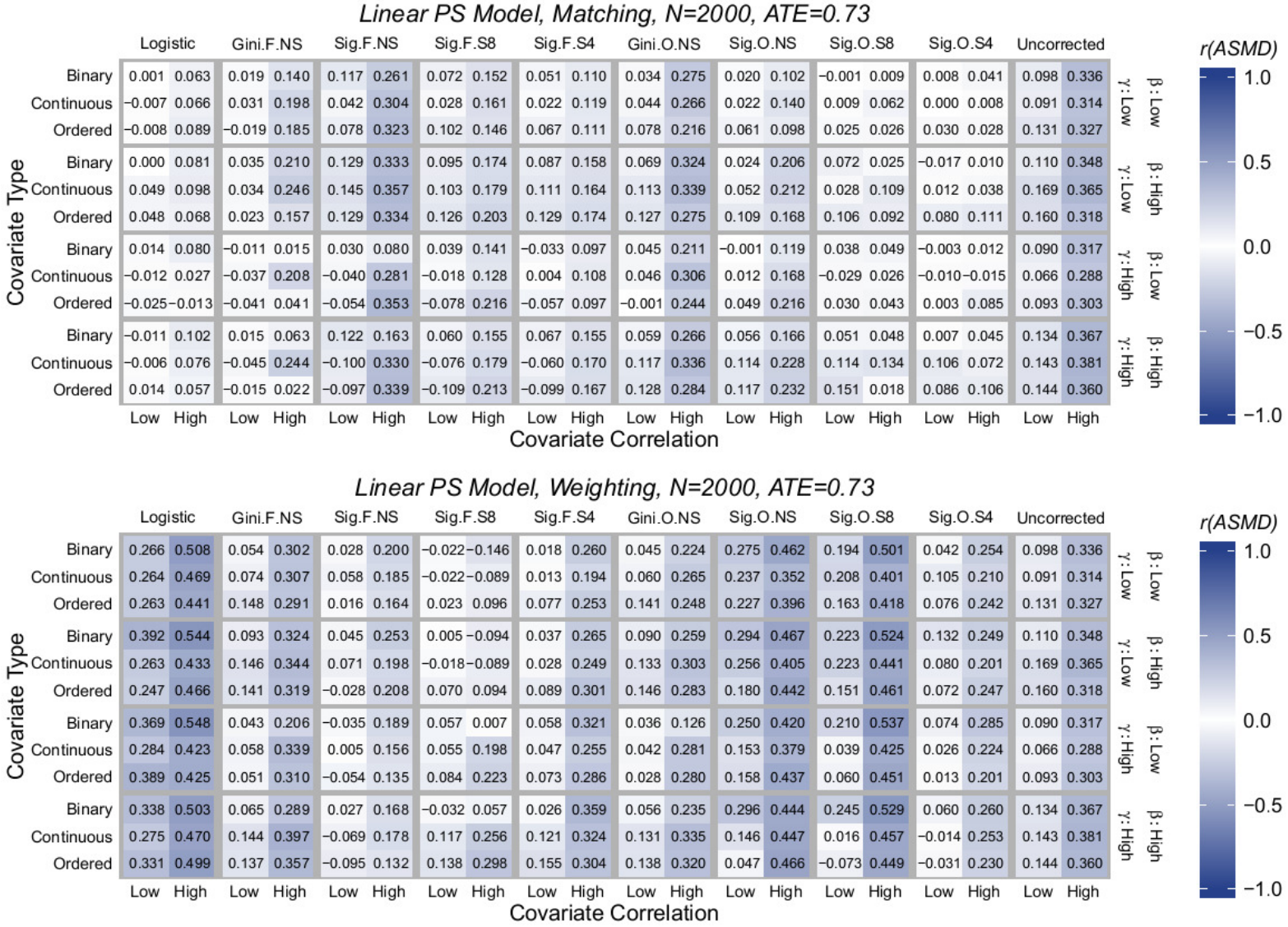
The pattern of the results in Figure 3 was similar between the *ATE* = 0 and 0.73 conditions. I present the results for *ATE* = 0.73, linear propensity score model, and *N* = 600 and 2000. Appendix B presents the figures for *ATE* = 0. Across all sample sizes *N* and propensity score models (linear, nonlinear), the covariates in the high level of covariate correlation condition had large uncorrected mean differences between treatment conditions and those in the low level covariate correlation condition had small uncorrected mean differences between treatment conditions (see Figure 4 below). For *N* = 600, when matching was used, only the Gini.F.NS and Sig.F.NS propensity scores had *r(ASMD)* > 0.3 among most covariates in the high level covariate correlation condition. When weighting was used, the logistic regression propensity scores had *r(ASMD)* > 0.3 in most covariate conditions. The covariates in the high level covariate correlation condition had higher *r(ASMD)* than the covariates in the low level covariate correlation

condition. Gini.F.NS, Sig.F.S8, Sig.F.S4, Gini.O.NS, and Sig.O.NS had *r(ASMD)* > 0.3 among most covariates in the high level covariate correlation condition.

For $N = 2000$, when matching was used, only the Sig.F.NS and Gini.O.NS propensity scores had *r(ASMD)* > 0.3 among most covariates in the high level covariate correlation condition. When weighting was used, the logistic regression propensity scores had *r(ASMD)* > 0.3 in most covariate conditions. The covariates in the high level covariate correlation condition had higher *r(ASMD)* than the covariates in the low level covariate correlation condition. Gini.F.NS, Sig.F.S4, Gini.O.NS, Sig.O.NS, and Sig.O.S8 had *r(ASMD)* > 0.3 among most covariates in the high level covariate correlation condition.


Linear PS Model, Matching, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Linear PS Model, Weighting, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0


Nonlinear PS Model, Matching, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
γ : High
β : Low
β : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Nonlinear PS Model, Weighting, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
γ : High
β : Low
β : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0


Linear PS Model, Matching, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Linear PS Model, Weighting, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

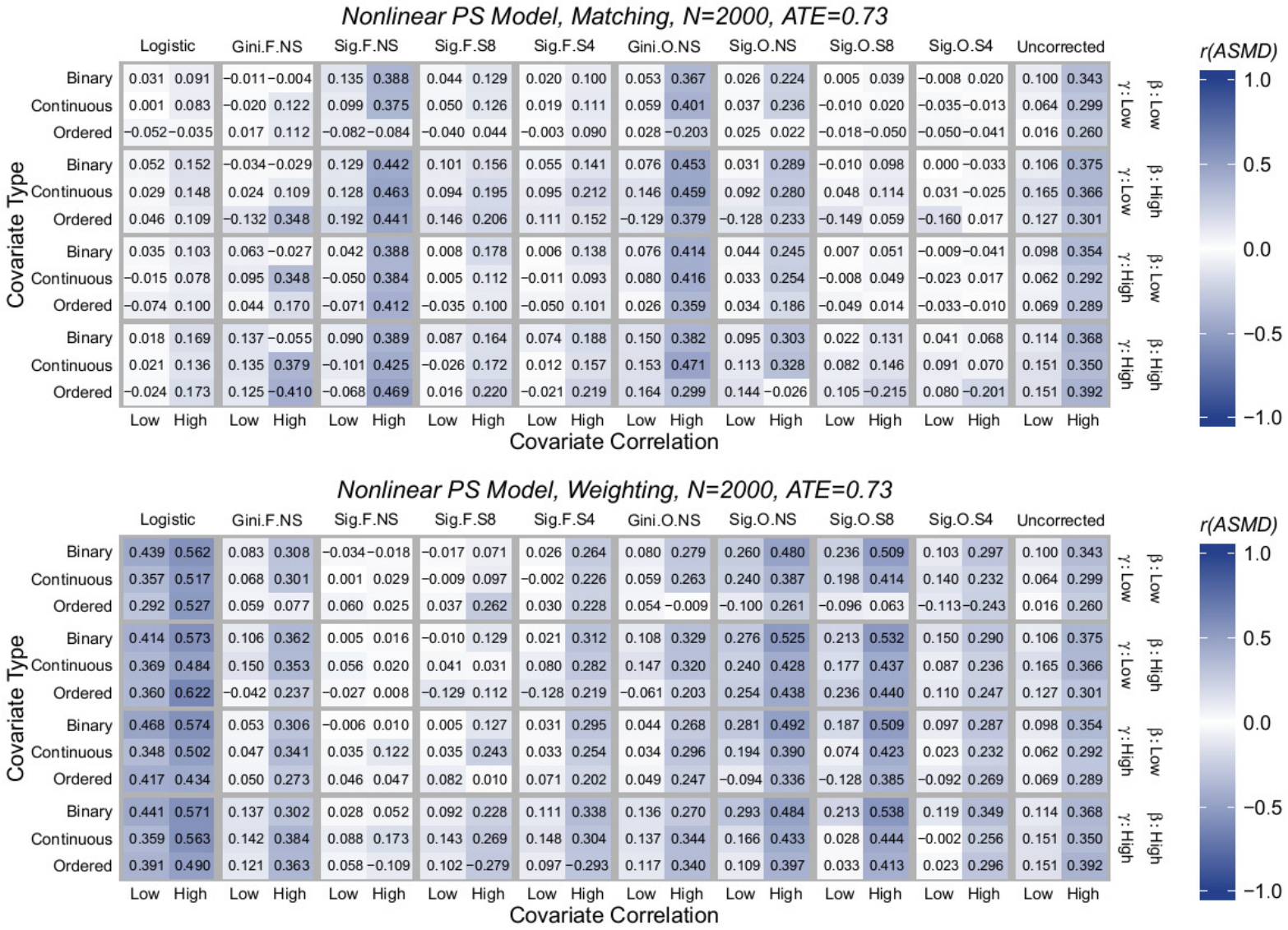

Nonlinear PS Model, Matching, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low
β : High
γ : Low
γ : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Nonlinear PS Model, Weighting, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low
β : High
γ : Low
γ : High
Low
High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

Figure 3. Correlation between absolute standardized mean difference (*ASMD*) and magnitude of bias of *ATE* estimate across replications. The numerical value in each tile is the correlation *r(ASMD)* of the condition. When *r(ASMD)* = 0.0, the tile is white in color. As *r(ASMD)* approaches 1.0 or -1.0, the tile becomes increasingly dark. .

Standardized mean difference and bias of *ATE* estimate between propensity scores. Figure 4 consists of tile plots of the mean of the absolute standardized mean difference (*ASMD – SMD* ignoring direction) of the covariates across replications. The arrangement of the tile plots is the same as those in Figure 3. The numerical value in each tile is the mean *ASMD* of the condition. As the *ASMD* of the condition becomes higher, the tile becomes increasingly dark. I loosely used Rubin's (2001) suggestion that *ASMD* < 0.25 is satisfactory. Although much information is provided by Figure 4, the tile plots were used primarily to investigate whether the mean *ASMD* is related to the bias of the *ATE* estimate between different methods of computing propensity scores and methods of equating groups (matching, weighting). Recall the conditions that produced an unbiased *ATE* estimate: $N = 600$, logistic regression (matching and weighting), and Sig.O.S8 (weighting); $N = 2000$, logistic regression (matching and weighting), Gini.F.NS (matching), Sig.F.NS (weighting), and Sig.O.S4 (weighting). The pattern of the results in Figure 4 was consistent for both the conditions in which the population *ATE* = 0 and 0.73. Below I focus on the results for *ATE* = 0.73, sample size $N = 600$ and 2000, and for the linear and nonlinear propensity score models. Appendix C presents the tile plots for *ATE* = 0.

For $N = 600$ and the linear propensity score model, the *ASMD* of the covariates was similar across covariate types, γ, and β values before adjustment. Before adjustment, the *ASMDs* in the high level covariate correlation condition were larger than those of the in the low level covariate correlation condition. All the conditions that yielded an unbiased *ATE* estimate produced a satisfactory *ASMD* in all covariate conditions, in the ascending order of logistic regression using matching (maximum *ASMD* = 0.063),

Sig.O.S8 using weighting (maximum *ASMD* = 0.163), and logistic regression using weighting (maximum *ASMD* = 0.182). Conditions under which the Random Forests propensity scores yielded biased *ATE* estimates included the following. Under the full sample specification and matching, the high level covariate correlation condition had a large *ASMD*. These propensity scores underestimated *ATE* with percent bias > 100% in magnitude. Under the out-of-bag sample specification and matching, all covariate conditions had satisfactory *ASMD* (maximum *ASMD* = 0.234). These propensity scores overestimated *ATE* with percent bias between 29.7% and 58.4%. Under weighting, Gini.F.NS and Gini.O.NS had large *ASMD*s in the high level covariate correlation condition. These two propensity scores overestimated *ATE* with percent bias > 80%. All of the conditional significance test propensity scores had a satisfactory *ASMD* with similar results to Sig.O.S8. These propensity scores had percent bias between 17.7% and 56.2% in magnitude.

The pattern of the results for $N = 600$ and the nonlinear propensity score model was similar to that for $N = 600$ and linear propensity score model, except for the aspects noted below. First, before adjustment, the *ASMD* of the ordered-categorical covariates in the low level covariate correlation condition was larger than those of the binary and continuous covariates in the low level covariate correlation condition. Across all Random Forests propensity scores and adjustment methods (including Sig.O.S8 using weighting that produced an unbiased *ATE* estimate), the *ASMD* of the ordered-categorical covariates were often large.

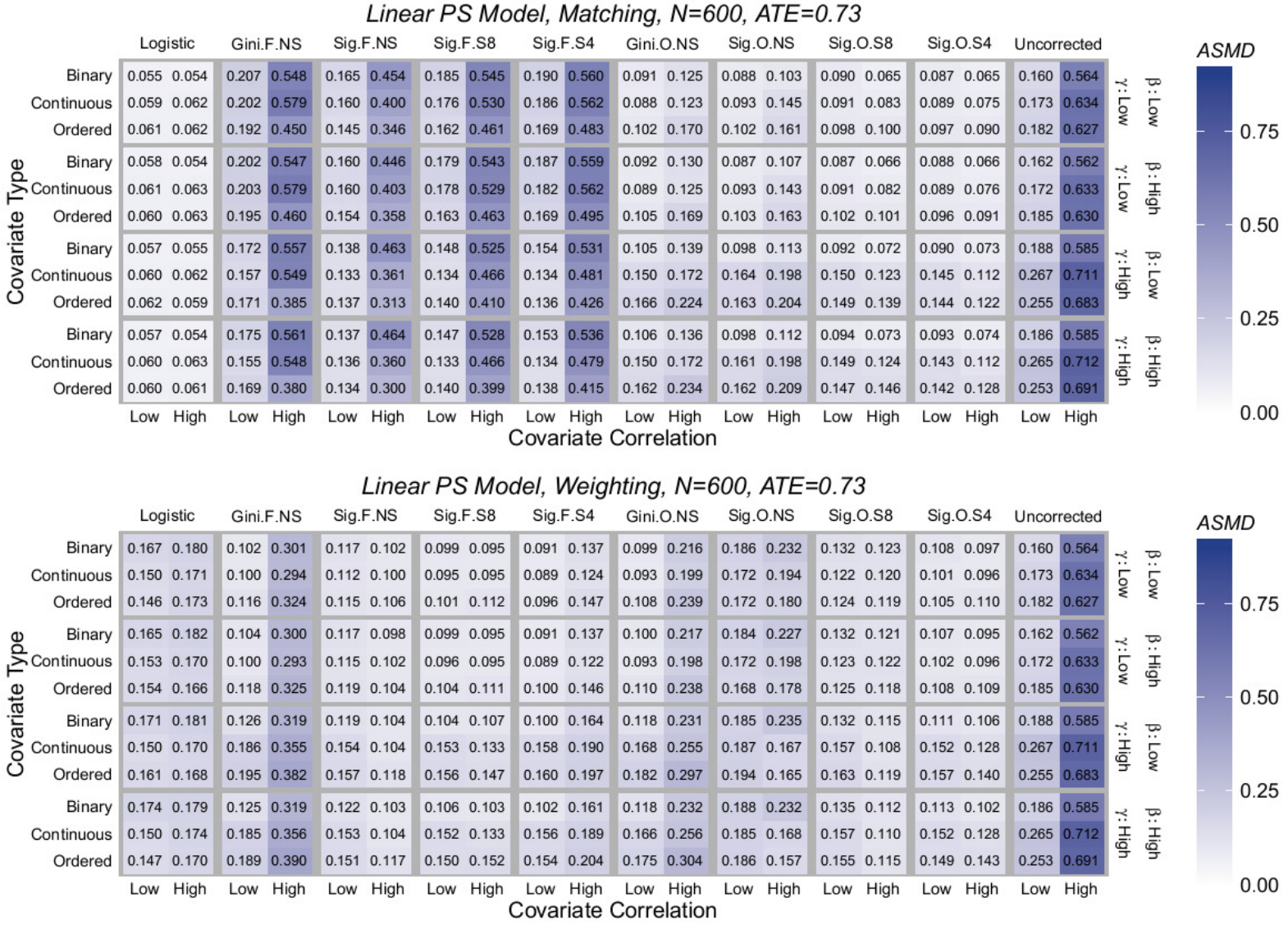
For $N = 2000$ and linear model, the *ASMD*s of the covariates were similar across covariate types, γ, and β values before adjustment. Before adjustment, the *ASMDs* of the

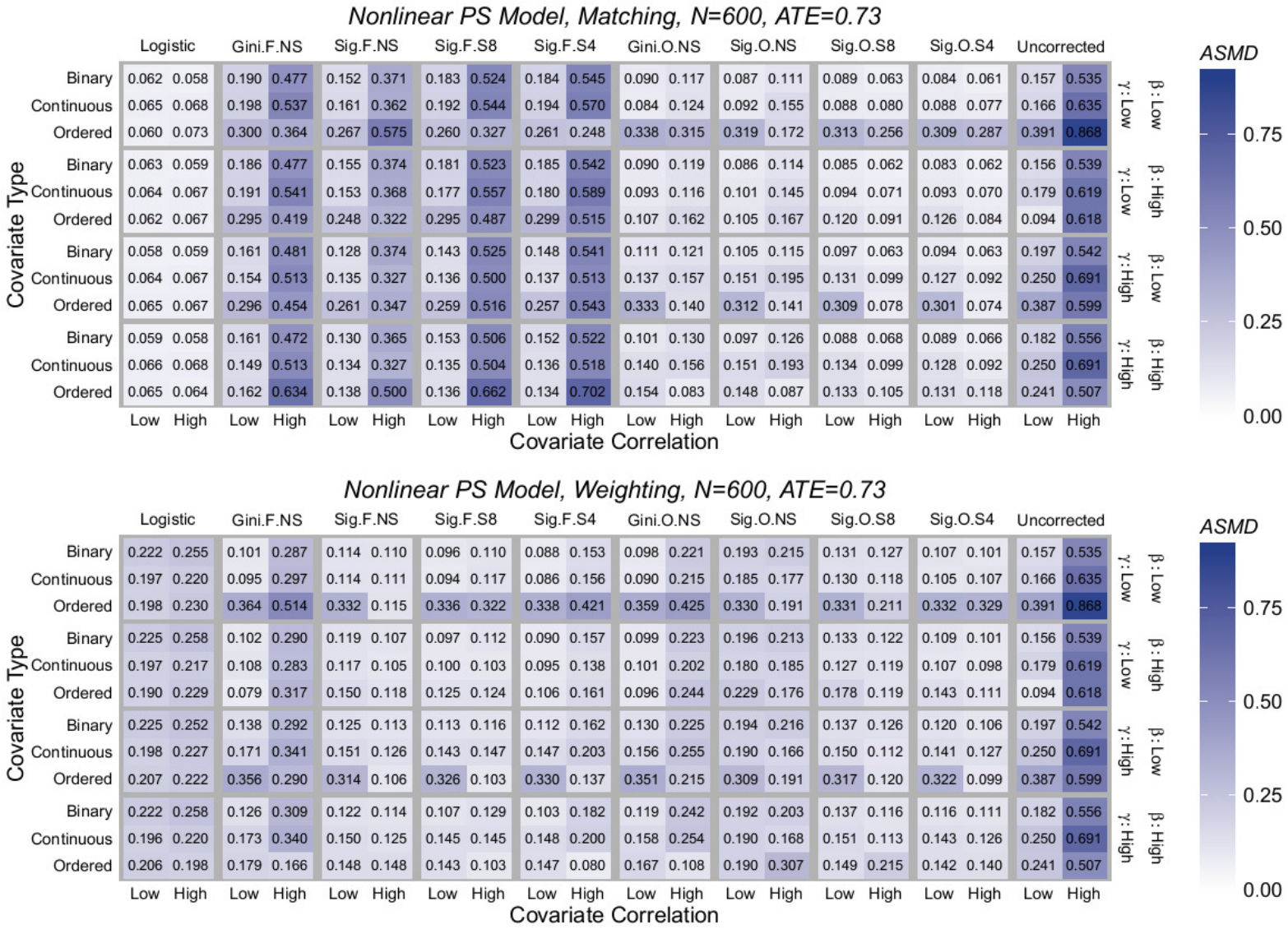
covariates in the high level covariate correlation condition were larger than those of the covariates in the low level covariate correlation condition. All of the propensity score conditions that yielded an unbiased *ATE* estimate produced a satisfactory *ASMD* in all covariate conditions, in the ascending order of logistic regression using matching (maximum *ASMD* = 0.042), logistic regression using weighting (maximum *ASMD* = 0.087), Sig.O.S4 using weighting (maximum *ASMD* = 0.121), Sig.F.NS using weighting (maximum *ASMD* = 0.136), and Gini.F.NS using matching (maximum *ASMD* = 0.175). Conditions under which the Random Forests propensity scores that yielded biased *ATE* estimates included the following. Under the full sample specification and matching, covariates in the high level covariate correlation condition had large *ASMDs*. These propensity scores underestimated *ATE* with percent bias > 97% in magnitude. Under the out-of-bag sample specification and matching, all covariate conditions had satisfactory *ASMD*s (maximum *ASMD* = 0.238). These propensity scores overestimated *ATE* with percent bias between 26.8% and 59.3%. Using weighting, Gini.F.NS, Gini.O.NS, Sig.O.NS had large *ASMD* for covariates in the high level covariate correlation condition. These propensity scores produced biased *ATE* estimates with percent bias > 63% in magnitude. Sig.F.S8, Sig.F.S4, and Sig.O.S8 had satisfactory *ASMD* with similar results to Sig.F.NS and Sig.O.S4. These propensity scores had percent bias in *ATE* estimates between 15.4% and 34.7% in magnitude.

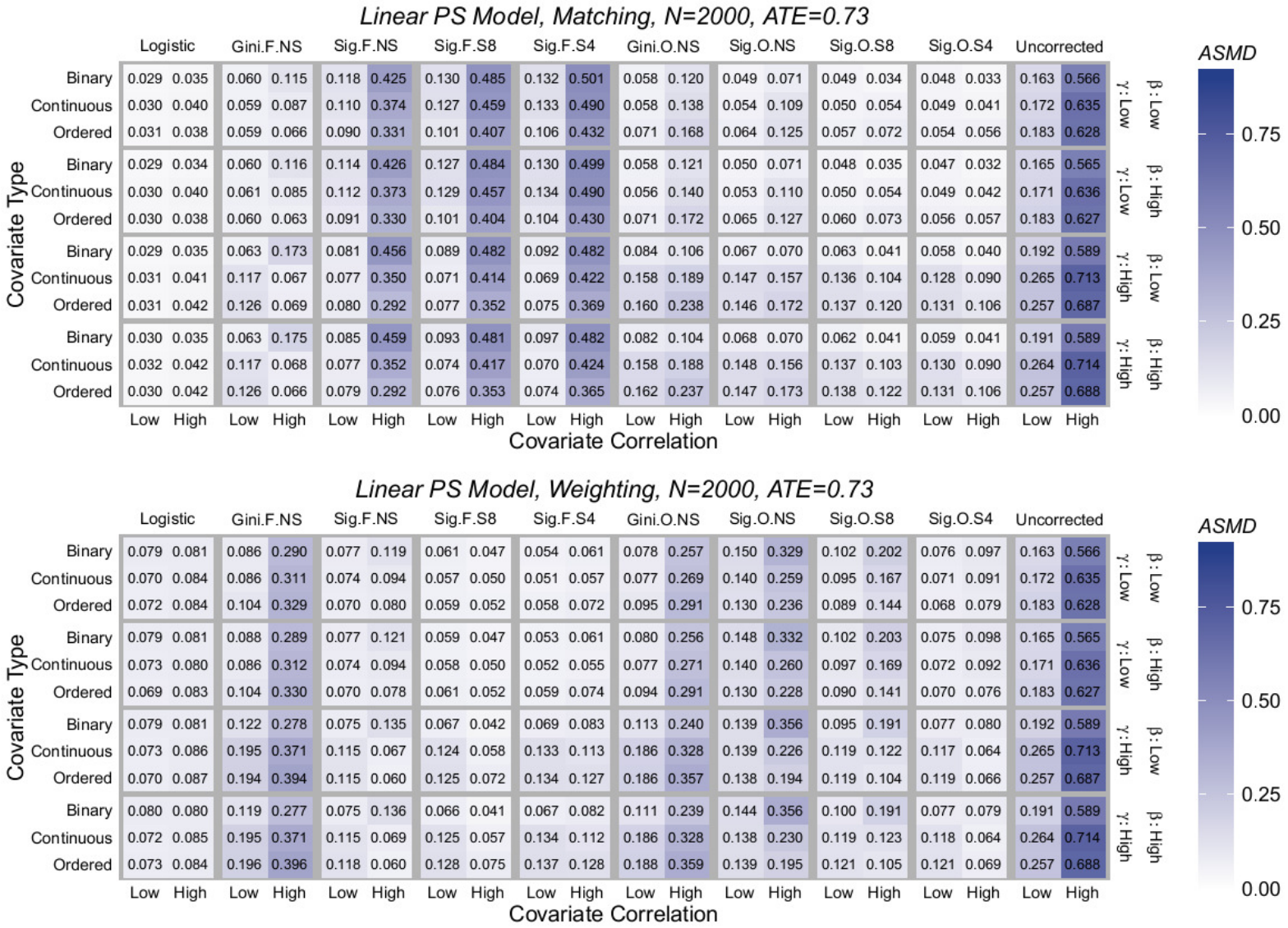
The pattern of the results for $N = 2000$ and the nonlinear model was similar to that for $N = 2000$ and the linear propensity score model except for the aspects noted below. First, before adjustment, the *ASMDs* of the ordered-categorical covariates were larger than those of the binary and continuous covariates. Across all Random Forests propensity

scores and adjustment methods (including Gini.F.NS using matching, Sig.F.NS using weighting, Sig.O.S8 using weighting that produced an unbiased *ATE* estimate), the *ASMD*s of the ordered-categorical covariates were often large.

In summary, the conditions that yielded an unbiased *ATE* estimate also produced a satisfactory *ASMD* for most but not all covariates. Second, larger *ASMD*s of some covariates were related to larger percent bias of the *ATE* estimates. Third, small *ASMD*s for most covariates of the Random Forests propensity scores were *not* necessarily related to an unbiased *ATE* estimate.


Linear PS Model, Matching, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Linear PS Model, Weighting, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00


Nonlinear PS Model, Matching, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Nonlinear PS Model, Weighting, N=600, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00


Linear PS Model, Matching, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary 0.029 0.035 0.060 0.115 0.118 0.425 0.130 0.485 0.132 0.501 0.058 0.120 0.049 0.071 0.049 0.034 0.048 0.033 0.163 0.566
Continuous 0.030 0.040 0.059 0.087 0.110 0.374 0.127 0.459 0.133 0.490 0.058 0.138 0.054 0.109 0.050 0.054 0.049 0.041 0.172 0.635
Ordered 0.031 0.038 0.059 0.066 0.090 0.331 0.101 0.407 0.106 0.432 0.071 0.168 0.064 0.125 0.057 0.072 0.054 0.056 0.183 0.628
γ : Low β : Low
Binary 0.029 0.034 0.060 0.116 0.114 0.426 0.127 0.484 0.130 0.499 0.058 0.121 0.050 0.071 0.048 0.035 0.047 0.032 0.165 0.565
Continuous 0.030 0.040 0.061 0.085 0.112 0.373 0.129 0.457 0.134 0.490 0.056 0.140 0.053 0.110 0.050 0.054 0.049 0.042 0.171 0.636
Ordered 0.030 0.038 0.060 0.063 0.091 0.330 0.101 0.404 0.104 0.430 0.071 0.172 0.065 0.127 0.060 0.073 0.056 0.057 0.183 0.627
γ : Low β : High
Binary 0.029 0.035 0.063 0.173 0.081 0.456 0.089 0.482 0.092 0.482 0.084 0.106 0.067 0.070 0.063 0.041 0.058 0.040 0.192 0.589
Continuous 0.031 0.041 0.117 0.067 0.077 0.350 0.071 0.414 0.069 0.422 0.158 0.189 0.147 0.157 0.136 0.104 0.128 0.090 0.265 0.713
Ordered 0.031 0.042 0.126 0.069 0.080 0.292 0.077 0.352 0.075 0.369 0.160 0.238 0.146 0.172 0.137 0.120 0.131 0.106 0.257 0.687
γ : High β : Low
Binary 0.030 0.035 0.063 0.175 0.085 0.459 0.093 0.481 0.097 0.482 0.082 0.104 0.068 0.070 0.062 0.041 0.059 0.041 0.191 0.589
Continuous 0.032 0.042 0.117 0.068 0.077 0.352 0.074 0.417 0.070 0.424 0.158 0.188 0.148 0.156 0.137 0.103 0.130 0.090 0.264 0.714
Ordered 0.030 0.042 0.126 0.066 0.079 0.292 0.076 0.353 0.074 0.365 0.162 0.237 0.147 0.173 0.138 0.122 0.131 0.106 0.257 0.688
γ : High β : High
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Linear PS Model, Weighting, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary 0.079 0.081 0.086 0.290 0.077 0.119 0.061 0.047 0.054 0.061 0.078 0.257 0.150 0.329 0.102 0.202 0.076 0.097 0.163 0.566
Continuous 0.070 0.084 0.086 0.311 0.074 0.094 0.057 0.050 0.051 0.057 0.077 0.269 0.140 0.259 0.095 0.167 0.071 0.091 0.172 0.635
Ordered 0.072 0.084 0.104 0.329 0.070 0.080 0.059 0.052 0.058 0.072 0.095 0.291 0.130 0.236 0.089 0.144 0.068 0.079 0.183 0.628
γ : Low β : Low
Binary 0.079 0.081 0.088 0.289 0.077 0.121 0.059 0.047 0.053 0.061 0.080 0.256 0.148 0.332 0.102 0.203 0.075 0.098 0.165 0.565
Continuous 0.073 0.080 0.086 0.312 0.074 0.094 0.058 0.050 0.052 0.055 0.077 0.271 0.140 0.260 0.097 0.169 0.072 0.092 0.171 0.636
Ordered 0.069 0.083 0.104 0.330 0.070 0.078 0.061 0.052 0.059 0.074 0.094 0.291 0.130 0.228 0.090 0.141 0.070 0.076 0.183 0.627
γ : Low β : High
Binary 0.079 0.081 0.122 0.278 0.075 0.135 0.067 0.042 0.069 0.083 0.113 0.240 0.139 0.356 0.095 0.191 0.077 0.080 0.192 0.589
Continuous 0.073 0.086 0.195 0.371 0.115 0.067 0.124 0.058 0.133 0.113 0.186 0.328 0.139 0.226 0.119 0.122 0.117 0.064 0.265 0.713
Ordered 0.070 0.087 0.194 0.394 0.115 0.060 0.125 0.072 0.134 0.127 0.186 0.357 0.138 0.194 0.119 0.104 0.119 0.066 0.257 0.687
γ : High β : Low
Binary 0.080 0.080 0.119 0.277 0.075 0.136 0.066 0.041 0.067 0.082 0.111 0.239 0.144 0.356 0.100 0.191 0.077 0.079 0.191 0.589
Continuous 0.072 0.085 0.195 0.371 0.115 0.069 0.125 0.057 0.134 0.112 0.186 0.328 0.138 0.230 0.119 0.123 0.118 0.064 0.264 0.714
Ordered 0.073 0.084 0.196 0.396 0.118 0.060 0.128 0.075 0.137 0.128 0.188 0.359 0.139 0.195 0.121 0.105 0.121 0.069 0.257 0.688
γ : High β : High
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00

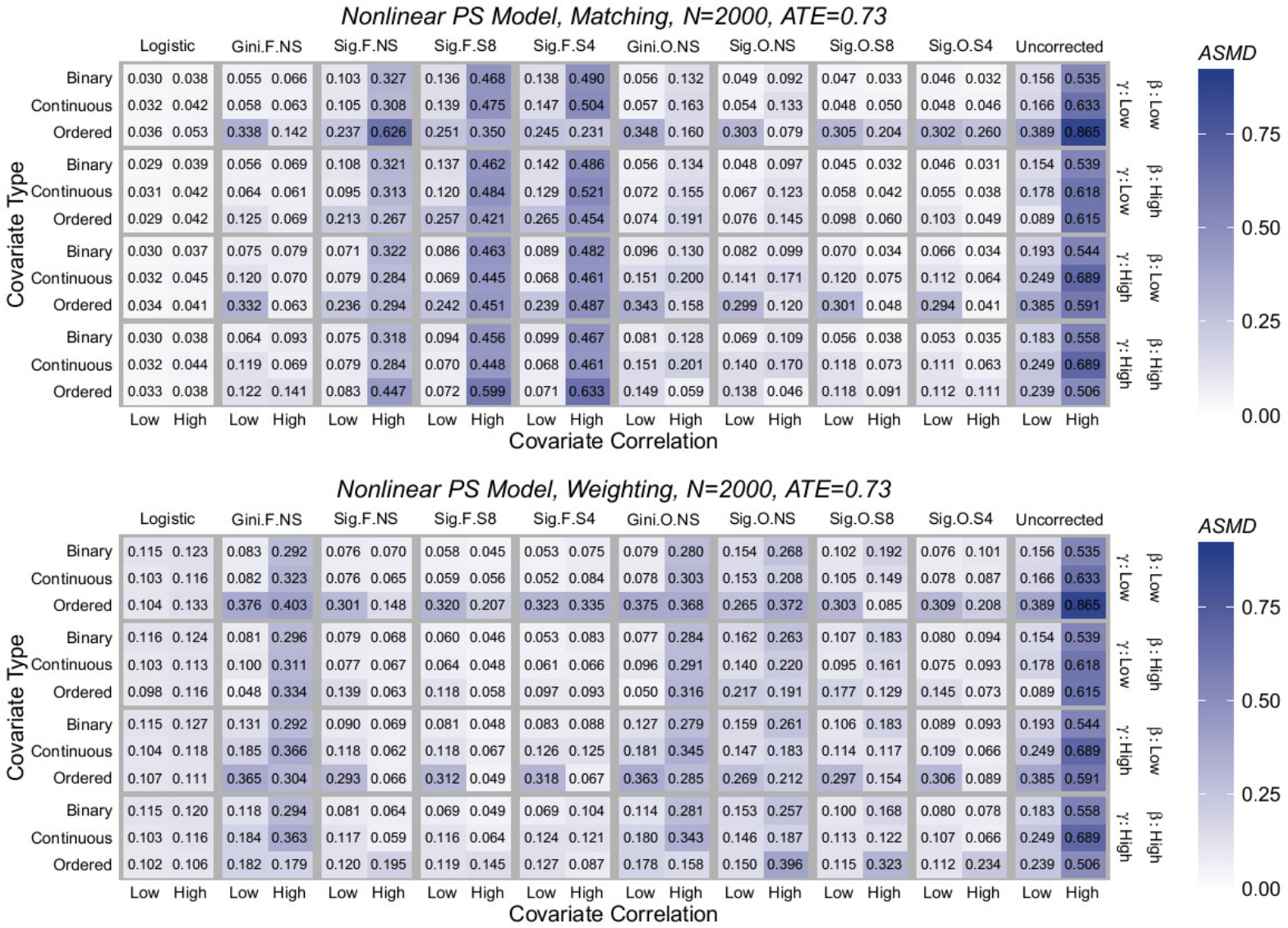

Nonlinear PS Model, Matching, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Nonlinear PS Model, Weighting, N=2000, ATE=0.73
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
γ : Low
β : Low
β : High
γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00

Figure 4. Mean of absolute standardized mean difference (*ASMD*) across replications. The numerical value in each tile is the mean *ASMD* of the condition. As *ASMD* increases, the tile becomes increasingly dark.

Chapter 4

DISCUSSION

In this chapter, the simulation study and its results are discussed in three sections. First, I describe in detail the previous simulation studies by Austin (2012) and Lee et al. (2010) which also investigated the performance of Random Forests method in propensity score estimation. I compare the conditions and procedures of their simulation studies with the simulation study in this dissertation. Second, I summarize and provide an interpretation of the most important findings of the present simulation study. I also compare and explain the similarities and differences of the present results with those of Austin (2012) and Lee et al. (2010). Third, I provide practical suggestions for applied researchers and future research directions.

Previous Simulation Studies

To the best of my knowledge, Austin (2012) and Lee et al. (2010) are the only studies that have investigated the performance of the Random Forests method in propensity score estimation. The simulation study in this dissertation differs from their simulation studies in several important aspects. First, Austin and Lee et al. investigated the effects of a very limited set of Random Forests model specifications, focusing particularly on the effects of random sampling of covariates. For no sampling (i.e., all covariates included), both studies used the R package ipred. To sample a random subset of covariates, both studies used the R package randomForest. Austin set the number of covariates in the random subset equal $k / 3$, where k is the total number of covariates. Lee et al. set the number of covariates in the random subset equal $\sqrt{k}$, following suggestions by Hastie et al. (2009, p. 592) and Liaw and Wiener (2002). However, as mentioned in

footnote 5 in this dissertation, the randomForest package uses a different procedure to calculate the Random Forests propensity scores than the ipred package. These different methods of calculating the Random Forests propensity scores in the two R packages potentially confound the interpretation of the results of the effect of this model specification in these two studies. Other than random sampling of covariates, these two studies used the default model specifications offered by these two software packages: (a) Gini index decision rule, (b) out-of-bag sample to estimate the propensity scores in each Classification Tree, (c) 500 Classification Trees for the Random Forests model, (d) random sampling of the data with replacement using same sample size as the original data, (e) maximum depth of Classification Tree was not controlled, and (f) the minimum terminal node size was set to 20 in ipred and 1 in randomForest. In contrast to the limited investigation of the effects of Random Forests model specifications in Austin and Lee et al., this dissertation provided a systematic investigation of the effects of a large array of potentially key Random Forests model specifications in propensity score estimation.

Second, the two simulation studies investigated two different weighting adjustment methods to equate the treatment and control groups. Austin (2012) investigated the inverse probability of treatment weighting method, which estimates the average treatment effect of the entire sample. Lee et al. (2010) investigated the weighting by the odds method, which estimates the average treatment effect of the participants who are actually assigned to the treatment group (the average treatment effect on the treated; see also footnote 3 in this dissertation). The present simulation study focused on two equating methods, nearest neighbor matching and weighting by the odds; both methods estimate the average treatment effect of the participants who are actually assigned to the

treatment group. Nearest neighbor matching has a potential disadvantage that, when not *all* treatment group participants can be successfully matched with control group participants in the sample data, the parameter estimate is not exactly the average treatment effect of the participants who are actually assigned to the treatment group.

Third, Austin (2012) and Lee et al. (2010) adopted the data generation model used by Setoguchi et al. (2008). In the population models for their simulation studies, there were 6 binary and 4 standard normal covariates. Among these 10 covariates, only 4 pairs of covariates were correlated. Both studies investigated the same set of the population propensity score models, which included a linear logistic regression model and four nonlinear logistic regression models with additional combinations of quadratic and two-way interaction effects. Lee et al. investigated two additional nonlinear logistic regression model specifications. Both studies investigated the same type of population treatment-outcome model, a linear regression model. Austin introduced a slight change by adding a normally distributed residual term in the population treatment-outcome model. Austin also investigated two nonlinear treatment-outcome models with additional combinations of quadratic and two-way interaction effects. The simulation study in this dissertation was designed based on the empirical example by Im et al. (2013). Compared with the population models investigated by Austin (2012) and Lee et al. (2010), the population models in this dissertation would be more challenging for the Random Forests method in the following ways. (a) More covariates with different types of covariates were included: 16 binary, 8 ordinal, and 40 standard normal covariates. (b) All the covariates were inter-correlated at manipulated levels of correlation in the population (low, high). (c) The regression coefficients of the covariates in the propensity score model and treatment-

outcome model was also systematically manipulated at two levels (low, high). These conditions were designed to make it challenging for the Random Forests method to identify the optimal covariate to select and split at each step of the construction of the Classification Tree.

Fourth, Austin (2012) studied a single sample size condition of 1000; Lee et al. (2010) studied sample size conditions of 500, 1000, and 2000. The simulation study in this dissertation investigated the sample size conditions of 600 and 2000.

Simulation Study Results

In this section, the simulation study results are discussed in three parts: benchmark methods, Random Forests methods in propensity score estimation, and the achieved level of balance in the covariates' distributions.

Benchmark methods. In the simulation study, there were three benchmark methods. The first benchmark method was the uncorrected method in which none of the covariates in the true population models were included in the linear regression model to estimate the average treatment effect. This uncorrected benchmark method showed that the population models successfully introduced bias in the *ATE* estimate and imbalance in the covariates' distributions. The uncorrected method overestimated the *ATE* for both sample sizes (600, 2000), propensity score models (linear, nonlinear), and magnitude of the unstandardized *ATE* effects (0, 0.73) conditions with a percent bias of about 200% (Appendix A). Figure 4 showed that there were larger mean differences between the treatment and control groups on the covariates in the high level covariate correlation condition than in the low level covariate correlation condition. Although not presented,

the uncorrected results also showed that there were no variance differences between the treatment and control groups on the covariates.

The second benchmark method was analysis of covariance (ANCOVA) which used the covariates and functional form of the covariates corresponding to the true population treatment-outcome (linear regression) model to estimate the *ATE*. The ANCOVA method represented the true population data generating model; it was theoretically expected to produce unbiased *ATE* estimates. The simulation results showed that, across all sample sizes, propensity score models, and *ATE* effects conditions, the ANCOVA method produced unbiased *ATE* estimates, the minimal sampling variability compared to other propensity score procedures (logistic regression, Random Forests; matching, weighting), and an unbiased estimate of the true standard error. Despite these appealing results in the ideal case, the ANCOVA method is not suggested in practice. In practice, analysts do not know the true population treatment-outcome model; the ANCOVA model is specified according to prior research, theory, and model modification based on the sample data set. The likelihood of misspecification of the true population treatment-outcome model is high, particularly when there are a large number of covariates. Permitting model modification based on results with the sample data set may risk modifications that favor the analysts' preferred results for the *ATE* (Rubin, 2001). On the other hand, examination and modification of the propensity score estimation models involve only the examination of balance diagnostics for the covariates; it does not involve the outcome variable. ANCOVA also permits extrapolation of the estimate of the *ATE* beyond the region of common support of the covariates, a highly risky practice if the treatment outcome model is not linear. Even though the ANCOVA uses the true

population treatment-outcome model, it estimates the average treatment effect of the entire sample, which is different than the *ATE* estimates using the matching and weighting adjustments. Schafer and Kang (2008) present a fuller discussion of the theoretical and practical strengths and limitations of ANCOVA.

The third benchmark method was the true population logistic regression propensity score models. This benchmark method provided three types of information. (a) Given asymptotic propensity score theory holds and nearest neighbor matching or weighting by the odds adjustment is conducted, this method was expected to produce unbiased *ATE* estimates. (b) Given the strong ignorability assumption was met and the propensity score estimation model was correctly specified, the method permitted an examination whether the *ATE* estimate was unbiased in the finite sample sizes considered in this dissertation, as not all treatment group participants would be successfully matched with a control group participant. (c) This method was theoretically expected to provide the *optimal* absolute standardized mean differences of the covariates after the matching and weighting adjustments (Figure 4).

Using both the nearest neighbor matching and weighting by the odds adjustment methods, the logistic regression propensity scores produced unbiased *ATE* estimates across all sample sizes, propensity score models, and *ATE* effects conditions. Although not all treatment group participants were successfully matched with a control group participant under nearest neighbor matching (not presented), this did not lead to biased *ATE* estimates. The sampling variability of the *ATE* estimate using nearest neighbor matching was smaller than that produced using weighting by the odds method. One possible explanation of the large sampling variability using the weighting by the odds

method is the extreme low or high values of weights of some control group participants (Schafer & Kang, 2008). As a check on the above statement, I randomly selected one simulated data set from the $N = 600$, linear propensity score model, $ATE = 0.73$ condition and compared the range of the weights for the control group participants of the logistic regression propensity scores and those produced for the true population propensity scores: logistic regression weights (range = 0.00 - 78.36), true population weights (range 0.00 - 12.96). Although the weights produced by the logistic regression propensity scores led to an unbiased *ATE* estimate, the extreme values of the weights of the control group participants led to large sampling variability of the *ATE* estimate. The standard error of the *ATE* estimate was overestimated using matching, whereas it was underestimated using weighting. The bias of the estimated standard error using matching (overestimation) and weighting (underestimation) occurred because the estimate of the standard error did not account for the uncertainty in the propensity scores estimation process that contributes to the sampling variability of the *ATE* estimate. The overestimation of the standard error using matching has been noted by Rubin and Thomas (1996), and the underestimation of the standard error using weighting has been noted by Schafer and Kang (2008). Although correction formulas for the estimated standard error for the logistic regression propensity scores using matching (Rubin & Thomas, 1996) and weighting (Robins, Rotnitzky, & Zhao, 1995; Schafer & Kang, 2008) have been developed, the formulas are complex and have not been implemented in any software package to date.

The logistic regression benchmark method provided the *minimal* absolute standardized mean differences of the covariates after the matching and weighting

adjustments. Figure 4 showed that, the logistic regression propensity scores after matching and weighting adjustments produced small absolute standardized mean differences (*ASMD*) that satisfied Rubin's (2001) guideline of *ASMD* < 0.25. The *ASMD*s were often larger in the $N = 600$ than those in the $N = 2000$ condition. Note also that the *ASMD*s using weighting were larger than those using matching across sample sizes, propensity score models, *ATE* effects, and covariates conditions. The results suggested that the analysts should *not* choose between nearest neighbor matching or weighting by the odds method by comparing the *ASMD* of the covariates using the same set of the estimated propensity scores.

Random Forests method in propensity score estimation. One of the primary goals of this dissertation is the effects of some potential key Random Forests model specifications on propensity score estimation. The investigation of the Random Forests model specifications provides information about procedures that can be implemented by researchers in practice. I hypothesized that the optimal empirical Random Forests model specification(s) would produce a minimally biased average treatment effect (*ATE*) estimate. I expected, based on previous literature on Random Forests (Berk, 2008; Breiman, 2001; Hastie et al., 2009; Hothorn, Hornik, et al., 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007; Strobl et al., 2009), that conditions under which the Random Forests specifications would produce less biased estimates would include: (1) the conditional significance test decision rule (compared to the Gini index), (2) random sampling of covariates (compared to using all covariates), and (3) the use of out-of-bag (hold out) samples (compared to the full sample data). The simulation results differed for the two sample sizes and for the matching and weighting methods. For sample size $N =$

600, Sig.O.S8 using weighting produced an unbiased *ATE* estimate. For $N = 2000$, Gini.F.NS using matching, Sig.F.NS using weighting, and Sig.O.S4 using weighting produced unbiased average treatment effect estimate.

Using the nearest neighbor matching method, the two hypotheses above were not supported. The Random Forests method was not likely to work well in combination with nearest neighbor matching to produce an unbiased *ATE* estimate. For the $N = 600$ condition, none of the Random Forests specifications produced unbiased *ATE* estimates. Given the biased *ATE* estimates, the results regarding statistical inference were not of the central interest. For the $N = 2000$ condition, Gini.F.NS produced an unbiased *ATE* estimate, but its standard error of the *ATE* estimate was underestimated. Two possible explanations for the disappointing results of the Random Forests method in combination with matching may be considered. First, the Random Forests model specifications did not consistently match all of the treatment group participants successfully with control group participants. But, this reason cannot provide a full explanation for the biased *ATE* estimates: The benchmark logistic regression propensity scores results showed that the *ATE* estimate was unbiased even when not all treatment group participants could be successfully matched with the control group participants in all replications. Second, it appears that the nearest neighbor matching setting in the present study may be very sensitive to the *optimal* Random Forests model specifications. One clue is provided by the number of matched participants by different Random Forests specifications. Across all sample sizes, propensity score models, and *ATE* effects conditions, the out-of-bag specification had more matched participants than the benchmark logistic regression propensity scores, whereas the full sample specification had fewer matched participants

than the logistic regression propensity scores. The Gini.F.NS was the only specification that produced an unbiased *ATE* estimate in the $N = 2000$ condition; in this specification the number of matched participants was close to that of the logistic regression benchmark method. These results suggest that the models specifications investigated were not optimal to match the performance of the logistic regression benchmark method. Further investigation is required to examine the optimal set of the Random Forest model specifications and the nearest neighbor matching settings. One key specification that may improve the performance of the Random Forest method with nearest neighbor matching is the minimum terminal node size. Yet it is uncertain whether increasing or decreasing the minimum terminal node size will improve the performance. Zhao (2008) simulation study showed that matching with replacement produced less biased *ATE* estimate, when the logistic regression propensity score model was misspecified. It is speculate that matching with replacement is also less sensitive to the Random Forests model specifications.

More support for the hypotheses was found using the weighting by the odds method. Under several model specifications in both the $N = 600$ and $N = 2000$ conditions, Random Forests successfully produced unbiased *ATE* estimates. Particularly encouraging was that the combination of model specification conditions of (a) the conditional significance test, (b) out-of-bag samples, and (c) random sampling of covariates produced unbiased *ATE* estimates. The results showed that the same Random Forests model specifications performed consistently across the linear and nonlinear propensity score models. What differed between these optimally performing conditions was the number of

covariates to be randomly sampled: 8 covariates for $N = 600$ and 4 covariates for $N = 2000$.

The simulation study by Lee et al. (2010) also investigated the Random Forests method using the weighting by the odds method. Their results showed that random sampling of covariates at $\sqrt{k}$ (where k is the total number of covariates) produced less biased *ATE* estimates (percent bias < 10%) than no random sampling of covariates (i.e., selecting all covariates) across all sample sizes (500, 1000, and 2000) and propensity score model conditions. Yet the percent bias of the *ATE* estimates of no random sampling of covariates was still around 10%. Based on the results in this dissertation and the results by Lee et al., the number of covariates to be randomly sampled appears to be sensitive to the sample size, number of covariates, and types of covariates of the data. I speculate that, with fewer covariates, the Random Forest method will be less sensitive to the number of covariates to be sampled. With more covariates, fewer covariates should be sampled as sample size increases. To the best of my knowledge, no study has systematically investigated the effects of the total number of covariates and the number to be randomly sampled using the Random Forests method. Further work is needed to study the performance of this combination of (a) the conditional significance test, (b) out-of-bag samples, and (c) random sampling of covariates with different numbers of covariates using a variety of different population models to more fully understand the performance of this Random Forests specification.

A second advantage of the Random Forests method when used with the weighting by the odds method is increased efficiency, as reflected in the standard error of the *ATE* estimate. The Random Forests method produced *ATE* estimates that had far less sampling

variability than even the logistic regression method using the true population propensity score model. One possible explanation of the decreased sampling variability of the Random Forest method is that it produced less extreme values of the weights of the control group participants (Schafer & Kang, 2008). As a check, I investigated the range of the weights of the control group participants of the propensity scores in one selected simulated data set above ($N = 600$, linear propensity score model, $ATE = 0.73$): Sig.O.S8 weights (0.00 - 18.56), logistic regression weights (0.00 - 78.36), true population weights (0.00 - 12.96). One reason why the weights of the Random Forests propensity scores were less extreme than those of the benchmark logistic regression approach is that the Random Forests propensity scores represent the average of the Classification Tree propensity scores from multiple random samples drawn from the original data.

The third advantage of the Random Forests method when used with the weighting by the odds method is that, the Random Forests model specifications that produced unbiased estimates of the *ATE* also produced unbiased the estimates of the standard error of the *ATE* across sample sizes, propensity scores models, and *ATE* effects conditions. Given an unbiased *ATE* estimate and its unbiased estimated standard error, statistical inferences based on either the confidence interval or hypothesis testing approaches were correct within sampling error at the specified α level. Lee et al. (2010) found that the coverage rate of *ATE* confidence interval was above the specified α level (conservative). On the other hand, in the present study the benchmark logistic regression method using the true population propensity score model *underestimated* the standard error of *ATE* estimate when weighting was used. As discussed earlier, the standard error estimation formulas used in the present simulation study did not correct for the uncertainty of the

estimated propensity scores. One possible reason is the variability of the Random Forests weights are similar to that of the true population weights. Using the same randomly selected data set described above, the standard deviation of the Random Forests weights = 2.06, true population weights = 1.81, logistic regression weights = 4.79.

Austin (2012) investigated the Random Forests method using the inverse probability of treatment weighting adjustment. His results showed that the *ATE* estimates were biased across different propensity score models. Based on the discussion above, it may be that the Random Forests model specification investigated by Austin (2012) was not optimal to produce unbiased ATE estimates using the inverse probability of treatment weighting adjustment.

Balance on covariates' distributions. I also investigated the ability of the Random Forests method to produce balance on the covariates following equating on propensity scores. Following Rosenbaum and Rubin (1983), it has widely been assumed that an informative procedure for comparing different propensity score estimation methods is to compare the balance on the covariates' distributions (e.g., in terms of means, variances, etc.) between treatment conditions (e.g., Dehijia & Wahba, 2002; Lee et al., 2010; Rubin, 2001; West et al., 2013). Given that the strong ignorability assumption is met in the population (as it was in the present simulation study) and that a specified method of equating groups (e.g., matching; weighting) is used, the method of estimating propensity scores that produces the better balance of the covariates' distributions has been believed to produce a less biased average treatment effect estimate. Consistent with this reasoning, Figure 4 of the present simulation results showed that, under the logistic regression benchmark method and optimal Random Forests model

specifications under which the *ATE* estimates were unbiased, there were no substantial mean differences of the covariates between treatment conditions following equating. In addition, those Random Forests model specifications and adjustment methods that produced a very large percentage bias in the *ATE* estimate (percent bias > 60% in magnitude) failed to balance the covariates means. However, the Random Forests model specifications that produced a smaller, but still unsatisfactory percent bias in the *ATE* estimates (percent bias ranging from 10% to 60% in magnitude) balanced the covariates means similarly to the optimal specifications. I conclude that comparing the balance on the covariates' means is a procedure that can detect estimation and adjustment methods that will produce a grossly biased estimate of the *ATE*. However, in the present study, the procedure was not sufficiently sensitive to differentiate between Random Forests models that produced optimal versus less biased, but still unsatisfactory estimates of the *ATE*. Lee et al. (2010) reached a similar conclusion that the degree of balance achieved was not sufficiently sensitive to differentiate between different propensity score estimation methods that produced versus did not produce unbiased *ATE* estimates. Although other procedures of investigating the balance of different aspects of the univariate and bivariate covariates' distributions (e.g., variances, univariate QQ plot, two-way interactions means) have been proposed (Austin, 2009), none appear to be capable of fully assessing the balance of the multivariate distribution of the covariates between the treatment conditions. The results of the present study call for the development of new procedures that are sufficiently sensitive to identify the optimal model specifications in simulation studies comparing different propensity score models.

This dissertation also investigated the question of whether the balance on the covariate means was related to the magnitude of bias of the *ATE* estimate across repeated samples within a Random Forests model specification. Figure 3 of the simulation results showed that for the optimal Random Forests specifications that produced unbiased *ATE* estimate, there were not substantial correlations between balance on the covariates means and the magnitude of bias in the *ATE* across repeated samples. One problem in the use of correlations as an index may be the reduction in range that occurs when the variance in the *ATE* estimates and the balance on the covariates means are reduced, severely attenuating the correlations. This problem may lead to insensitivity of the correlation coefficient after the propensity score equating procedure has produced comparable groups. Thus, the degree of balance on the covariates following propensity score equating provides little information about which sample would have smaller bias of the average treatment effect estimate.

Conclusion

In conclusion, the results of my dissertation results provided evidence that the Random Forests method can be successfully applied to propensity score analysis to produce an unbiased average treatment effect, unbiased standard error estimate, and proper statistical inference. Given the use of weighting to equate groups, the combination of (a) the conditional significance test, (b) out-of-bag samples, and (c) random sampling of covariates can produce unbiased *ATE* estimates. Additional research is needed to clarify the number of covariates that should be included in the random sample. New procedures beyond assessing the balance of different aspects of the univariate and

bivariate covariates' distributions are needed to help identify the optimal Random Forest specifications for propensity score analysis.

REFERENCES

Amatya, A., & Demirtas, H. (2012). BinNor: Simultaneous generation of multivariate binary and normal variates. (R package version 2.0). [Computer software]. Retrieved March 20, 2013, from http://cran.r-project.org/web/packages/BinNor/index.html

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*(25), 3083–3107.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. *Multivariate Behavioral Research*, *47*(1), 115–135.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Bostrom, H. (2007). Estimating class probabilities in Random Forests (pp. 211–216). Presented at the Sixth International Conference on Machine Learning and Applications, Cincinnati, OH. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4457233

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman and Hall.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). NJ: Hillsdale: Lawrence Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

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

Demirtas, H., & Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*, *22*(2), 223–236.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*(4), 1231–1236.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055–2061.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Im, M. H., Hughes, J. N., Kwok, O-M., Puckett, S., & Cerda, C. A. (2013). *Effect of retention in elementary grades on transition to middle school*. Manuscript submitted for publication.

Kim, H., & Loh, W. Y. (2001). Classification Trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589–604.

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

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lohr, S. (2010). *Sampling: Design and analysis* (2nd ed.). Boston: Brooks/Cole, Cengage Learning.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores. *Evaluation Review*, *29*(6), 530–558.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.

Lumley, T. (2012). *survey: analysis of complex survey samples*. (R package version 3.29). [Computer software]. Retrieved March 7, 2013, from http://cran.r-project.org/web/packages/survey/index.html

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*(302), 415–434.

Peters, A., & Hothorn, T. (2012). *ipred: improved predictors*. (R package version 0.9-1). [Computer software]. Retrieved March 7, 2013, from http://cran.r-project.org/web/packages/ipred/index.html

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved March 7, 2013, from http://www.R-project.org/

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*(429), 106-121.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*(1), 159-183.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3), 169–188.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, *17*(2), 241–264.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

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*(6), 546–555.

Strobl, C., Boulesteix, A. L., & Augustin, T. (2007). Unbiased split selection for Classification Trees based on the Gini index. *Computational Statistics & Data Analysis*, *52*(1), 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: Illustrations, 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*, *25*(1), 1–21.

Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan (2011). IVEware: Imputation and variance estimation software. (Version 0.2) [Computer software]. Retrieved January 20, 2013, from http://www.isr.umich.edu/src/smp/ive/

West, S. G., Cham, H., & Liu, Y. (in press). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (2nd Ed.). New York: Cambridge University Press.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Weiler, M., & Schulze, J. (2013). *Propensity scores for equating groups: Basic principles and applications in clinical treatment outcome research*. Manuscript submitted for publication.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, *15*(1), 18–37.

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*(8), 826–833.

White, A. P., & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, *15*(3), 321–329.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *WJ-III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications.*Economics Letters*, *98*(3), 309-319.

APPENDIX A

SIMULATION STUDY:

AVERAGE TREATMENT EFFECT ESTIMATION RESULTS

$N = 600$, $ATE = 0$

Propensity Score Model	Adjustment Method	Propensity Score	Mean *ATE* Estimate	Percent Bias of *ATE* Estimate	Mean Square Error of *ATE* Estimate	Percent Bias of Estimated SE of *ATE* Estimate	Coverage Rate	Empirical Type I Error Rate
Linear	Matching	ANCOVA	0.000	N/A	0.007	2.2%	0.959	0.041
Linear	Matching	Logistic	0.080	N/A	0.021	37.4%	0.975	0.025
Linear	Matching	Gini.F.NS	-1.066	N/A	1.208	-9.7%	0.018	0.982
Linear	Matching	Sig.F.NS	-0.746	N/A	0.606	-11.2%	0.060	0.940
Linear	Matching	Sig.F.S8	-0.977	N/A	0.991	3.0%	0.000	1.000
Linear	Matching	Sig.F.S4	-1.032	N/A	1.098	7.3%	0.001	0.999
Linear	Matching	Gini.O.NS	0.407	N/A	0.178	25.8%	0.133	0.867
Linear	Matching	Sig.O.NS	0.427	N/A	0.196	22.6%	0.123	0.877
Linear	Matching	Sig.O.S8	0.256	N/A	0.078	30.9%	0.606	0.394
Linear	Matching	Sig.O.S4	0.212	N/A	0.056	33.1%	0.735	0.265
Linear	Matching	Uncorrected	1.501	N/A	2.267	-2.4%	0.000	1.000
Linear	Weighting	ANCOVA	0.000	N/A	0.007	2.2%	0.959	0.041
Linear	Weighting	Logistic	0.061	N/A	0.143	-30.8%	0.813	0.187
Linear	Weighting	Gini.F.NS	0.799	N/A	0.652	17.0%	0.001	0.999
Linear	Weighting	Sig.F.NS	0.135	N/A	0.051	6.4%	0.849	0.151
Linear	Weighting	Sig.F.S8	0.287	N/A	0.105	13.6%	0.583	0.417
Linear	Weighting	Sig.F.S4	0.417	N/A	0.192	15.7%	0.236	0.764
Linear	Weighting	Gini.O.NS	0.597	N/A	0.374	15.7%	0.035	0.965
Linear	Weighting	Sig.O.NS	-0.188	N/A	0.132	-16.4%	0.914	0.086
Linear	Weighting	Sig.O.S8	0.039	N/A	0.050	-1.1%	0.913	0.087
Linear	Weighting	Sig.O.S4	0.229	N/A	0.083	8.6%	0.698	0.302
Linear	Weighting	Uncorrected	1.501	N/A	2.267	-2.4%	0.000	1.000
Nonlinear	Matching	ANCOVA	-0.002	N/A	0.007	0.9%	0.954	0.046
Nonlinear	Matching	Logistic	0.079	N/A	0.024	32.5%	0.973	0.027
Nonlinear	Matching	Gini.F.NS	-0.992	N/A	1.058	-14.0%	0.031	0.969
Nonlinear	Matching	Sig.F.NS	-0.669	N/A	0.514	-22.5%	0.158	0.842
Nonlinear	Matching	Sig.F.S8	-1.033	N/A	1.100	9.0%	0.000	1.000
Nonlinear	Matching	Sig.F.S4	-1.093	N/A	1.228	8.7%	0.000	1.000
Nonlinear	Matching	Gini.O.NS	0.368	N/A	0.147	27.4%	0.194	0.806
Nonlinear	Matching	Sig.O.NS	0.407	N/A	0.181	14.2%	0.160	0.840
Nonlinear	Matching	Sig.O.S8	0.199	N/A	0.051	31.4%	0.754	0.246
Nonlinear	Matching	Sig.O.S4	0.157	N/A	0.036	26.8%	0.862	0.138
Nonlinear	Matching	Uncorrected	1.448	N/A	2.109	3.0%	0.000	1.000
Nonlinear	Weighting	ANCOVA	-0.002	N/A	0.007	0.9%	0.954	0.046
Nonlinear	Weighting	Logistic	0.103	N/A	0.239	-41.8%	0.713	0.287
Nonlinear	Weighting	Gini.F.NS	0.751	N/A	0.580	18.1%	0.006	0.994
Nonlinear	Weighting	Sig.F.NS	0.210	N/A	0.077	9.2%	0.772	0.228
Nonlinear	Weighting	Sig.F.S8	0.310	N/A	0.117	17.7%	0.519	0.481
Nonlinear	Weighting	Sig.F.S4	0.436	N/A	0.206	20.9%	0.184	0.816
Nonlinear	Weighting	Gini.O.NS	0.580	N/A	0.353	19.5%	0.039	0.961
Nonlinear	Weighting	Sig.O.NS	-0.144	N/A	0.134	-17.6%	0.918	0.082
Nonlinear	Weighting	Sig.O.S8	0.020	N/A	0.057	-3.4%	0.914	0.086
Nonlinear	Weighting	Sig.O.S4	0.205	N/A	0.076	8.0%	0.747	0.253
Nonlinear	Weighting	Uncorrected	1.448	N/A	2.109	3.0%	0.000	1.000

$N = 600$, $ATE = 0.73$

Propensity Score Model	Adjustment Method	Propensity Score	Mean *ATE* Estimate	Percent Bias of *ATE* Estimate	Mean Square Error of *ATE* Estimate	Percent Bias of Estimated SE of *ATE* Estimate	Coverage Rate	Empirical Type I Error Rate
Linear	Matching	ANCOVA	0.733	0.4%	0.008	-1.9%	0.942	1.000
Linear	Matching	Logistic	0.805	10.3%	0.019	38.9%	0.983	1.000
Linear	Matching	Gini.F.NS	-0.342	-146.8%	1.213	-5.4%	0.012	0.310
Linear	Matching	Sig.F.NS	-0.019	-102.6%	0.608	-9.3%	0.057	0.087
Linear	Matching	Sig.F.S8	-0.243	-133.3%	0.978	11.0%	0.000	0.216
Linear	Matching	Sig.F.S4	-0.291	-139.9%	1.073	12.6%	0.000	0.289
Linear	Matching	Gini.O.NS	1.140	56.1%	0.180	26.9%	0.122	1.000
Linear	Matching	Sig.O.NS	1.156	58.4%	0.195	24.5%	0.125	1.000
Linear	Matching	Sig.O.S8	0.990	35.6%	0.080	30.5%	0.578	1.000
Linear	Matching	Sig.O.S4	0.946	29.7%	0.059	29.1%	0.723	1.000
Linear	Matching	Uncorrected	2.230	205.4%	2.260	4.9%	0.000	1.000
Linear	Weighting	ANCOVA	0.733	0.4%	0.008	-1.9%	0.942	1.000
Linear	Weighting	Logistic	0.766	4.9%	0.170	-36.0%	0.803	0.783
Linear	Weighting	Gini.F.NS	1.524	108.7%	0.644	20.5%	0.000	1.000
Linear	Weighting	Sig.F.NS	0.859	17.7%	0.049	7.0%	0.864	0.964
Linear	Weighting	Sig.F.S8	1.011	38.4%	0.101	12.6%	0.589	0.998
Linear	Weighting	Sig.F.S4	1.141	56.2%	0.186	16.1%	0.227	1.000
Linear	Weighting	Gini.O.NS	1.320	80.9%	0.366	17.1%	0.047	1.000
Linear	Weighting	Sig.O.NS	0.539	-26.1%	0.128	-14.3%	0.913	0.611
Linear	Weighting	Sig.O.S8	0.762	4.4%	0.053	-3.6%	0.922	0.869
Linear	Weighting	Sig.O.S4	0.953	30.6%	0.082	5.3%	0.700	0.962
Linear	Weighting	Uncorrected	2.230	205.4%	2.260	4.9%	0.000	1.000
Nonlinear	Matching	ANCOVA	0.727	-0.4%	0.007	-0.9%	0.948	1.000
Nonlinear	Matching	Logistic	0.812	11.2%	0.023	37.9%	0.980	1.000
Nonlinear	Matching	Gini.F.NS	-0.285	-139.0%	1.105	-15.3%	0.024	0.259
Nonlinear	Matching	Sig.F.NS	0.037	-94.9%	0.542	-20.0%	0.110	0.127
Nonlinear	Matching	Sig.F.S8	-0.323	-144.3%	1.145	5.8%	0.000	0.353
Nonlinear	Matching	Sig.F.S4	-0.375	-151.4%	1.254	8.6%	0.000	0.484
Nonlinear	Matching	Gini.O.NS	1.091	49.5%	0.144	16.9%	0.217	1.000
Nonlinear	Matching	Sig.O.NS	1.135	55.5%	0.179	15.5%	0.160	1.000
Nonlinear	Matching	Sig.O.S8	0.924	26.6%	0.050	27.3%	0.770	1.000
Nonlinear	Matching	Sig.O.S4	0.879	20.4%	0.034	28.6%	0.870	1.000
Nonlinear	Matching	Uncorrected	2.174	197.8%	2.097	-1.8%	0.000	1.000
Nonlinear	Weighting	ANCOVA	0.727	-0.4%	0.007	-0.9%	0.948	1.000
Nonlinear	Weighting	Logistic	0.855	17.1%	0.238	-41.5%	0.697	0.785
Nonlinear	Weighting	Gini.F.NS	1.474	101.9%	0.570	11.0%	0.004	1.000
Nonlinear	Weighting	Sig.F.NS	0.932	27.7%	0.076	4.4%	0.760	0.980
Nonlinear	Weighting	Sig.F.S8	1.034	41.6%	0.115	10.5%	0.544	0.999
Nonlinear	Weighting	Sig.F.S4	1.158	58.7%	0.202	12.4%	0.209	1.000
Nonlinear	Weighting	Gini.O.NS	1.301	78.2%	0.345	10.9%	0.049	1.000
Nonlinear	Weighting	Sig.O.NS	0.585	-19.9%	0.132	-19.1%	0.912	0.623
Nonlinear	Weighting	Sig.O.S8	0.750	2.7%	0.055	-4.2%	0.907	0.843
Nonlinear	Weighting	Sig.O.S4	0.930	27.4%	0.075	4.4%	0.736	0.959
Nonlinear	Weighting	Uncorrected	2.174	197.8%	2.097	-1.8%	0.000	1.000

$N = 2000$, $ATE = 0$

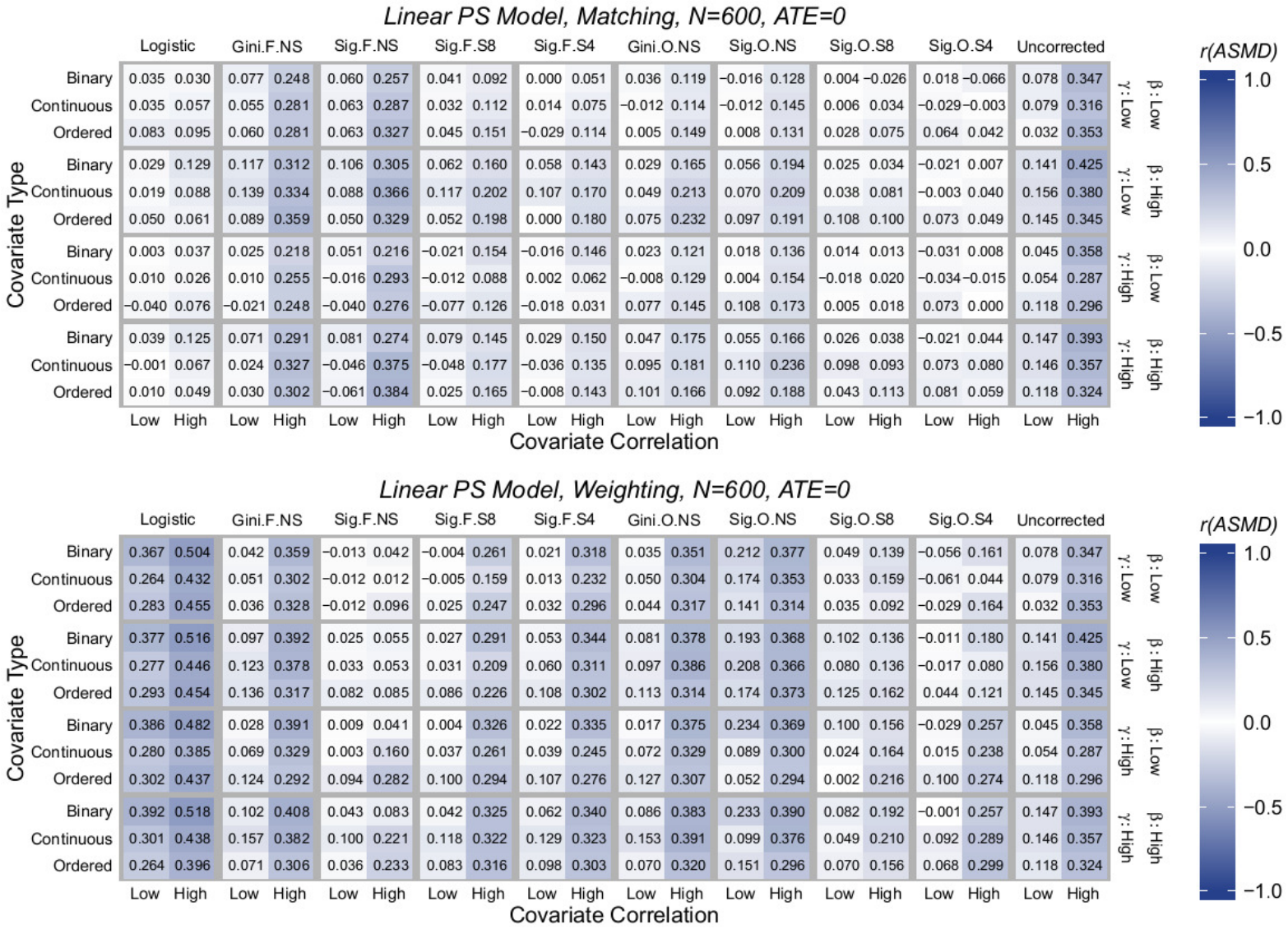
Propensity Score Model	Adjustment Method	Propensity Score	Mean *ATE* Estimate	Percent Bias of *ATE* Estimate	Mean Square Error of *ATE* Estimate	Percent Bias of Estimated SE of *ATE* Estimate	Coverage Rate	Empirical Type I Error Rate
Linear	Matching	ANCOVA	0.001	N/A	0.002	-0.3%	0.947	0.053
Linear	Matching	Logistic	0.079	N/A	0.010	40.9%	0.927	0.073
Linear	Matching	Gini.F.NS	-0.038	N/A	0.015	-21.3%	0.858	0.142
Linear	Matching	Sig.F.NS	-0.710	N/A	0.518	-8.5%	0.000	1.000
Linear	Matching	Sig.F.S8	-0.852	N/A	0.735	9.7%	0.000	1.000
Linear	Matching	Sig.F.S4	-0.895	N/A	0.810	7.1%	0.000	1.000
Linear	Matching	Gini.O.NS	0.433	N/A	0.193	6.8%	0.000	1.000
Linear	Matching	Sig.O.NS	0.356	N/A	0.131	24.4%	0.002	0.998
Linear	Matching	Sig.O.S8	0.237	N/A	0.059	36.6%	0.086	0.914
Linear	Matching	Sig.O.S4	0.195	N/A	0.042	34.0%	0.235	0.765
Linear	Matching	Uncorrected	1.502	N/A	2.259	1.9%	0.000	1.000
Linear	Weighting	ANCOVA	0.001	N/A	0.002	-0.3%	0.947	0.053
Linear	Weighting	Logistic	0.025	N/A	0.040	-18.0%	0.885	0.115
Linear	Weighting	Gini.F.NS	0.815	N/A	0.668	21.4%	0.000	1.000
Linear	Weighting	Sig.F.NS	-0.061	N/A	0.016	14.0%	0.965	0.035
Linear	Weighting	Sig.F.S8	0.117	N/A	0.021	20.7%	0.811	0.189
Linear	Weighting	Sig.F.S4	0.256	N/A	0.072	21.9%	0.204	0.796
Linear	Weighting	Gini.O.NS	0.728	N/A	0.534	22.9%	0.000	1.000
Linear	Weighting	Sig.O.NS	-0.458	N/A	0.267	-14.4%	0.357	0.643
Linear	Weighting	Sig.O.S8	-0.210	N/A	0.073	-3.3%	0.837	0.163
Linear	Weighting	Sig.O.S4	-0.006	N/A	0.016	7.0%	0.960	0.040
Linear	Weighting	Uncorrected	1.502	N/A	2.259	1.9%	0.000	1.000
Nonlinear	Matching	ANCOVA	0.000	N/A	0.002	0.6%	0.953	0.047
Nonlinear	Matching	Logistic	0.088	N/A	0.012	38.0%	0.909	0.091
Nonlinear	Matching	Gini.F.NS	0.088	N/A	0.027	-37.6%	0.703	0.297
Nonlinear	Matching	Sig.F.NS	-0.578	N/A	0.352	-19.6%	0.004	0.996
Nonlinear	Matching	Sig.F.S8	-0.911	N/A	0.840	7.5%	0.000	1.000
Nonlinear	Matching	Sig.F.S4	-0.963	N/A	0.936	2.4%	0.000	1.000
Nonlinear	Matching	Gini.O.NS	0.452	N/A	0.213	-15.9%	0.000	1.000
Nonlinear	Matching	Sig.O.NS	0.373	N/A	0.144	10.8%	0.001	0.999
Nonlinear	Matching	Sig.O.S8	0.175	N/A	0.034	28.2%	0.361	0.639
Nonlinear	Matching	Sig.O.S4	0.134	N/A	0.022	27.1%	0.588	0.412
Nonlinear	Matching	Uncorrected	1.447	N/A	2.098	-0.6%	0.000	1.000
Nonlinear	Weighting	ANCOVA	0.000	N/A	0.002	0.6%	0.953	0.047
Nonlinear	Weighting	Logistic	0.019	N/A	0.087	-33.1%	0.824	0.176
Nonlinear	Weighting	Gini.F.NS	0.795	N/A	0.636	17.5%	0.000	1.000
Nonlinear	Weighting	Sig.F.NS	0.050	N/A	0.016	12.0%	0.932	0.068
Nonlinear	Weighting	Sig.F.S8	0.143	N/A	0.028	24.1%	0.723	0.277
Nonlinear	Weighting	Sig.F.S4	0.275	N/A	0.082	25.7%	0.172	0.828
Nonlinear	Weighting	Gini.O.NS	0.755	N/A	0.574	18.4%	0.000	1.000
Nonlinear	Weighting	Sig.O.NS	-0.367	N/A	0.210	-20.2%	0.662	0.338
Nonlinear	Weighting	Sig.O.S8	-0.222	N/A	0.081	-2.2%	0.850	0.150
Nonlinear	Weighting	Sig.O.S4	-0.033	N/A	0.020	6.8%	0.971	0.029
Nonlinear	Weighting	Uncorrected	1.447	N/A	2.098	-0.6%	0.000	1.000

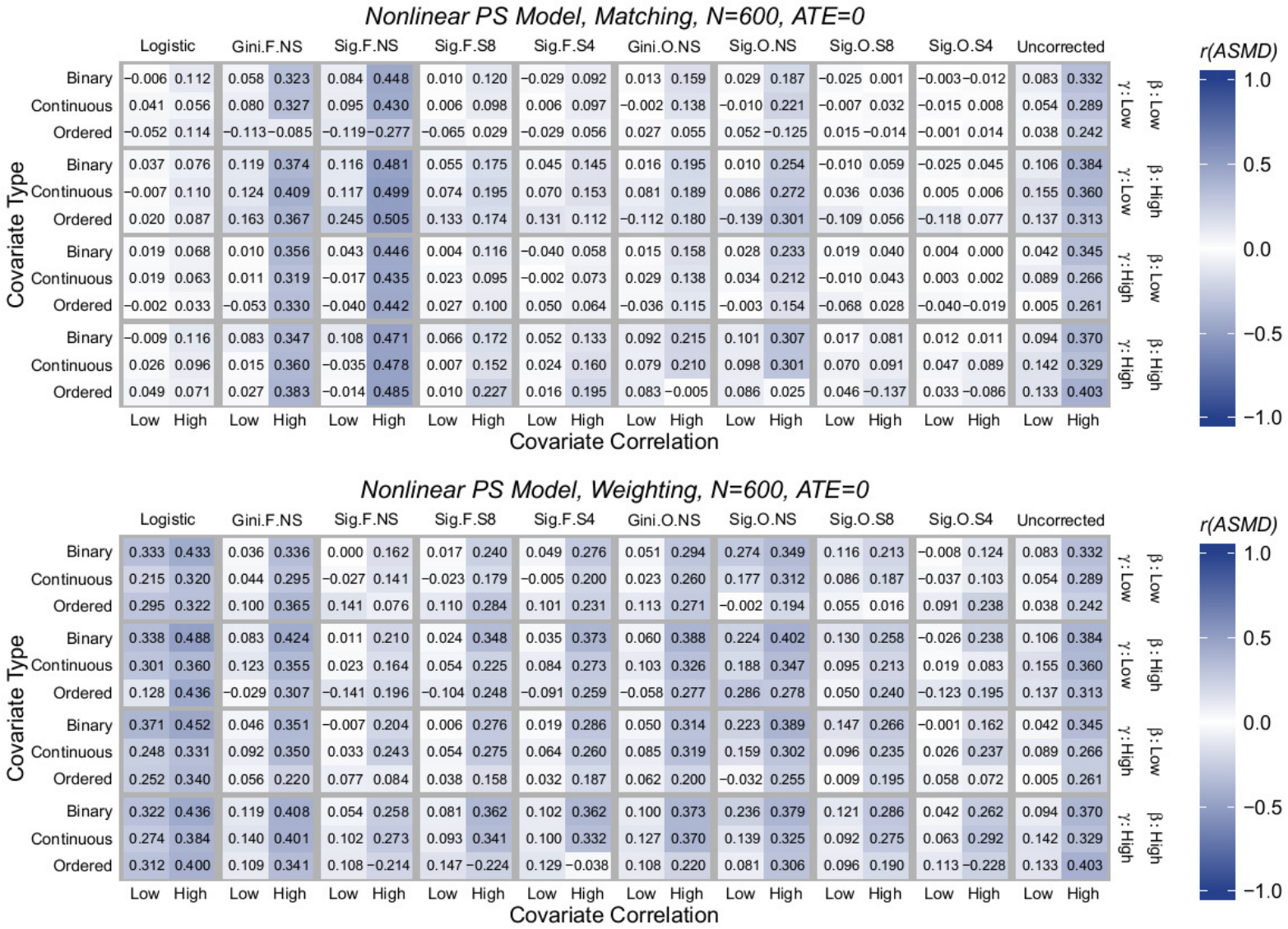
$N = 2000$, $ATE = 0.73$

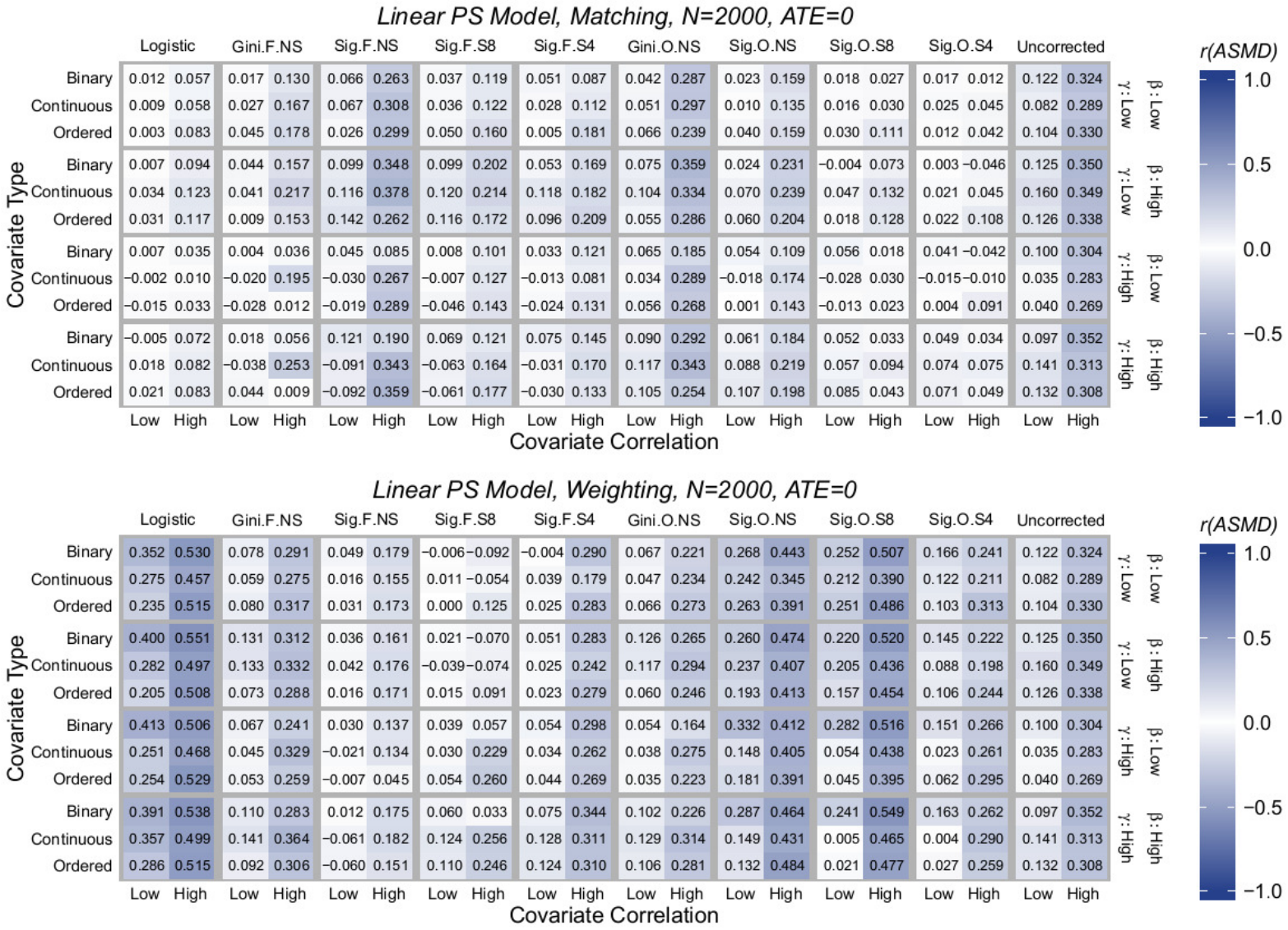
Propensity Score Model	Adjustment Method	Propensity Score	Mean *ATE* Estimate	Percent Bias of *ATE* Estimate	Mean Square Error of *ATE* Estimate	Percent Bias of Estimated SE of *ATE* Estimate	Coverage Rate	Empirical Type I Error Rate
Linear	Matching	ANCOVA	0.729	-0.2%	0.002	-0.7%	0.949	1.000
Linear	Matching	Logistic	0.810	11.0%	0.010	38.5%	0.905	1.000
Linear	Matching	Gini.F.NS	0.691	-5.4%	0.014	-20.7%	0.863	1.000
Linear	Matching	Sig.F.NS	0.017	-97.7%	0.522	-7.3%	0.000	0.080
Linear	Matching	Sig.F.S8	-0.123	-116.8%	0.735	11.7%	0.000	0.187
Linear	Matching	Sig.F.S4	-0.166	-122.8%	0.812	11.8%	0.000	0.364
Linear	Matching	Gini.O.NS	1.163	59.3%	0.193	5.1%	0.000	1.000
Linear	Matching	Sig.O.NS	1.083	48.4%	0.129	22.4%	0.002	1.000
Linear	Matching	Sig.O.S8	0.967	32.4%	0.060	28.0%	0.099	1.000
Linear	Matching	Sig.O.S4	0.926	26.8%	0.042	32.7%	0.226	1.000
Linear	Matching	Uncorrected	2.235	206.1%	2.268	-4.6%	0.000	1.000
Linear	Weighting	ANCOVA	0.729	-0.2%	0.002	-0.7%	0.949	1.000
Linear	Weighting	Logistic	0.743	1.7%	0.043	-17.3%	0.900	0.902
Linear	Weighting	Gini.F.NS	1.544	111.5%	0.667	17.8%	0.000	1.000
Linear	Weighting	Sig.F.NS	0.665	-8.9%	0.017	11.0%	0.964	0.996
Linear	Weighting	Sig.F.S8	0.842	15.4%	0.021	16.4%	0.805	1.000
Linear	Weighting	Sig.F.S4	0.983	34.7%	0.071	18.0%	0.235	1.000
Linear	Weighting	Gini.O.NS	1.457	99.6%	0.533	19.5%	0.000	1.000
Linear	Weighting	Sig.O.NS	0.270	-63.1%	0.272	-15.9%	0.372	0.451
Linear	Weighting	Sig.O.S8	0.507	-30.5%	0.079	-1.4%	0.842	0.779
Linear	Weighting	Sig.O.S4	0.715	-2.1%	0.017	6.8%	0.950	0.962
Linear	Weighting	Uncorrected	2.235	206.1%	2.268	-4.6%	0.000	1.000
Nonlinear	Matching	ANCOVA	0.730	0.0%	0.002	-1.0%	0.949	1.000
Nonlinear	Matching	Logistic	0.813	11.3%	0.011	39.9%	0.920	1.000
Nonlinear	Matching	Gini.F.NS	0.809	10.8%	0.022	-31.7%	0.734	1.000
Nonlinear	Matching	Sig.F.NS	0.148	-79.7%	0.355	-16.5%	0.002	0.304
Nonlinear	Matching	Sig.F.S8	-0.179	-124.5%	0.835	9.6%	0.000	0.415
Nonlinear	Matching	Sig.F.S4	-0.233	-132.0%	0.937	10.2%	0.000	0.652
Nonlinear	Matching	Gini.O.NS	1.176	61.1%	0.206	-9.9%	0.000	1.000
Nonlinear	Matching	Sig.O.NS	1.098	50.4%	0.141	11.1%	0.002	1.000
Nonlinear	Matching	Sig.O.S8	0.905	24.0%	0.034	25.3%	0.352	1.000
Nonlinear	Matching	Sig.O.S4	0.864	18.4%	0.022	26.9%	0.585	1.000
Nonlinear	Matching	Uncorrected	2.175	197.9%	2.092	-3.1%	0.000	1.000
Nonlinear	Weighting	ANCOVA	0.730	0.0%	0.002	-1.0%	0.949	1.000
Nonlinear	Weighting	Logistic	0.757	3.7%	0.080	-32.3%	0.832	0.870
Nonlinear	Weighting	Gini.F.NS	1.521	108.4%	0.631	12.4%	0.000	1.000
Nonlinear	Weighting	Sig.F.NS	0.776	6.3%	0.015	12.3%	0.940	0.999
Nonlinear	Weighting	Sig.F.S8	0.870	19.2%	0.027	20.7%	0.732	1.000
Nonlinear	Weighting	Sig.F.S4	1.002	37.3%	0.080	22.7%	0.159	1.000
Nonlinear	Weighting	Gini.O.NS	1.481	102.8%	0.568	13.6%	0.000	1.000
Nonlinear	Weighting	Sig.O.NS	0.363	-50.3%	0.201	-15.0%	0.646	0.523
Nonlinear	Weighting	Sig.O.S8	0.508	-30.3%	0.078	-0.4%	0.841	0.801
Nonlinear	Weighting	Sig.O.S4	0.697	-4.6%	0.018	9.9%	0.971	0.958
Nonlinear	Weighting	Uncorrected	2.175	197.9%	2.092	-3.1%	0.000	1.000

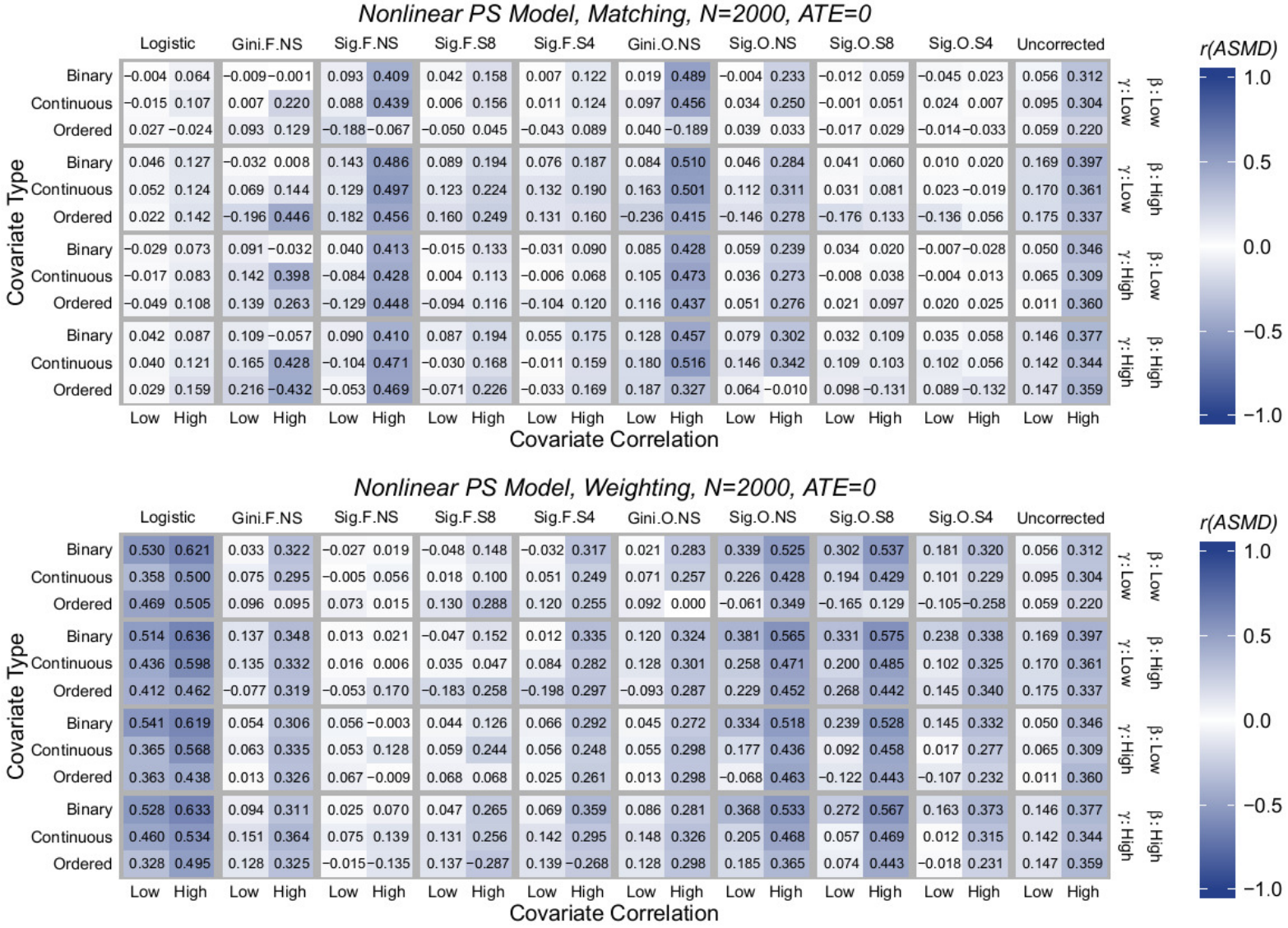
APPENDIX B

CORRELATION BETWEEN ABSOLUTE STANDARDIZED MEAN DIFFERENCE (ASMD) AND MAGNITUDE OF BIAS OF ATE ESTIMATE ACROSS REPLICATIONS (ATE = 0)


Linear PS Model, Matching, N=600, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Covariate Type
Binary 0.035 0.030 0.077 0.248 0.060 0.257 0.041 0.092 0.000 0.051 0.036 0.119 −0.016 0.128 0.004 −0.026 0.018 −0.066 0.078 0.347
Continuous 0.035 0.057 0.055 0.281 0.063 0.287 0.032 0.112 0.014 0.075 −0.012 0.114 −0.012 0.145 0.006 0.034 −0.029 −0.003 0.079 0.316
Ordered 0.083 0.095 0.060 0.281 0.063 0.327 0.045 0.151 −0.029 0.114 0.005 0.149 0.008 0.131 0.028 0.075 0.064 0.042 0.032 0.353
Binary 0.029 0.129 0.117 0.312 0.106 0.305 0.062 0.160 0.058 0.143 0.029 0.165 0.056 0.194 0.025 0.034 −0.021 0.007 0.141 0.425
Continuous 0.019 0.088 0.139 0.334 0.088 0.366 0.117 0.202 0.107 0.170 0.049 0.213 0.070 0.209 0.038 0.081 −0.003 0.040 0.156 0.380
Ordered 0.050 0.061 0.089 0.359 0.050 0.329 0.052 0.198 0.000 0.180 0.075 0.232 0.097 0.191 0.108 0.100 0.073 0.049 0.145 0.345
Binary 0.003 0.037 0.025 0.218 0.051 0.216 −0.021 0.154 −0.016 0.146 0.023 0.121 0.018 0.136 0.014 0.013 −0.031 0.008 0.045 0.358
Continuous 0.010 0.026 0.010 0.255 −0.016 0.293 −0.012 0.088 0.002 0.062 −0.008 0.129 0.004 0.154 −0.018 0.020 −0.034 −0.015 0.054 0.287
Ordered −0.040 0.076 −0.021 0.248 −0.040 0.276 −0.077 0.126 −0.018 0.031 0.077 0.145 0.108 0.173 0.005 0.018 0.073 0.000 0.118 0.296
Binary 0.039 0.125 0.071 0.291 0.081 0.274 0.079 0.145 0.029 0.150 0.047 0.175 0.055 0.166 0.026 0.038 −0.021 0.044 0.147 0.393
Continuous −0.001 0.067 0.024 0.327 −0.046 0.375 −0.048 0.177 −0.036 0.135 0.095 0.181 0.110 0.236 0.098 0.093 0.073 0.080 0.146 0.357
Ordered 0.010 0.049 0.030 0.302 −0.061 0.384 0.025 0.165 −0.008 0.143 0.101 0.166 0.092 0.188 0.043 0.113 0.081 0.059 0.118 0.324
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Linear PS Model, Weighting, N=600, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Covariate Type
Binary 0.367 0.504 0.042 0.359 −0.013 0.042 −0.004 0.261 0.021 0.318 0.035 0.351 0.212 0.377 0.049 0.139 −0.056 0.161 0.078 0.347
Continuous 0.264 0.432 0.051 0.302 −0.012 0.012 −0.005 0.159 0.013 0.232 0.050 0.304 0.174 0.353 0.033 0.159 −0.061 0.044 0.079 0.316
Ordered 0.283 0.455 0.036 0.328 −0.012 0.096 0.025 0.247 0.032 0.296 0.044 0.317 0.141 0.314 0.035 0.092 −0.029 0.164 0.032 0.353
Binary 0.377 0.516 0.097 0.392 0.025 0.055 0.027 0.291 0.053 0.344 0.081 0.378 0.193 0.368 0.102 0.136 −0.011 0.180 0.141 0.425
Continuous 0.277 0.446 0.123 0.378 0.033 0.053 0.031 0.209 0.060 0.311 0.097 0.386 0.208 0.366 0.080 0.136 −0.017 0.080 0.156 0.380
Ordered 0.293 0.454 0.136 0.317 0.082 0.085 0.086 0.226 0.108 0.302 0.113 0.314 0.174 0.373 0.125 0.162 0.044 0.121 0.145 0.345
Binary 0.386 0.482 0.028 0.391 0.009 0.041 0.004 0.326 0.022 0.335 0.017 0.375 0.234 0.369 0.100 0.156 −0.029 0.257 0.045 0.358
Continuous 0.280 0.385 0.069 0.329 0.003 0.160 0.037 0.261 0.039 0.245 0.072 0.329 0.089 0.300 0.024 0.164 0.015 0.238 0.054 0.287
Ordered 0.302 0.437 0.124 0.292 0.094 0.282 0.100 0.294 0.107 0.276 0.127 0.307 0.052 0.294 0.002 0.216 0.100 0.274 0.118 0.296
Binary 0.392 0.518 0.102 0.408 0.043 0.083 0.042 0.325 0.062 0.340 0.086 0.383 0.233 0.390 0.082 0.192 −0.001 0.257 0.147 0.393
Continuous 0.301 0.438 0.157 0.382 0.100 0.221 0.118 0.322 0.129 0.323 0.153 0.391 0.099 0.376 0.049 0.210 0.092 0.289 0.146 0.357
Ordered 0.264 0.396 0.071 0.306 0.036 0.233 0.083 0.316 0.098 0.303 0.070 0.320 0.151 0.296 0.070 0.156 0.068 0.299 0.118 0.324
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

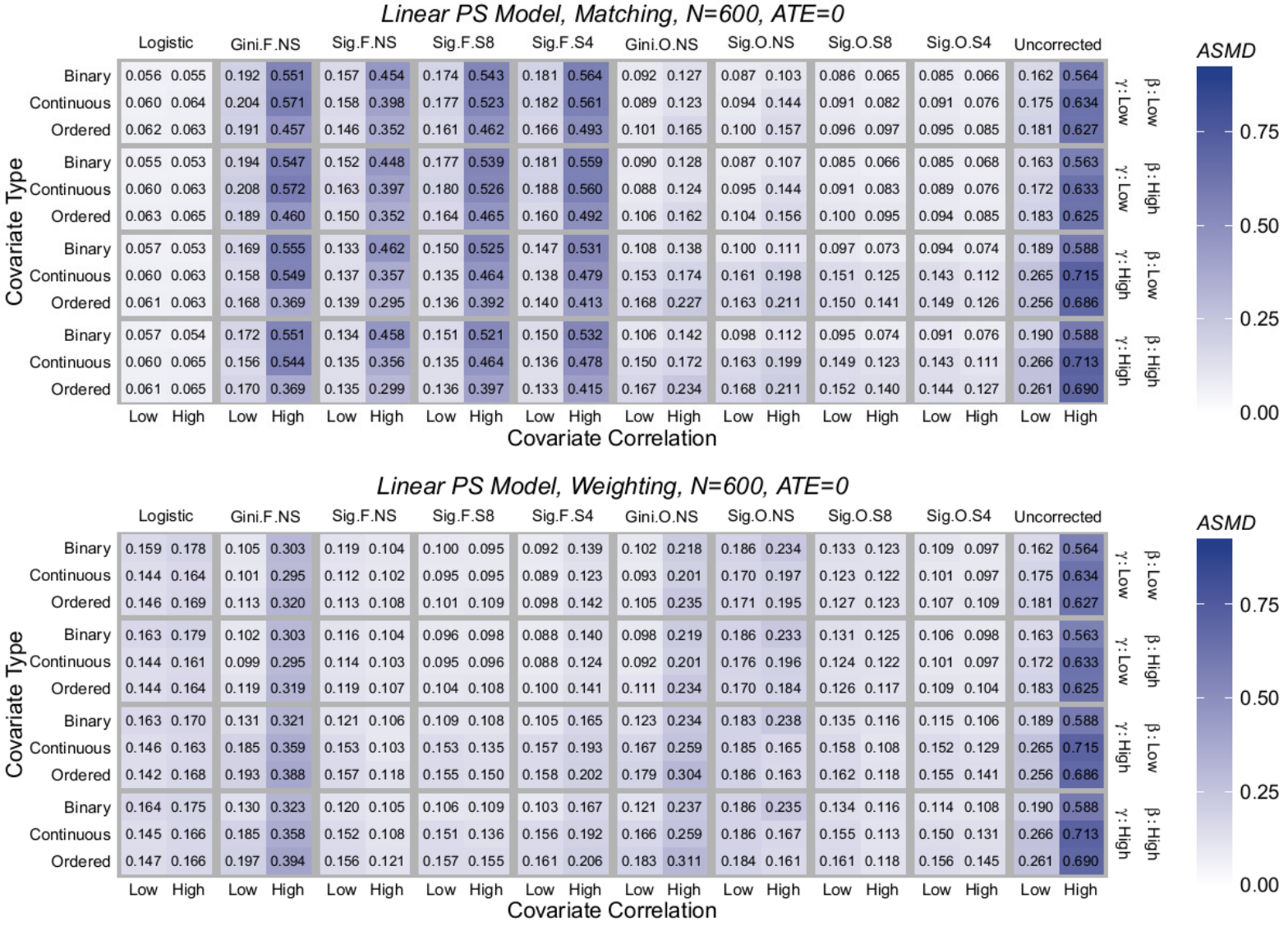

Nonlinear PS Model, Matching, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Nonlinear PS Model, Weighting, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

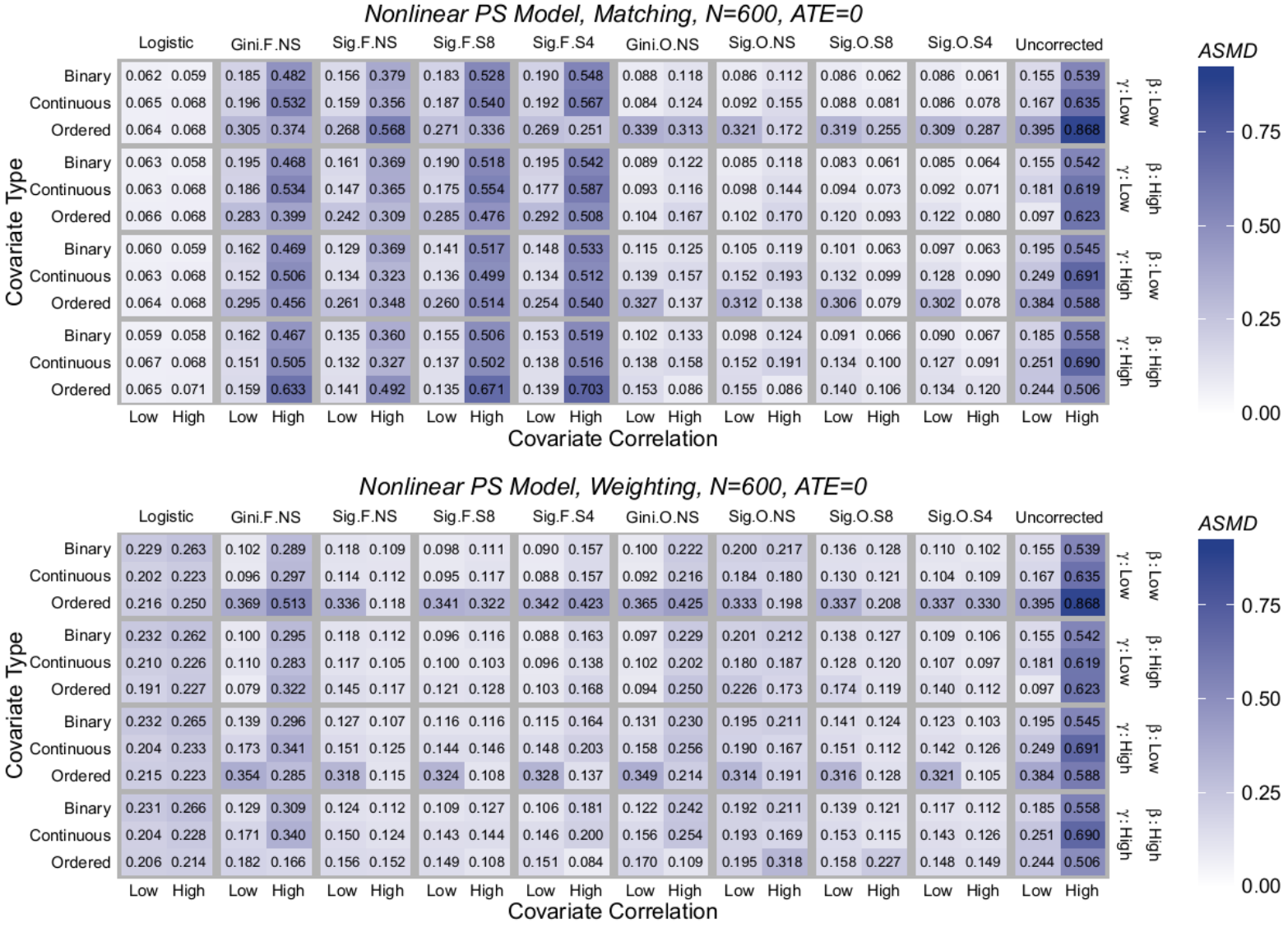

Linear PS Model, Matching, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.012 0.057 0.017 0.130 0.066 0.263 0.037 0.119 0.051 0.087 0.042 0.287 0.023 0.159 0.018 0.027 0.017 0.012 0.122 0.324
Continuous 0.009 0.058 0.027 0.167 0.067 0.308 0.036 0.122 0.028 0.112 0.051 0.297 0.010 0.135 0.016 0.030 0.025 0.045 0.082 0.289
Ordered 0.003 0.083 0.045 0.178 0.026 0.299 0.050 0.160 0.005 0.181 0.066 0.239 0.040 0.159 0.030 0.111 0.012 0.042 0.104 0.330
Binary 0.007 0.094 0.044 0.157 0.099 0.348 0.099 0.202 0.053 0.169 0.075 0.359 0.024 0.231 −0.004 0.073 0.003 −0.046 0.125 0.350
Continuous 0.034 0.123 0.041 0.217 0.116 0.378 0.120 0.214 0.118 0.182 0.104 0.334 0.070 0.239 0.047 0.132 0.021 0.045 0.160 0.349
Ordered 0.031 0.117 0.009 0.153 0.142 0.262 0.116 0.172 0.096 0.209 0.055 0.286 0.060 0.204 0.018 0.128 0.022 0.108 0.126 0.338
Binary 0.007 0.035 0.004 0.036 0.045 0.085 0.008 0.101 0.033 0.121 0.065 0.185 0.054 0.109 0.056 0.018 0.041 −0.042 0.100 0.304
Continuous −0.002 0.010 −0.020 0.195 −0.030 0.267 −0.007 0.127 −0.013 0.081 0.034 0.289 −0.018 0.174 −0.028 0.030 −0.015 −0.010 0.035 0.283
Ordered −0.015 0.033 −0.028 0.012 −0.019 0.289 −0.046 0.143 −0.024 0.131 0.056 0.268 0.001 0.143 −0.013 0.023 0.004 0.091 0.040 0.269
Binary −0.005 0.072 0.018 0.056 0.121 0.190 0.069 0.121 0.075 0.145 0.090 0.292 0.061 0.184 0.052 0.033 0.049 0.034 0.097 0.352
Continuous 0.018 0.082 −0.038 0.253 −0.091 0.343 −0.063 0.164 −0.031 0.170 0.117 0.343 0.088 0.219 0.057 0.094 0.074 0.075 0.141 0.313
Ordered 0.021 0.083 0.044 0.009 −0.092 0.359 −0.061 0.177 −0.030 0.133 0.105 0.254 0.107 0.198 0.085 0.043 0.071 0.049 0.132 0.308
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Type
Covariate Correlation
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Linear PS Model, Weighting, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.352 0.530 0.078 0.291 0.049 0.179 −0.006 −0.092 −0.004 0.290 0.067 0.221 0.268 0.443 0.252 0.507 0.166 0.241 0.122 0.324
Continuous 0.275 0.457 0.059 0.275 0.016 0.155 0.011 −0.054 0.039 0.179 0.047 0.234 0.242 0.345 0.212 0.390 0.122 0.211 0.082 0.289
Ordered 0.235 0.515 0.080 0.317 0.031 0.173 0.000 0.125 0.025 0.283 0.066 0.273 0.263 0.391 0.251 0.486 0.103 0.313 0.104 0.330
Binary 0.400 0.551 0.131 0.312 0.036 0.161 0.021 −0.070 0.051 0.283 0.126 0.265 0.260 0.474 0.220 0.520 0.145 0.222 0.125 0.350
Continuous 0.282 0.497 0.133 0.332 0.042 0.176 −0.039 −0.074 0.025 0.242 0.117 0.294 0.237 0.407 0.205 0.436 0.088 0.198 0.160 0.349
Ordered 0.205 0.508 0.073 0.288 0.016 0.171 0.015 0.091 0.023 0.279 0.060 0.246 0.193 0.413 0.157 0.454 0.106 0.244 0.126 0.338
Binary 0.413 0.506 0.067 0.241 0.030 0.137 0.039 0.057 0.054 0.298 0.054 0.164 0.332 0.412 0.282 0.516 0.151 0.266 0.100 0.304
Continuous 0.251 0.468 0.045 0.329 −0.021 0.134 0.030 0.229 0.034 0.262 0.038 0.275 0.148 0.405 0.054 0.438 0.023 0.261 0.035 0.283
Ordered 0.254 0.529 0.053 0.259 −0.007 0.045 0.054 0.260 0.044 0.269 0.035 0.223 0.181 0.391 0.045 0.395 0.062 0.295 0.040 0.269
Binary 0.391 0.538 0.110 0.283 0.012 0.175 0.060 0.033 0.075 0.344 0.102 0.226 0.287 0.464 0.241 0.549 0.163 0.262 0.097 0.352
Continuous 0.357 0.499 0.141 0.364 −0.061 0.182 0.124 0.256 0.128 0.311 0.129 0.314 0.149 0.431 0.005 0.465 0.004 0.290 0.141 0.313
Ordered 0.286 0.515 0.092 0.306 −0.060 0.151 0.110 0.246 0.124 0.310 0.106 0.281 0.132 0.484 0.021 0.477 0.027 0.259 0.132 0.308
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Type
Covariate Correlation
γ : Low β : Low
γ : Low β : High
γ : High β : Low
γ : High β : High
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

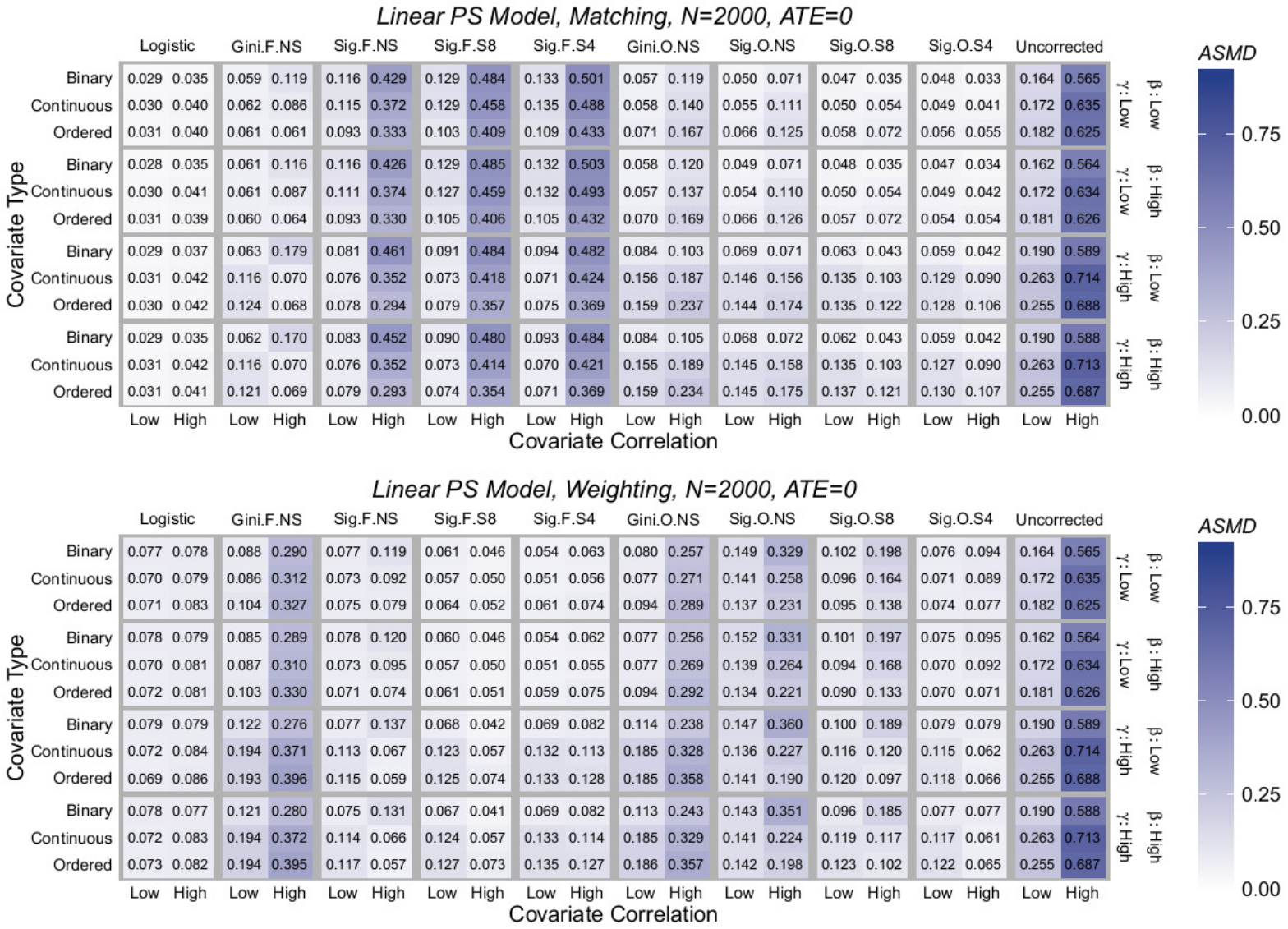

Nonlinear PS Model, Matching, N=2000, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0
Nonlinear PS Model, Weighting, N=2000, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low High
Covariate Correlation
r(ASMD)
1.0
0.5
0.0
−0.5
−1.0

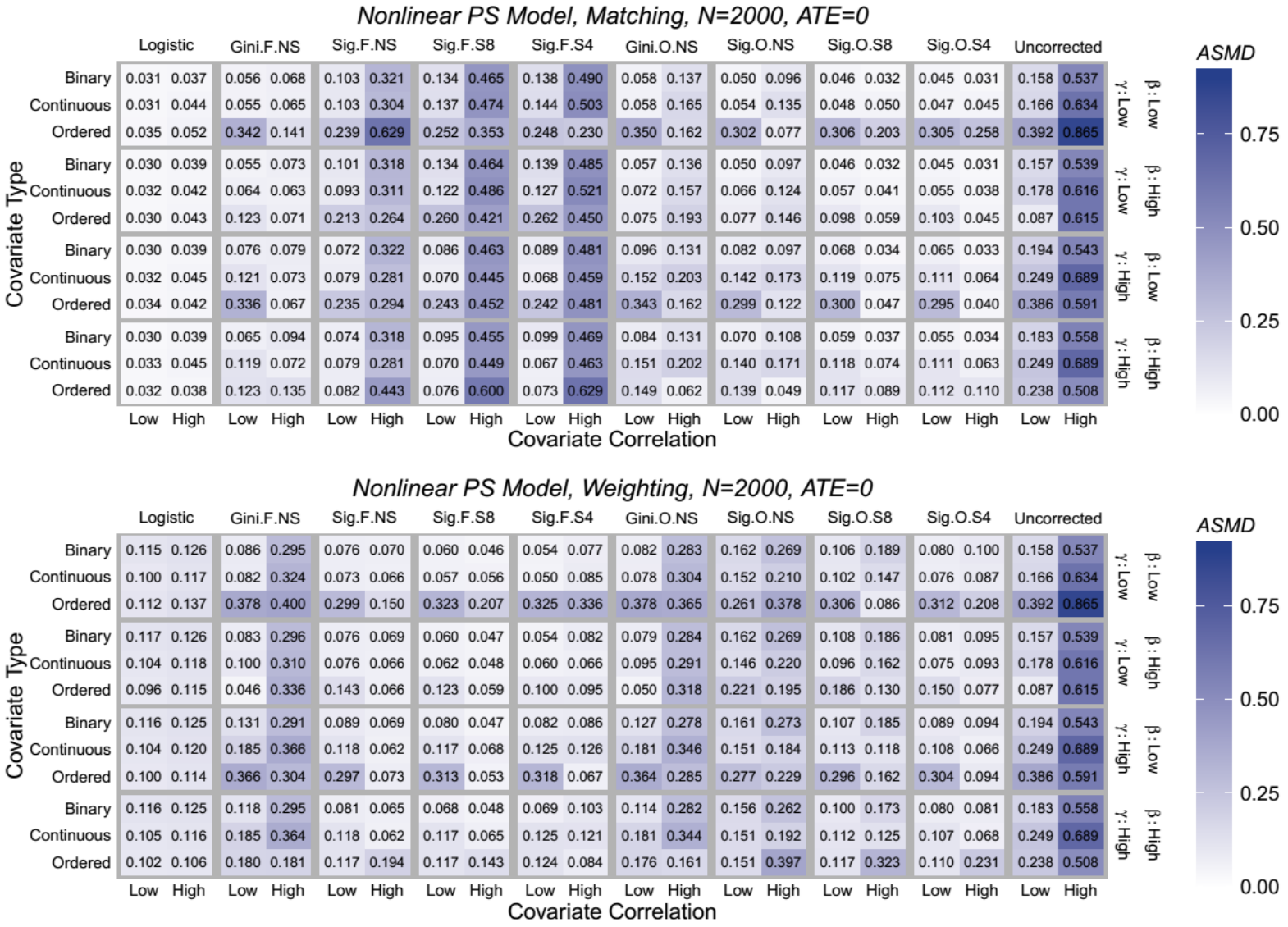
APPENDIX C

MEAN OF ABSOLUTE STANDARDIZED MEAN DIFFERENCE (ASMD) ACROSS REPLICATIONS (ATE = 0)


Linear PS Model, Matching, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary 0.056 0.055 0.192 0.551 0.157 0.454 0.174 0.543 0.181 0.564 0.092 0.127 0.087 0.103 0.086 0.065 0.085 0.066 0.162 0.564
Continuous 0.060 0.064 0.204 0.571 0.158 0.398 0.177 0.523 0.182 0.561 0.089 0.123 0.094 0.144 0.091 0.082 0.091 0.076 0.175 0.634
Ordered 0.062 0.063 0.191 0.457 0.146 0.352 0.161 0.462 0.166 0.493 0.101 0.165 0.100 0.157 0.096 0.097 0.095 0.085 0.181 0.627
Binary 0.055 0.053 0.194 0.547 0.152 0.448 0.177 0.539 0.181 0.559 0.090 0.128 0.087 0.107 0.085 0.066 0.085 0.068 0.163 0.563
Continuous 0.060 0.063 0.208 0.572 0.163 0.397 0.180 0.526 0.188 0.560 0.088 0.124 0.095 0.144 0.091 0.083 0.089 0.076 0.172 0.633
Ordered 0.063 0.065 0.189 0.460 0.150 0.352 0.164 0.465 0.160 0.492 0.106 0.162 0.104 0.156 0.100 0.095 0.094 0.085 0.183 0.625
Binary 0.057 0.053 0.169 0.555 0.133 0.462 0.150 0.525 0.147 0.531 0.108 0.138 0.100 0.111 0.097 0.073 0.094 0.074 0.189 0.588
Continuous 0.060 0.063 0.158 0.549 0.137 0.357 0.135 0.464 0.138 0.479 0.153 0.174 0.161 0.198 0.151 0.125 0.143 0.112 0.265 0.715
Ordered 0.061 0.063 0.168 0.369 0.139 0.295 0.136 0.392 0.140 0.413 0.168 0.227 0.163 0.211 0.150 0.141 0.149 0.126 0.256 0.686
Binary 0.057 0.054 0.172 0.551 0.134 0.458 0.151 0.521 0.150 0.532 0.106 0.142 0.098 0.112 0.095 0.074 0.091 0.076 0.190 0.588
Continuous 0.060 0.065 0.156 0.544 0.135 0.356 0.135 0.464 0.136 0.478 0.150 0.172 0.163 0.199 0.149 0.123 0.143 0.111 0.266 0.713
Ordered 0.061 0.065 0.170 0.369 0.135 0.299 0.136 0.397 0.133 0.415 0.167 0.234 0.168 0.211 0.152 0.140 0.144 0.127 0.261 0.690
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
ASMD
0.75
0.50
0.25
0.00
Linear PS Model, Weighting, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary 0.159 0.178 0.105 0.303 0.119 0.104 0.100 0.095 0.092 0.139 0.102 0.218 0.186 0.234 0.133 0.123 0.109 0.097 0.162 0.564
Continuous 0.144 0.164 0.101 0.295 0.112 0.102 0.095 0.095 0.089 0.123 0.093 0.201 0.170 0.197 0.123 0.122 0.101 0.097 0.175 0.634
Ordered 0.146 0.169 0.113 0.320 0.113 0.108 0.101 0.109 0.098 0.142 0.105 0.235 0.171 0.195 0.127 0.123 0.107 0.109 0.181 0.627
Binary 0.163 0.179 0.102 0.303 0.116 0.104 0.096 0.098 0.088 0.140 0.098 0.219 0.186 0.233 0.131 0.125 0.106 0.098 0.163 0.563
Continuous 0.144 0.161 0.099 0.295 0.114 0.103 0.095 0.096 0.088 0.124 0.092 0.201 0.176 0.196 0.124 0.122 0.101 0.097 0.172 0.633
Ordered 0.144 0.164 0.119 0.319 0.119 0.107 0.104 0.108 0.100 0.141 0.111 0.234 0.170 0.184 0.126 0.117 0.109 0.104 0.183 0.625
Binary 0.163 0.170 0.131 0.321 0.121 0.106 0.109 0.108 0.105 0.165 0.123 0.234 0.183 0.238 0.135 0.116 0.115 0.106 0.189 0.588
Continuous 0.146 0.163 0.185 0.359 0.153 0.103 0.153 0.135 0.157 0.193 0.167 0.259 0.185 0.165 0.158 0.108 0.152 0.129 0.265 0.715
Ordered 0.142 0.168 0.193 0.388 0.157 0.118 0.155 0.150 0.158 0.202 0.179 0.304 0.186 0.163 0.162 0.118 0.155 0.141 0.256 0.686
Binary 0.164 0.175 0.130 0.323 0.120 0.105 0.106 0.109 0.103 0.167 0.121 0.237 0.186 0.235 0.134 0.116 0.114 0.108 0.190 0.588
Continuous 0.145 0.166 0.185 0.358 0.152 0.108 0.151 0.136 0.156 0.192 0.166 0.259 0.186 0.167 0.155 0.113 0.150 0.131 0.266 0.713
Ordered 0.147 0.166 0.197 0.394 0.156 0.121 0.157 0.155 0.161 0.206 0.183 0.311 0.184 0.161 0.161 0.118 0.156 0.145 0.261 0.690
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
ASMD
0.75
0.50
0.25
0.00


Nonlinear PS Model, Matching, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Nonlinear PS Model, Weighting, N=600, ATE=0
Logistic
Gini.F.NS
Sig.F.NS
Sig.F.S8
Sig.F.S4
Gini.O.NS
Sig.O.NS
Sig.O.S8
Sig.O.S4
Uncorrected
Covariate Type
Binary
Continuous
Ordered
β : Low γ : Low
β : High γ : Low
β : Low γ : High
β : High γ : High
Low
High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00


Linear PS Model, Matching, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.029 0.035 0.059 0.119 0.116 0.429 0.129 0.484 0.133 0.501 0.057 0.119 0.050 0.071 0.047 0.035 0.048 0.033 0.164 0.565
Continuous 0.030 0.040 0.062 0.086 0.115 0.372 0.129 0.458 0.135 0.488 0.058 0.140 0.055 0.111 0.050 0.054 0.049 0.041 0.172 0.635
Ordered 0.031 0.040 0.061 0.061 0.093 0.333 0.103 0.409 0.109 0.433 0.071 0.167 0.066 0.125 0.058 0.072 0.056 0.055 0.182 0.625
Binary 0.028 0.035 0.061 0.116 0.116 0.426 0.129 0.485 0.132 0.503 0.058 0.120 0.049 0.071 0.048 0.035 0.047 0.034 0.162 0.564
Continuous 0.030 0.041 0.061 0.087 0.111 0.374 0.127 0.459 0.132 0.493 0.057 0.137 0.054 0.110 0.050 0.054 0.049 0.042 0.172 0.634
Ordered 0.031 0.039 0.060 0.064 0.093 0.330 0.105 0.406 0.105 0.432 0.070 0.169 0.066 0.126 0.057 0.072 0.054 0.054 0.181 0.626
Binary 0.029 0.037 0.063 0.179 0.081 0.461 0.091 0.484 0.094 0.482 0.084 0.103 0.069 0.071 0.063 0.043 0.059 0.042 0.190 0.589
Continuous 0.031 0.042 0.116 0.070 0.076 0.352 0.073 0.418 0.071 0.424 0.156 0.187 0.146 0.156 0.135 0.103 0.129 0.090 0.263 0.714
Ordered 0.030 0.042 0.124 0.068 0.078 0.294 0.079 0.357 0.075 0.369 0.159 0.237 0.144 0.174 0.135 0.122 0.128 0.106 0.255 0.688
Binary 0.029 0.035 0.062 0.170 0.083 0.452 0.090 0.480 0.093 0.484 0.084 0.105 0.068 0.072 0.062 0.043 0.059 0.042 0.190 0.588
Continuous 0.031 0.042 0.116 0.070 0.076 0.352 0.073 0.414 0.070 0.421 0.155 0.189 0.145 0.158 0.135 0.103 0.127 0.090 0.263 0.713
Ordered 0.031 0.041 0.121 0.069 0.079 0.293 0.074 0.354 0.071 0.369 0.159 0.234 0.145 0.175 0.137 0.121 0.130 0.107 0.255 0.687
Covariate Type
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00
Linear PS Model, Weighting, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.077 0.078 0.088 0.290 0.077 0.119 0.061 0.046 0.054 0.063 0.080 0.257 0.149 0.329 0.102 0.198 0.076 0.094 0.164 0.565
Continuous 0.070 0.079 0.086 0.312 0.073 0.092 0.057 0.050 0.051 0.056 0.077 0.271 0.141 0.258 0.096 0.164 0.071 0.089 0.172 0.635
Ordered 0.071 0.083 0.104 0.327 0.075 0.079 0.064 0.052 0.061 0.074 0.094 0.289 0.137 0.231 0.095 0.138 0.074 0.077 0.182 0.625
Binary 0.078 0.079 0.085 0.289 0.078 0.120 0.060 0.046 0.054 0.062 0.077 0.256 0.152 0.331 0.101 0.197 0.075 0.095 0.162 0.564
Continuous 0.070 0.081 0.087 0.310 0.073 0.095 0.057 0.050 0.051 0.055 0.077 0.269 0.139 0.264 0.094 0.168 0.070 0.092 0.172 0.634
Ordered 0.072 0.081 0.103 0.330 0.071 0.074 0.061 0.051 0.059 0.075 0.094 0.292 0.134 0.221 0.090 0.133 0.070 0.071 0.181 0.626
Binary 0.079 0.079 0.122 0.276 0.077 0.137 0.068 0.042 0.069 0.082 0.114 0.238 0.147 0.360 0.100 0.189 0.079 0.079 0.190 0.589
Continuous 0.072 0.084 0.194 0.371 0.113 0.067 0.123 0.057 0.132 0.113 0.185 0.328 0.136 0.227 0.116 0.120 0.115 0.062 0.263 0.714
Ordered 0.069 0.086 0.193 0.396 0.115 0.059 0.125 0.074 0.133 0.128 0.185 0.358 0.141 0.190 0.120 0.097 0.118 0.066 0.255 0.688
Binary 0.078 0.077 0.121 0.280 0.075 0.131 0.067 0.041 0.069 0.082 0.113 0.243 0.143 0.351 0.096 0.185 0.077 0.077 0.190 0.588
Continuous 0.072 0.083 0.194 0.372 0.114 0.066 0.124 0.057 0.133 0.114 0.185 0.329 0.141 0.224 0.119 0.117 0.117 0.061 0.263 0.713
Ordered 0.073 0.082 0.194 0.395 0.117 0.057 0.127 0.073 0.135 0.127 0.186 0.357 0.142 0.198 0.123 0.102 0.122 0.065 0.255 0.687
Covariate Type
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
ASMD
0.75
0.50
0.25
0.00


Nonlinear PS Model, Matching, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.031 0.037 0.056 0.068 0.103 0.321 0.134 0.465 0.138 0.490 0.058 0.137 0.050 0.096 0.046 0.032 0.045 0.031 0.158 0.537
Continuous 0.031 0.044 0.055 0.065 0.103 0.304 0.137 0.474 0.144 0.503 0.058 0.165 0.054 0.135 0.048 0.050 0.047 0.045 0.166 0.634
Ordered 0.035 0.052 0.342 0.141 0.239 0.629 0.252 0.353 0.248 0.230 0.350 0.162 0.302 0.077 0.306 0.203 0.305 0.258 0.392 0.865
Binary 0.030 0.039 0.055 0.073 0.101 0.318 0.134 0.464 0.139 0.485 0.057 0.136 0.050 0.097 0.046 0.032 0.045 0.031 0.157 0.539
Continuous 0.032 0.042 0.064 0.063 0.093 0.311 0.122 0.486 0.127 0.521 0.072 0.157 0.066 0.124 0.057 0.041 0.055 0.038 0.178 0.616
Ordered 0.030 0.043 0.123 0.071 0.213 0.264 0.260 0.421 0.262 0.450 0.075 0.193 0.077 0.146 0.098 0.059 0.103 0.045 0.087 0.615
Binary 0.030 0.039 0.076 0.079 0.072 0.322 0.086 0.463 0.089 0.481 0.096 0.131 0.082 0.097 0.068 0.034 0.065 0.033 0.194 0.543
Continuous 0.032 0.045 0.121 0.073 0.079 0.281 0.070 0.445 0.068 0.459 0.152 0.203 0.142 0.173 0.119 0.075 0.111 0.064 0.249 0.689
Ordered 0.034 0.042 0.336 0.067 0.235 0.294 0.243 0.452 0.242 0.481 0.343 0.162 0.299 0.122 0.300 0.047 0.295 0.040 0.386 0.591
Binary 0.030 0.039 0.065 0.094 0.074 0.318 0.095 0.455 0.099 0.469 0.084 0.131 0.070 0.108 0.059 0.037 0.055 0.034 0.183 0.558
Continuous 0.033 0.045 0.119 0.072 0.079 0.281 0.070 0.449 0.067 0.463 0.151 0.202 0.140 0.171 0.118 0.074 0.111 0.063 0.249 0.689
Ordered 0.032 0.038 0.123 0.135 0.082 0.443 0.076 0.600 0.073 0.629 0.149 0.062 0.139 0.049 0.117 0.089 0.112 0.110 0.238 0.508
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
Covariate Type
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
ASMD
0.75
0.50
0.25
0.00
Nonlinear PS Model, Weighting, N=2000, ATE=0
Logistic Gini.F.NS Sig.F.NS Sig.F.S8 Sig.F.S4 Gini.O.NS Sig.O.NS Sig.O.S8 Sig.O.S4 Uncorrected
Binary 0.115 0.126 0.086 0.295 0.076 0.070 0.060 0.046 0.054 0.077 0.082 0.283 0.162 0.269 0.106 0.189 0.080 0.100 0.158 0.537
Continuous 0.100 0.117 0.082 0.324 0.073 0.066 0.057 0.056 0.050 0.085 0.078 0.304 0.152 0.210 0.102 0.147 0.076 0.087 0.166 0.634
Ordered 0.112 0.137 0.378 0.400 0.299 0.150 0.323 0.207 0.325 0.336 0.378 0.365 0.261 0.378 0.306 0.086 0.312 0.208 0.392 0.865
Binary 0.117 0.126 0.083 0.296 0.076 0.069 0.060 0.047 0.054 0.082 0.079 0.284 0.162 0.269 0.108 0.186 0.081 0.095 0.157 0.539
Continuous 0.104 0.118 0.100 0.310 0.076 0.066 0.062 0.048 0.060 0.066 0.095 0.291 0.146 0.220 0.096 0.162 0.075 0.093 0.178 0.616
Ordered 0.096 0.115 0.046 0.336 0.143 0.066 0.123 0.059 0.100 0.095 0.050 0.318 0.221 0.195 0.186 0.130 0.150 0.077 0.087 0.615
Binary 0.116 0.125 0.131 0.291 0.089 0.069 0.080 0.047 0.082 0.086 0.127 0.278 0.161 0.273 0.107 0.185 0.089 0.094 0.194 0.543
Continuous 0.104 0.120 0.185 0.366 0.118 0.062 0.117 0.068 0.125 0.126 0.181 0.346 0.151 0.184 0.113 0.118 0.108 0.066 0.249 0.689
Ordered 0.100 0.114 0.366 0.304 0.297 0.073 0.313 0.053 0.318 0.067 0.364 0.285 0.277 0.229 0.296 0.162 0.304 0.094 0.386 0.591
Binary 0.116 0.125 0.118 0.295 0.081 0.065 0.068 0.048 0.069 0.103 0.114 0.282 0.156 0.262 0.100 0.173 0.080 0.081 0.183 0.558
Continuous 0.105 0.116 0.185 0.364 0.118 0.062 0.117 0.065 0.125 0.121 0.181 0.344 0.151 0.192 0.112 0.125 0.107 0.068 0.249 0.689
Ordered 0.102 0.106 0.180 0.181 0.117 0.194 0.117 0.143 0.124 0.084 0.176 0.161 0.151 0.397 0.117 0.323 0.110 0.231 0.238 0.508
Low High Low High Low High Low High Low High Low High Low High Low High Low High Low High
Covariate Correlation
Covariate Type
γ: Low β: Low
γ: Low β: High
γ: High β: Low
γ: High β: High
ASMD
0.75
0.50
0.25
0.00