International Journal of Turkish Educational Studies,
2023, 11 (21), 989-1031
https://doi.org/10.46778/goputeb.1312865

ISSN: 2148-2314

Comparison of Propensity Score Weighting Methods to Remove Selection Bias in Average Treatment Effect Estimates *

Sungur Gürel[a] Walter Lana Leite [b]

[a] *Assist. Prof. Dr., Siirt University, Siirt, Türkiye, s.gurel@siirt.edu.tr*

[b] *Prof., University of Florida, Florida, United States, walter.leite@coe.ufl.edu*

ABSTRACT

In this Monte Carlo simulation study, the performance of six different propensity score methods implemented through weighting cases was investigated: inverse probability of treatment weighting, truncated inverse probability of treatment weighting, propensity score stratification, marginal mean weighting through propensity score stratification, optimal full propensity score matching, and marginal mean weighting through optimal full propensity score matching. These methods aim to reduce selection bias in estimates of the average treatment effect (ATE) in observational studies. For the estimation of standard errors of the ATE with weights, three methods were compared: weighted least squares (WLS), Taylor series linearization (TSL), and jackknife (JK). Results indicated that covariance adjustment extensions of the investigated propensity score methods, in combination with TSL and JK standard error estimation methods, remove the selection bias appropriately and provide the most accurate standard errors under the simulated conditions.

Article Type
Research

Article Background
Received:
13.06.2023
Accepted:
22.09.2023

Keywords
Observational Study,
Propensity Score Methods,
Average Treatment Effect,
Standard Error

To cite this article: Gürel, S. & Leite, W. L. (2023). Comparison of propensity score weighting methods to remove selection bias in average treatment effect estimates. *International Journal of Turkish Educational Sciences, 11* (21), 989-1031.

Corresponding Author: Sungur Gürel, e-mail: s.gurel@siirt.edu.tr

* This study is produced from an unpublished Master's Thesis titled "The performance of propensity score methods to estimate the average treatment effect in observational studies with selection bias: A Monte Carlo simulation study" defended in 2012 at University of Florida. This study is also partially presented at the Annual Meeting of the American Educational Research Association in 2012 with the title of "Comparison of Inverse Probability of Treatment Weighting and Optimal Full Matching Methods to Estimate the Average Treatment Effect: A Monte Carlo Simulation Study"

Introduction

Estimating the effects of educational interventions using secondary data has become common in educational research because of the availability of various nationally representative databases collected by agencies such as the National Center for Education Statistics (NCES) and the National Science Foundation (NSF) (Strayhorn, 2009). However, because the assignment of participants to interventions in these national studies is not random, estimates of the effects of interventions are vulnerable to selection bias due to both observed and unobserved covariates (Shadish et al., 2002). In the last four decades, several methods emerged for estimating treatment effects and dealing with selection bias in studies that lack random assignment to treatment conditions (Heckman, 1978; Rosenbaum & Rubin, 1983; Abadie & Imbens, 2006; Heckman et al., 1997), which are referred collectively as observational studies. Propensity score methods are among the most commonly used methods in social science research to analyze observational studies (Thoemmes & Kim, 2011). In order to deal with selection bias in treatment effect estimates, propensity score methods attempt to balance pre-existing differences between treated and untreated participants on observed covariates. Rosenbaum and Rubin (1983) used the term propensity score (PS) for the first time and defined it as the predicted probability of treatment assignment given observed covariates. They proved that if selection into treatment depends on observed covariates, the observed difference in treatment and control at a propensity score level is an unbiased estimate of the average treatment effect (ATE) at that level (Rosenbaum & Rubin, 1983). Propensity scores can be utilized to reduce selection bias in the ATE estimates by matching observations based on their similarity in PS, weighting observations with the inverse of the PS, and stratifying observations into homogenous groups based on PS (Stuart, 2010). Except for one-to-one matching, all PS methods produce observation weights, and therefore, estimation methods for survey data with sampling weights are applicable to the estimation of treatment effects with PS weights (Leite, 2016). Lunceford and Davidian (2004) show estimators of the treatment effect for propensity score weighting and stratification that are equivalent to estimators discussed in the sampling literature (e.g., Lohr, 1999). Also, methods to estimate standard errors, such as bootstrapping, jackknife, and Taylor-Series linearization with complex survey data (Stapleton, 2008), can be applied to data with propensity score weights.

Although several types of treatment effects have been defined in the literature (Guo & Fraser, 2010), the estimates most commonly found in the social sciences literature are the average treatment effect (ATE) and the average treatment effect on the treated (ATT) (Thoemmes & Kim, 2011). The specific implementation of a PS method differs depending on whether the ATE or ATT are of interest. There have been several studies comparing implementations of PS methods to estimate the ATT (Gu & Rosenbaum, 1993; Cepeda et al., 2003; Austin, 2010b; Harder et al., 2010), but there has not been a study comparing major PS methods for the estimation of the ATE. Therefore, the first objective of this study is to compare inverse probability of treatment weighting (IPTW), truncated inverse probability of treatment weighting (TIPTW), propensity score stratification (PSS), marginal mean weighting through stratification (MMWS), optimal full matching (OFM), marginal mean weighting through full matching (MMWFM), and their covariance adjustment extensions for their ability to reduce selection bias in estimates of the ATE. Because most studies comparing PS methods focused on treatment effect estimates and did not address the estimation of standard errors, the second objective of this study is to compare weighted least squares regression (WLS), Taylor series linearization (TSL), and jackknife (JK) for estimating standard errors for the ATE estimates obtained with each PS method.

The Potential Outcomes Framework

Rubin's potential outcomes framework (1974) is commonly used to address the selection bias issue in observational studies. Its basic principle is that treated and control group participants have potential outcomes in both the presence and absence of treatment. For instance, let the observed outcome of a treated participant *i* be Y_{iT}^{t} , while Y_{iT}^{c} is the potential outcome if this participant had been placed in the control group. Similarly, Y_{iC}^{c} is the observed outcome for the control group participant *i*, and Y_{iC}^{t} is the potential outcome if this participant had been placed in the treatment group. In other words, a control group participant has a potential outcome under the treatment condition. Conversely, a treatment group participant has a potential outcome under the control condition. The ATE is $E[Y_i^t] - E[Y_i^c]$, which is the difference between the potential outcome for all individuals if they were exposed to the treatment condition and the potential outcome for all individuals if they were exposed to the untreated condition (Winship & Morgan, 1999). In randomized studies, the ATE is $E[Y_{iT}^{t}]$ - $E[Y_{iC}^{c}]$, but this is not true in observational studies because the assumption that $E[Y_{iT}^{c}] = E[Y_{iC}^{c}]$ and $E[Y_{iC}^{t}] = E[Y_{iT}^{t}]$ cannot be made. Further, while in randomized experimental designs, the ATT, defined as $E[Y_{iT}^{t}] - E[Y_{iT}^{c}]$, is identical to the ATE, this is not necessarily true in observational studies. Therefore, the estimator $\bar{Y}_{iT}^{t} - \bar{Y}_{iC}^{c}$ of the ATE (i.e., the difference between the means of the observed outcomes of the treated and control group individuals) will only be unbiased if the assignment to the treatment is independent of the potential outcomes. More formally, $T \perp \{Y_i^t, Y_i^c\} \mid X$ where Y_i^t is the potential outcome if treated, Y_i^c is the potential outcome if untreated, *X* is the all potential confounders, and *T* is the treatment assignment. This condition is known as the strong ignorability of treatment assignment (SITA) (Rubin, 1974). It is also necessary that the stable unit treatment value assumption (SUTVA) is met, which requires that the potential outcome of one unit is not affected by the particular treatment assignment or potential outcomes of other units (Rubin, 2007). In observational studies, both SITA and SUTVA may be violated, which leads to biased estimates of the ATE and poor internal validity of the study (Shadish, 2002). In observational studies, PS methods are used to attempt to achieve strong ignorability of treatment assignment, under the assumption that SUTVA holds, by balancing the distributions of observed covariates between treatment and control groups. Violations of SUTVA, for instance, whena parent's decision to enroll the student in an educational intervention is affected by the enrollment status of the his/her classmates, require special considerations in the estimation of propensity scores and implementation of the PS methods that are discussed elsewhere (Arpino & Mealli, 2011; Hong & Hong, 2008; Thoemmes & West, 2011; Leite et al., 2015).

Propensity Score Methods for Reducing Selection Bias in ATE

The use of any PS method requires a multiple-step process that starts with selecting observed covariates related to both selection into treatment conditions and the outcome (see Brookhart et al., 2006, for a discussion of variable selection). The second step is to estimate propensity scores, which is most commonly accomplished with logistic regression, but other parametric or non-parametric models (McCaffrey et al., 2004) can be used. The third step is to evaluate the common support area of the estimated propensity scores, which is the area of the propensity score distribution where values exist for both treatment and control groups (Guo & Fraser, 2010). Lack of common support for a particular area of the propensity score distribution restricts the generalizability of the estimates

only to the sub-population for which common support exists. The fourth step is to verify the balance of the distribution of covariates given the propensity score method of choice. The fifth step is to combine the PS method with either parametric or non-parametric estimators of the ATE and its standard error and reach conclusions about the statistical significance of the ATE. The last step is to evaluate the sensitivity of the results to the possible omission of important covariates (Rosenbaum, 2010).

Combining propensity scores with parametric, model-based estimators of the ATE (Ho et al., 2006) has the advantage of reducing bias of the ATE estimates because it allows the researcher to evaluate complex hypotheses about the outcome with linear models, generalized linear models, mixed-effects models, and structural equation models. Propensity score methods can be incorporated into model-based estimation by creating weights, which are used similarly to sampling weights in model estimation from finite survey samples (Leite, 2016). Below, we describe commonly used propensity score methods and the calculation of weights with each method.

Inverse Probability of Treatment Weighting

Inverse probability weighting was introduced around the middle of the 20th century by Horvitz and Thompson (1952) to account for the effect of the sampling design in survey estimates. Robins et al. (2000) extended this concept to IPTW, to control for selection bias in observational studies. The idea behind IPTW is to weight subjects by the inverse of the conditional probability of being in the group that they are actually in. Formally, let T_i be the treatment indicator, with $T_i = 1$ indicating a member of the treatment group and $T_i = 0$ indicating a member of the control group. $\hat{e}_i$ is the estimated propensity score. To estimate the ATE, for individual *i* the weight w_i is (Stuart, 2010):

$$w_i = \frac{T_i}{\hat{e}_i} + \frac{1 - T_i}{1 - \hat{e}_i} \tag{1}$$

IPTW creates a pseudo population where observations are replicated based on the weights so that participants not only account for themselves, but also for those who have similar characteristics in the other group (Hernan et al., 2004). Neugebauer and van der Laan (2005) discovered that the performance of IPTW depends on the experimental treatment assignment assumption, which requires all of the weights to be different from zero. They also found that if any treatment probability is close to zero, the new weighted sample may not represent the target population.

Truncated Inverse Probability of Treatment Weighting

The IPTW method has been criticized for its performance when the weights or propensity scores are extreme (Freedman & Berk, 2008). The extreme weights create overly influential observations and inflate the sampling variability of estimates. Several researchers came up with different solutions to solve this problem. Bembom and van der Laan (2008) developed a data-adaptive selection of truncation level for IPTW estimators. They were able to gain up to *7%* efficiency in the mean square error of estimates. Freedman and Berk (2008) replaced the weights greater than *20* with *20* and trimmed observations greater than *20*. However, they concluded that neither method could reduce the selection bias. Sturmer et al. (2010) found that trimming up to propensity scores that are more extreme than 2.5^{th} and 97.5^{th} percentiles reduces selection bias compared to not trimming any observations and trimming more observations. Crump et al. (2009) proposed estimating the ATE using only the subsample of the data that meets the optimal selection rule to reduce the asymptotic

variance of the treatment effect estimation without introducing additional bias to the ATE estimate. The optimal selection rule simply puts upper and lower bounds to keep observations without extreme propensity scores. Researchers also concluded that [.1, .9] the thumb selection rule works as well as the optimal selection rule when the propensity scores follow various beta distributions. Instead of dealing with possible extreme weights that extreme propensity scores would cause, we suggest a more straightforward approach to deal with unintended consequences of extreme weights. After IPTW weights are calculated, truncated inverse probability of treatment weights are obtained by truncating weights that are greater than the *99th* percentile of the IPTW. Those extreme weights are truncated to the value of the weight that represents the *99th* percentile.

Propensity Score Stratification

PSS consists of creating strata containing similar individuals with respect to propensity scores (Stuart, 2010), where each strata should contain at least one treated and one untreated individual. PSS is usually accomplished by dividing the distribution of propensity scores into intervals of equal size. Stratification based on a single covariate to reduce selection bias was proposed by Cochran (1968). However, Rosenbaum and Rubin (1984) showed that stratification into five strata based on propensity scores reduces about *90%* of the selection bias. In applied social science research, Thoemmes and Kim (2011) found that most studies use between *5* and *10* strata. Obtaining strata to estimate the ATE requires that all members of the sample are placed into a stratum, while stratum containing only untreated observations may be dropped in the estimation of ATT. Furthermore, estimating ATE requires the cases to be weighted by the number of individuals in each stratum based on the following formula:

$$w_i = 0.5\left[T_i \frac{n^s}{n_s^t} + (1-T_i)\frac{n^s}{n_s^c} \right] \qquad (2)$$

where subscript s indexes the strata membership, *t* and *c* index the treatment status. n^s is the sample size of stratum *s*, n_s^t represents the number of treated participants in the stratum *s*, and n_s^c represents the number of untreated participants in stratum s. In contrast, ATT weights are created with respect to the number of treated individuals in each stratum.

Optimal Full Matching

Full matching is a method of stratification in which the number of subclasses is shaped based on the observed data (Leite, 2016). When picking a treated unit and a control unit randomly from the same subclass, the expected difference between those two with respect to a certain measure of distance is Δ. The OFM algorithm was established based on network flow theory to minimize Δ within matched sets by finding a minimum cost flow for the whole sample (Rosenbaum, 1989: 1991). If OFM is most frequently performed with propensity scores, where $\Delta_{ij} = \left|\hat{e}_{it} - \hat{e}_{jc}\right|$ is the distance between the propensity scores of the participant *i* of the treatment group and the participant j of the control group. Rosenbaum (1991) found that there is always a full matching that is optimal so that $\sum_{i,j}^{T,C} \Delta_{ij}$ is minimized where *T* is the number of treated participants, and *C* is the number of control participants in a sample. Estimating the ATE with OFM requires weights calculated like in PSS (see Equation 2).

Marginal Mean Weighting through Stratification

Hong (2012) proposed MMWS to adjust each treated or control unit as if the sample had been randomly assigned to treatment within each stratum. In combination with strong ignorability of the treatment assignment assumption, the weighted data is representative of a pseudo-population where treated and control groups are equivalent with respect to covariates. The following formula represents how weights are calculated:

$$w_i = T_i \Pr(\mathrm{T}=1)\frac{n_s}{n_s^t} + (1-T_i)\Pr(\mathrm{T}=0)\frac{n_s}{n_s^c} \quad (3)$$

where $\Pr(\mathrm{T}=1)$ and $\Pr(\mathrm{T}=0)$ are the proportion treated and untreated in the entire sample, respectively, and the other terms are as defined previously.

Marginal Mean Weighting through Full Matching

We propose MMWFM as a direct extension of MMWS. This method requires weighting observations using Hong's (2012) approach but obtaining strata based on the OFM rather than the PSS method. Once strata are shaped using OFM, weights are calculated using the Equation (3) rather than the Equation (2).

When we compared PSS to MMWS and OFM to MMWFM, we discovered that using the original weights or marginal mean weights produced the same ATE and standard error of the estimated ATE. This is because only a single term differs between the weight in Equation 2 for PSS and Equation 3 for MMWTS-PSS. The *0.5* constant in Equation 2 specifies that half of the sample would receive each condition if there is no selection bias. On the other hand, marginal mean weights in Equation 3 contain $\Pr(\mathrm{T}=1)$ and $\Pr(\mathrm{T}=0)$, which are the marginal proportions of individuals who received treatment or control conditions. We found that both PSS weights and MMWS weights have exactly the same relationship across strata, and the same is true for OFM compared to MMWFM. Therefore, MMWS and MMWFM are excluded from further analyses, and results obtained from PSS and OFM are generalizable to the MMWS and MMWFM, respectively. On the other hand, conditioning on propensity score creates a difference between the PSS and MMWS as well as OFM and MMWFM due to the propensity score differences across strata.

Covariance Adjustment (CA)

The intuitive need for the covariance adjustment (CA) using propensity score comes from the double robustness property. Observed covariates included in the propensity score estimation model can also added into parametric or non-parametric estimators of ATE (Robins & Rotnitzky, 2001; Funk, et al., 2011). Controlling for the same covariates in both treatment assignment and outcome models protects against misspecification of either the propensity score model or the outcome model (but not both simultaneously). CA using the propensity score is commonly used in medical research by including the treatment indicator and the propensity score in the outcome regression model. (Weitzen et al., 2004). CA is recommended in the standards of the What Works Clearinghouse (U.S. Department of Education et al., 2013) for quasi-experimental educational studies in addition to matching cases on values of baseline covariates. CA using the propensity score by itself without implementing any propensity score method is not recommended because it implies that the propensity score is linearly related to the outcome, and the estimated coefficient of the treatment indicator cannot be interpreted as the ATE (Schafer & Kang, 2008). Furthermore, Austin et al. (2007)

concluded that using only CA with the propensity score can result in a biased treatment effect estimation. We utilize CA with the propensity score combined with the PS methods that we investigate throughout the paper. In this combination, CA is expected to improve inferences by removing small residual covariate imbalances between treatment and control groups not removed by the PS method (Stuart, 2010). We add "-CA" to the abbreviation for propensity score method to indicate the covariance adjustment extensions of original propensity score methods as IPTW-CA, TIPTW-CA, PSS-CA, MMWS-CA, OFM-CA, and MMWFM-CA.

As we have discussed earlier, Equations 2 and 3 differ by a single term in the weight equation, and individuals in the same strata and treatment condition will be weighted by the same number, so the relationships between strata remain constant. Hence, the ATE and standard error estimates will remain constant if either Equation 2 or 3 are used to calculate weights. However, PSS-CA and MMWS-CA, or OFM-CA and MMWFM-CA do not behave equivalently because individuals do not necessarily have the same or linearly transformed propensity score in a stratum.

Estimation of Treatment Effects

Once weights are calculated with Equations 1, 2, or 3, the ATE can be estimated as the difference between weighted means (Schafer & Kang, 2008; Lunceford & Davidian, 2004):

$$\Delta = \frac{\sum_{i=1}^{nt} w_{it} y_{it}}{\sum_{i=1}^{n_t} w_{it}} - \frac{\sum_{i=1}^{n_c} w_{jc} y_{jc}}{\sum_{j=1}^{n_c} w_{jc}} \tag{4}$$

where w_{it} , w_{jc} and y_{it} , y_{jc} are weights and outcomes for treated and control group participants, respectively. Alternatively, the treatment effect can be obtained by fitting the weighted regression model $Y_i = \beta_0 + \beta_1 T_i + e_i$, where Y_i is the outcome, β_0 is the intercept, β_1 is the ATE, and e_i is the residual (Leite, 2016).

Standard Error Estimation

Weighted Least Squares Regression

Weighted least squares regression (WLS) can be used to obtain ATE estimates and standard errors with all PS methods presented above (Schafer & Kang, 2008). Standard errors with WLS are estimated with the following formula (Fox, 2008):

$$SE(\hat{B}_1) = \sqrt{\frac{\sum_{i=1}^{n} \left(e_i / w_i\right)^2}{n}} \tag{5}$$

Taylor Series Linearization

Several other methods can be used to obtain standard errors of ATE estimates from PS methods, such as Taylor Series Linearization, Jackknife, and Bootstrapping (Rodgers, 1999). These methods are used extensively in finite sample surveys but have not been researched extensively with PS methods. Taylor series linearization (TSL) can be used to obtain the variance of a statistic via approximating the estimator by a linear function of observations (Wolter, 2007). Formally, let

$U_i(ATE)$ be a function of the data for observation *i* and ATE, and the true population ATE^* solves the following equation:

$$\sum_{i=1}^{N} U_i(ATE^*) = 0 \tag{6}$$

Then, in a complex sample we are able to define $\hat{ATE}$ as solving the weighted sample equation:

$$\sum_{i=1}^{N} \hat{U}_i(ATE) = 0 \tag{7}$$

The variance of the ATE is defined as follows, applying delta method (Binder, 1983):

$$\hat{\text{var}}[\hat{ATE}] \approx \left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE} \right)^{-1} \text{cov}\left[\sum_{i=1}^{n} \hat{U}_i(\hat{ATE}) \right] \left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE} \right)^{-1} \tag{8}$$

For the observation *i*, let $\hat{B}_0$ be the intercept, $\hat{B}_1$ be the slope of the regression equation where $\hat{B}_1$ is estimated ATE, w_i be the weight, and $\hat{\bar{x}}$ be the mean of *x*. Using Taylor series linearization, the standard error of ATE is defined as (Lohr, 1999):

$$SE(\hat{B}_1) = \sqrt{\frac{\hat{V}\left(\sum_{i=1}^{n} w_i (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})\right)}{\left[\sum_{i=1}^{n} w_i x_i^2 - \frac{\left(\sum_{i=1}^{n} w_i x_i \right)^2}{\sum_{i=1}^{n} w_i} \right]}} \tag{9}$$

Jackknife

JK and bootstrapping are both based on resampling from the original data. The most common implementation of the jackknife is the delete-1 jackknife, where at each iteration, one member of the sample is removed randomly, and the parameters of interest are estimated using replicated weights, which are re-calculated after removing the observation. For delete-1 jackknife, let the w_i be the initial weight for an observation *i* and *n* be the sample size. Depending on the PS method selected, w_i may be IPTW weight obtained from Equation 1 or truncated weights:

$$w_{ik} = \begin{cases} 0 & \text{if the observation unit i is deleted at iteration k} \\ \frac{n}{n-1} w_i & \text{if the observation unit i is not deleted at iteration k} \end{cases} \tag{10}$$

At each iteration, the ATE ($\hat{B}_{1k}$), which is the parameter of interest, is re-calculated. The standard error will be as follows (Lohr, 1999):

$$SE(\hat{B}_1) = \sqrt{\frac{n-1}{n} \sum_{k=1}^{k} (\hat{B}_{1k} - \hat{B}_1)^2} \tag{11}$$

Comparison of Propensity Score Methods

Gu and Rosenbaum (1993) and Cepeda et al. (2003) found that optimal matching consistently outperforms matching with a greedy algorithm, which is the most commonly used algorithm for PS matching. Austin (2009a) found that matching on the propensity score within a specified caliper and IPTW methods removes more systematic differences between groups than PSS and covariance adjustment. Austin (2010a) also found that IPTW-CA method works better than PSS, matching on propensity score, IPTW, and covariance adjustment methods in terms of bias, variance estimation, coverage of confidence intervals, mean squared error, and Type I error rates. However, Austin only compared these methods under the condition of a binary outcome. Furthermore, Austin (2009b) evaluated standard error estimation methods for propensity score matching and found that methods that considered the matched nature of the data resulted in a smaller bias of standard errors and actual Type I error rates closer to the nominal Type I error rate. Leite et al. (2019) compared IPTW, OFM, and MMWS for the estimation of treatment effects of multivalued treatments and discovered that IPTW produced the lowest level of bias, followed by OFM and MMWS. They also found that standard errors of treatment effects were unbiased with IPTW but overestimated with OFM and MMWS regardless of whether TSL, JK, or bootstrapping had been used to estimate them. As summarized above, several studies have compared the relative performance of numerous PS and SE estimation methods in the literature. However, each specific research includes only a few methods. Very little is known about the performance of TIPTW and MMWFM compared to the performance of IPTW, PSS, MMWS, and OFM. Moreover, very few studies exists regarding -CA extensions of those PS methods combined with SE estimation methods. This study will fill the gap in the literature in three ways. Firstly, the performance of two not widely used PS methods (i.e., TIPTW and MMWFM) will be compared to more commonly used PS methods (IPTW, PSS, MMWS, OFM). Secondly, the performance of the -CA extensions of PS methods will be evaluated. Lastly, the performance of various SE estimation methods in combination with PS methods will be evaluated on a large scale. Given the scarcity of comparisons between these PS methods, the following research questions will be addressed in the current study:

1. Which propensity score method (IPTW, TIPTW, PSS, MMWS, OFM, and MMWFM) performs best with respect to unbiased estimation of ATE under conditions with different sample sizes and proportions treated?

2. Does the covariance adjustment using estimated propensity scores improve the performance of the propensity score methods (IPTW-CA, TIPTW-CA, PSS-CA, MMWS-CA, OFM-CA, and MMWFM-CA) with respect to the estimation of the ATE?

3. Which method (WLS, TSL, and JK) produces the most accurate standard errors when in combination with different propensity score methods?

4. Which propensity score method and standard error estimator combinations lead to the most power to detect the ATE?

Method

Data Simulation

In order to answer the research questions, a Monte Carlo simulation study was conducted using the R program (R Development Core Team, 2011). In order to obtain realistic population parameters to simulate data, we took estimates from the 2007-2008 School Survey on Crime and Safety (SSOCS) survey results (National Center for Education Statistics, 2010). The treatment variable was whether an outside school disciplinary plan was available, and the outcome was the total number of students involved in specified offenses. The covariates were the number of students transferred from another school, the typical number of classroom changes, the percentage of students below the 15th percentile in standardized tests, and the total number of transfers to specialized schools.

We generated multivariate-normally distributed covariates for the simulation study using the *MASS* package in R (Venables & Ripley, 2002). The first step of data simulation was to simulate the covariates $X_{1i}, X_{2i}, X_{3i}, X_{4i}$, which were normally distributed with population means of zero and population covariance matrix (obtained from SSOCS dataset) equal to:

$$\begin{bmatrix} 1.00 & .145 & -.004 & .125 \\ .145 & 1.00 & .001 & .467 \\ -.004 & .001 & 1.00 & .061 \\ .125 & .467 & .061 & 1.00 \end{bmatrix} \quad (12)$$

Secondly, we simulated residuals of the outcome regression. Residuals were simulated from a normal distribution with the mean of zero and the standard deviation of *166.278*. The population standard deviation of the residuals was defined so that the population R^2 for the outcome regression was *.211*. Once the covariates and the residual of the outcome were simulated, we obtained the potential control outcomes Y_C and potential treatment outcomes Y_T for all individuals in the sample based on following equations:

$$\begin{aligned} Y_{Ci} &= \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i \\ Y_{Ti} &= Y_{Ci} + ATE \end{aligned} \quad (13)$$

The population values of the coefficients β_0, β_1, β_2, β_3, and β_4 were *0*, *16.221*, *58.642*, *15.704*, and *33.601*. The population value of the ATE was *20*, which corresponds to a Cohen's effect size of *.085*, indicating that this is a small effect. The next step was to determine which individuals in the simulated samples were exposed to treatment. The population model for treatment assignment was:

$$\text{logit}(\text{P}(T_i = 1 | X_{1i}, X_{2i}, X_{3i}, X_{4i})) = \log(rt/(1-rt)) + \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i} \quad (14)$$

where the population values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3, \text{and } \gamma_4$ were *0*, *.127*, *.137*, *.166*, and *.101*, and *rt* is the proportion treated. The strength of the selection bias was defined based on the McKelvey and Zavoina pseudo R^2 (1975), and its population value for the simulated data was *.028*. We included the odds of being in the treated group $rt/(1-rt)$ in the treatment assignment model to control the proportion treated.

We manipulated sample size and the proportion of treated individuals only because those two characteristics (i.e., total sample size and the size of the treatment group relative to the total sample size) guide many researchers in terms of a quantitative research design. Monte Carlo simulation studies to compare PS methods by Gu and Rosenbaum (1993), Freedman and Berk (2008), and Austin (2009a) used *1000* as the only sample size. By manipulating the sample size, we were able to determine whether there were differences between the PS methods in terms of power to test the

ATE. We simulated data with sample sizes equal to *500*, *1000*, and *2000*.

We generated data where the proportion of the treated sample was set at *1/10*, *1/7*, *1/4*, *1/3*, and *1/2*. These conditions are an extension of Gu and Rosenbaum's (1993) study, which only examined ratios of *1/7*, *1/4*, and *1/3*.

We did not manipulate the number of covariates because Gu and Rosenbaum (1993) concluded that as long as the treatment assignment mechanism is modeled completely, the number of covariates does not affect the performance of the propensity score method, except for the potential problems of multicollinearity of covariates and convergence problems. We used four continuous covariates related to both the outcome and treatment assignment in this simulation study. Since we simulated data based on four covariates and estimated propensity scores using all four covariates, the assumption that the treatment assignment is modeled completely was met for all conditions in the study. Manipulated conditions in the data simulation are summarized in Table 1.

Table 1

Summary of the Manipulated Conditions in Data Simulation

Condition	Levels
Sample size	*500*, *1000*, and *2000*
Proportion of treated individuals	*1/10*, *1/7*, *1/4*, *1/3*, and *1/2*

Estimation of ATE and Standard Error

We simulated *1000* datasets per condition and analyzed the simulated datasets according to four steps for each dataset: 1) Estimate the PS for each individual using logistic regression; 2) Calculate the area of common support; 3) Estimate the ATE and standard error, ignoring selection bias to represent the baseline with an unweighted OLS regression model; 4) Implement propensity score methods. Using Equation 16, we estimated ATE with IPTW. After implementation of IPTW, we replaced the weights greater than 99^{th} percentile of the IPTW with the 99^{th} percentile to create TIPTW weights. In order to implement PSS, we grouped the treated and control individuals into five strata, based on similarity in PSs using the *MatchIt* package (Ho, et al., 2007) in R. We used five strata in that it is the most commonly-used number of strata in applications of PS methods in the social sciences (Thoemmes & Kim, 2011). In order to implement MMWS, weights were estimated using Equation 3. For the OFM method, we stratified the treated and control individuals based on similarity into a data-defined number of strata using the OFM algorithm implemented in the *optmach* package (Hansen & Klopfer, 2006) in R. In order to implement MMWFM, the same strata of OFM were used, but weights were calculated using Equation 11.

The ATE was estimated with weights using $Y_i = \beta_0 + \beta_1 T_i + e_i$, where β_1 is the ATE. To add covariance adjustment, we estimated the ATE with the model $Y_i = \beta_0 + \beta_1 T_i + \beta_2 PS + e_i$ where PS is the propensity score included as a covariate. ATE estimates with WLS standard errors were obtained using the *lm* function in R, while TLS standard errors and delete-1 JK standard errors were obtained with the *svyglm* function of the *survey* package (Lumley, 2011) in R.

Analysis

To measure the common support area of the propensity score, we used an overlap measure similar to Cohen's (1988) U_1 function, which is the proportion of non-overlap of the distributions. Let A and C be the non-overlap area at the lower and upper ends of the logit of propensity scores, and B

be the overlap area. Then,

$$U_1 = \frac{A+C}{A+B+C} \tag{15}$$

where increases in U_1 correspond to decreases in the area of common support. In the simulated conditions, the mean U_1 ranged from *.001* to *.604* with the mean and standard deviation of *.190* and *.091*, respectively. As the sample size or the proportion of the treated increased, the overlap also improved.

We compared the PS methods in terms of relative bias of ATE estimates and percent bias reduction. The relative bias of the ATE was calculated with $B(\hat{\theta}) = (\bar{\hat{\theta}} - \theta)/\theta$, where $\bar{\hat{\theta}}$ is the mean of the ATE estimates for all iterations of one condition, and θ is the population ATE. If the absolute value of the estimated $B(\hat{\theta})$ is larger than .05, the bias is considered unacceptable (Hoogland & Boomsma, 1998). We also evaluated the percent bias reduction, which is defined as:

$$PBR(\hat{\theta}) = \frac{B(\hat{\theta})_{baseline} - B(\hat{\theta})_{method}}{B(\hat{\theta})_{baseline}} x100 \tag{16}$$

where $B(\hat{\theta})_{method}$ is the mean relative bias of using a particular method, and $B(\hat{\theta})_{baseline}$ is the initial bias (Cochran & Rubin, 1973; Steiner et al., 2010). To be consistent with the Cochran and Rubin (1973) cutoff, we expected a PS method to remove at least *90%* of the initial bias to be considered successful.

We also compared the standard error estimation methods in terms of the relative bias of standard errors. The relative bias of the standard error is $B(S_{\hat{\theta}}) = [\bar{S}_{\hat{\theta}} - SD(\hat{\theta})]/SD(\hat{\theta})$, where $\bar{S}_{\hat{\theta}}$ the mean of the estimated standard errors of ATE is and $SD(\hat{\theta})$ is the empirical standard error, which is the standard deviation of estimated ATE. If the absolute value of the estimated $B(S_{\hat{\theta}})$ was larger than *.10*, the bias was considered unacceptable (Hoogland & Boomsma, 1998). For the methods that provided accurate ATE and standard error estimates, we estimated the power to test the ATE by calculating the proportion of ATE that were statistically significant at $\alpha = .05$ level for each condition.

Results

Comparison of Propensity Score Methods

We estimated the relative and the percent bias reduction for each one of the six propensity score methods. We ran two split plot ANOVAs where the relative bias of the estimated ATE and proportion bias reduction in the Estimated ATE were the outcome, the propensity score method was the within subjects factor, while sample size and the ratio of treated to untreated participants were between subjects factors. All possible interactions were also included. We used the generalized eta squared (η^2) as a measure of effect size to compare manipulated conditions. The propensity score method factor had $\eta^2 = .023$, and all other effect sizes were smaller than *.01*, indicating that relative and proportion bias reduction performance of PS methods are very similar at simulated sample sizes and proportion of treated individuals for any given PS method. Therefore, we collapsed the relative

bias results across sample size and treated to untreated ratio and show the relative bias and percent bias reduction in Table 2.

Table 2

Relative Bias and Percent Bias Reduction of the ATE Estimates by PS Method

ATE Estimation Method	Outcome model without covariance adjustment		Outcome model with covariance adjustment	
	Relative Bias	Percent Bias Reduction	Relative Bias	Percent Bias Reduction
IPTW	0.001	99.97%	-0.001	100.16%
TIPTW	0.086	92.54%	0.017	98.59%
PSS	0.116	89.96%	-0.002	100.24%
MMWS	0.116	89.96%	-0.004	100.37%
OFM	0.009	99.30%	0.002	99.90%
MMWFM	0.009	99.30%	0.002	99.90%

Note. The baseline relative bias across all iterations was 1.153.

We found that TIPTW, PSS, and MMWS methods exceed the *.05* relative bias criterion with relative bias of *0.086*, *0.116*, and *0.116*, respectively. In addition, PSS and MMWS removed close to *90%* of bias, while IPTW, OFM and MMWFM removed almost *99%* of bias. Truncation reduced bias removal, and TIPTW removed around *92%* of bias. Therefore, the relative and the percent bias reduction obtained with IPTW, TIPTW, OFM, and MMWFM were within acceptable levels. We also observed that the covariance adjustment with the propensity score improved the accuracy of the ATE estimates in all PS methods investigated; therefore, all of the propensity score methods, in combination with covariance adjustment, provided acceptable levels of relative bias and percent bias reduction in ATE estimates.

Comparison of Standard Error Estimation Methods

When examining the standard error estimates, we observed that some standard error estimates were unbiased while others were not. In order to investigate factors affecting the relative bias of the standard error estimates, we ran a split-plot ANOVA where the outcome was the relative bias of the standard error estimates, the between subjects factors were sample size and proportion treated, and within subjects factors were the PS method, the standard error estimation method, and covariance adjustment. We compared conditions based on the eta squared effect size measure (Olejnik & Algina, 2003). Table 3 shows the summary ANOVA table for the factors affecting the bias of the standard error estimates.

Table 3

Summary Effect Size Estimates for the Relative Bias of the Standard Error Estimates

Source	η^2
Between Subject Effects	
Ratio	0.109
Within Subject Effects	
PS Method	0.035
PS Method * Ratio	0.017
Covariance adjustment	0.047
standard error Method	0.327
standard error Method * Ratio	0.122
PS Method * standard error Method	0.062

PS Method * standard error Method *Ratio	0.030

Note. Effects that are smaller than .01 are not reported.

Results indicated that the standard error estimation method greatly affected the relative bias of the standard error estimates with generalized $\eta^2$of *.327*. However, there is also an interaction of the standard error estimation method and the proportion treated, which had a generalized η^2 of *.122*. The proportion treated had a generalized η^2 of *.109*. In addition to these effects, we have also observed various small effects. First, the two-way interaction regarding the PS and standard error method had a generalized η^2 of *.062*. Second, the three-way interaction of PS method, standard error method, and proportion treated had a generalized $\eta^2$of *.030,* and lastly, PS method and covariance adjustment had minor effects with a generalized η^2 of *.035* and *.047*. Given that the main effect and interactions involving sample size were negligible, we collapsed the relative bias of the standard errors over the three sample sizes.

Figure 1

Relative Bias of the Standard Error Estimates by Treated Ratio with IPTW and TIPTW

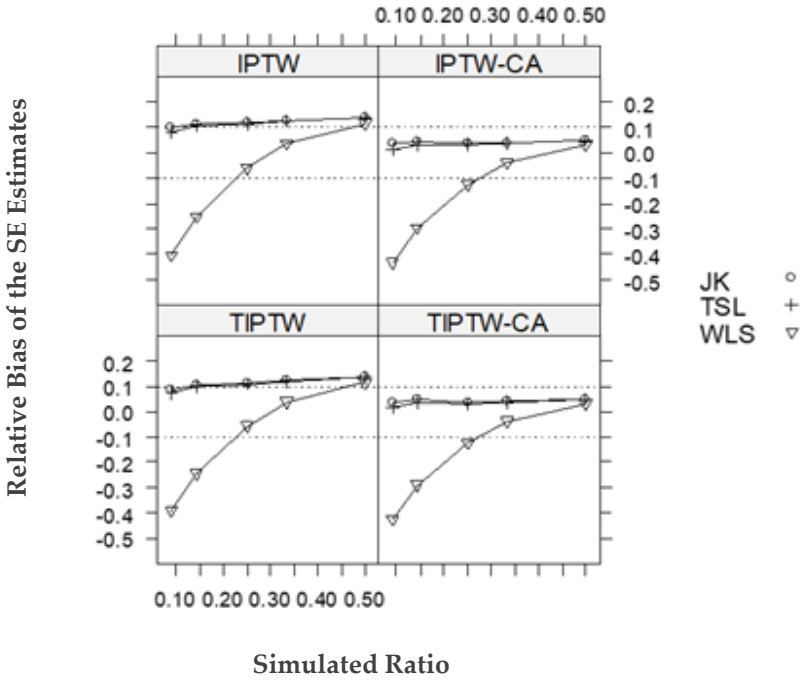
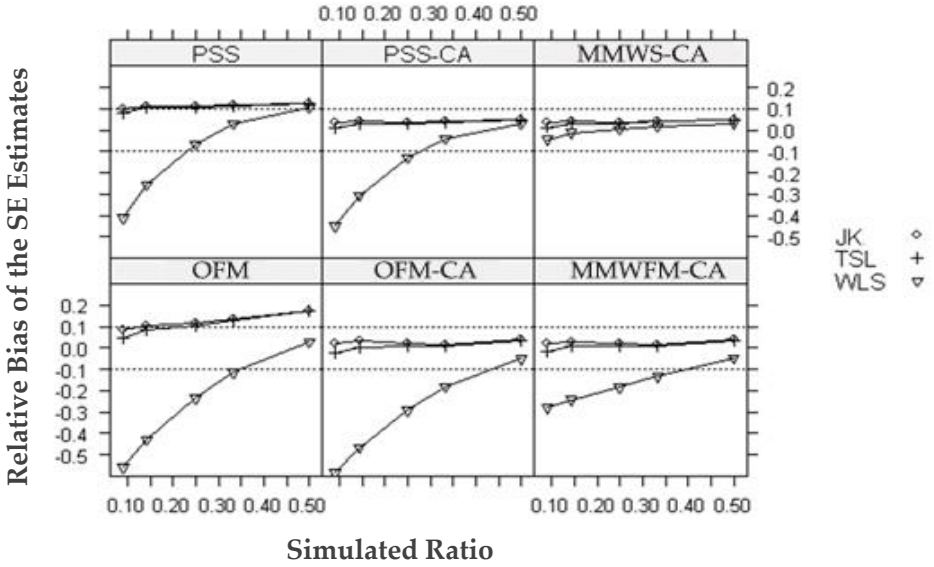


Figure 2

Relative Bias of the Standard Error Estimates by Treated Ratio for Stratification-Based PS Methods

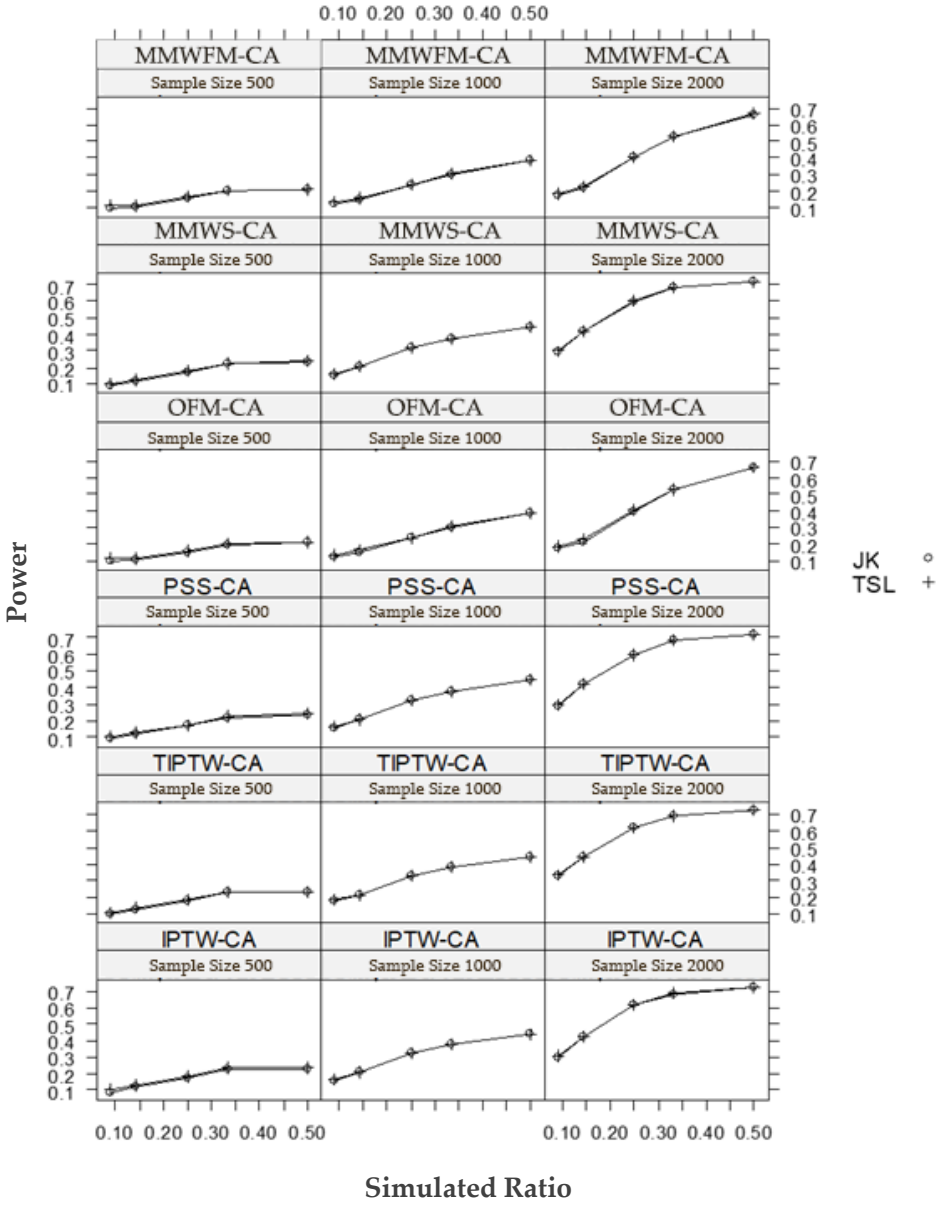


Relative biases of the standard error estimates are summarized in Figure 1 and Figure 2. We found that TSL and JK provided similar standard errors across all simulated conditions. The relative bias of the standard errors using TSL and JK methods for any PS method by proportion treated ranged from *-.024* to *.177* with the mean of *.058*. Most of this relative bias of the standard error estimates did not exceed the (*-.1,.1*) range, but as the proportion treated increased from *0.1* to *0.5*, we found that TSL and JK tended to overestimate the standard errors marginally to moderately. However, adding covariance adjustment to the outcome model for each PS method reduced the bias of the standard error estimates down to acceptable levels. Overall, TSL resulted in slightly less biased standard errors than JK. WLS standard errors ranged from *-.585* to *.117* with the mean of *-.146*. WLS underestimated standard errors with the lowest three proportions treated except MMWS-CA and some PS methods when the simulated ratio was *.25.* The MMWS-CA method with all three standard error estimation methods provided acceptable standard errors under any simulated condition. As the proportion treated increased from *0.1* to *0.5*, we saw a quadratic improvement in relative bias of the estimated standard errors. PS methods based on weighting with covariance adjustment in the outcome model combined with WLS estimation provided marginally more biased standard errors than PS methods without covariance adjustment. This relationship was inverted on the matching-based PS and MMWT methods providing less biased estimates compared to non-MMWT methods. When the proportion treated was *0.5*, WLS provided less biased standard errors than TSL and JK across all simulated conditions.

Considering that some of the PS method and standard error estimation method combinations provided biased standard errors for various simulated conditions, we restricted the examination of power to PS methods with outcome model with covariance adjustment combined with standard error estimation with TLI and JK because they removed the initial bias and provided unbiased standard errors across all simulated conditions.

Figure 3

Power to Detect the ATE as a Function of Treated Ratio, Sample Size, and Propensity Score Method



Based on the results plotted in Figure 3, we observed that the sample size and proportion treated are the two most important factors affecting power. As the design gets balanced or sample size increases, all PS methods with covariance adjustment paired with TSL and JK standard error estimation provided increased power. We have also observed that the increase in power as the design gets balanced was greater for larger sample sizes. In order to compare the power of TSL and JK in combination with six PS methods with covariance adjustment in the outcome model, we collapsed the power across the proportion treated and sample size. We found that the overall differences in power with TSL and JK standard error estimation methods with different PS methods were less than *.01* in favor of the TSL. Finally, we collapsed the power across the proportion treated, sample size, and standard error estimation method to observe the overall differences among the PS methods. Results indicated that MMWFM-CA and OFM-CA were the least powerful procedures, with averaged power of *.266* and *.267*, respectively. The remaining four PS methods were more powerful, with the overall power levels ranging from *.339* to *.349*.

Discussion and Conclusion

We investigated the appropriateness of the ATE estimate (i.e., relative bias and percent bias reduction), the standard error for the ATE estimate (i.e., relative bias), and power for different combinations of estimations methods and standard error estimation methods. Our first research question was about the appropriateness of ATE estimates using different PS methods. With IPTW, we observed that all of the initial bias was removed. Trimming the extreme values greater than the 99th percentile of the IPTW weights (i.e., TIPTW) performed acceptably but poorer than IPTW. This finding is consistent with the findings of Austin (2009a; 2010a), whereas contrast to findings of Freedman and Berk (2008) and Sturmer et al. (2010). In both Freedman and Berk (2008) and Sturmer et al. (2010), data was simulated in such a way that the extreme weights were highly influential regarding the treatment effect estimate. In our study, weights greater than *20*, which was the trimming rule investigated by Freedman and Berk (2008), occurred in less than 1% of the simulated iterations. Furthermore, we only trimmed the upper end of the weight distribution. The poorer performance of the TIPTW as compared to the IPTW is due to the fact that cases with large weights in the control group are similar to the treated group concerning covariate distributions. Therefore, cases with large weights are most important in balancing covariate distributions. Given these new results and the existing literature, it is recommended that applied researchers examine their weights to determine whether extreme weights occurred and how extreme they are and to compare estimates with and without truncating extreme weights.

Parallel to findings from Gu and Rosenbaum (1993) and Cepeda et al. (2003), we observed that OFM removed almost all of the initial bias in the ATE estimate. Similar to the findings from Cochan (1968) and Rosenbaum and Rubin (1984), PSS with five strata removed about *90%* of the initial bias. Austin (2009a; 2010a) and Lunceford and Davidian (2004) concluded that IPTW removes more systematic differences compared to PSS, whichwas observed in our study as well. OFM outperformed PSS as expected because OFM is a type of stratification into a maximum number of strata containing at least one treated and one untreated observation. As we expected, based on the definition of marginal mean weights, results from OFM and MMWFM, and PSS and MMWS were identical. Similar to the findings of Austin (2010a) and claims of Stuart (2010), performing covariance adjustment for the propensity score in the outcome model after applying a propensity score method removed more systematic differences for methods with poorer performance (i.e., TIPTW and PSS). However, adding covariance adjustment had no effect on the methods that removed almost all of the initial bias (i.e., IPTW and OFM). Therefore, we observed minimal differences between OFM-CA and MMWFM-CA as well as between PSS and MMWS-CA. All of the PS methods investigated with covariance adjustment removed almost all of the initial bias of the ATE.

Our second research question was about obtaining accurate standard error estimates using different combinations of PS methods and standard error estimation methods for different combinations of proportion treated and overall sample size. Our ANOVA indicated that PS method, standard error method, covariance adjustment, and the proportion treated affect the accuracy of the standard error estimates. First of all, we observed a severe under-estimation of the standard errors using WLS for the smallest two simulated ratios for all PS methods except MMWS-CA. As the proportion treated approaches *0.5*, the bias of standard errors with WLS declined. Comparing TSL and JK, we observed that the standard errors were similar to each other, but TSL standard errors were marginally less biased than JK standard errors. Third, we observed marginally or moderately over-estimation of standard errors using TSL and JK when the proportion treated was *0.5* in combination with any of

the PS methods. The overestimation was not extreme but large enough to make the significance test of the treatment effect artificially conservative. When we included the propensity score as a covariate in the treatment effect estimation model, we observed improvement in the accuracy of the standard errors regardless of the combination of PS method, standard error estimation method, and proportion treated. The improvement was not large enough for WLS standard errors to become unbiased, but all of the TSL and JK standard errors were unbiased when covariance adjustment was used under all simulated conditions. Therefore, we conclude that WLS standard errors should not be used with PS weights. However, WLS is a default in some software when weights are provided, such as the *lm* function in the R software package, so applied researchers must be careful about how weights are handled in standard error estimation. This study did not examine bootstrapping methods to estimate standard errors, but because of its similarity to jackknife, it is reasonable to expect it to perform similarly.

Given the results of this study, we recommend models for estimation of the ATE with PS weights and additional propensity score covariance adjustment because it not only provided additional bias removal of ATE estimates but also improved estimates of standard errors. However, a limitation of the simple outcome model used in this simulation study for covariance adjustment is that it assumes that the propensity score is linearly related to the outcome. In practice, it is safer to avoid this assumption by also including the square and cube of the propensity score in the model or dummy-coded strata of the propensity score. Furthermore, the outcome model should be expanded to include covariates that are strongly related to the outcome, such as a pre-test outcome measure, since it would reduce residual variability and increase power. An expanded outcome model with key covariates also has the double-robustness property, where bias may be removed by the covariates being correctly modeled in either the PS model or the outcome model. One advantage of PS weighting methods is that they can be used in combination with any outcome model as long as an estimator that incorporates weights, such as pseudo-maximum likelihood estimation (Asparouhov, 2006), is available.

Direct adjustment of the propensity score or covariates in the outcome model in addition to PS weighting is recommended because it is simple to implement and familiar to applied researchers. However, other doubly robust estimators of the ATE are available, which are described in Bang and Robins (2005), Kang and Schafer (2007), and Lunceford and Davidian (2004). Future research could compare these doubly robust estimators regarding bias reduction and robustness to model misspecification.

This study is also limited in that it only looked at conditions where SUTVA is met. Research on how to handle violations of SUTVA has been incipient, with Hong and Raudenbush (2006) being a rare example where SUTVA is addressed directly. Besides, datasets in educational research commonly have a clustered structure. However there have been few studieson propensity score weighting methods for data with multilevel structure, such as Arpino and Mealli (2011), Leite et al. (2015), and Thoemmes and West (2011),which means future research should expand on this area. Further research may include different data structures with distinct covariance matrices and covariate distributions.

Ethics Committee Approval: The study was conducted as a simulation study. In addition, studies produced from master's/doctorate theses before 2020 do not require an ethics committee report.

Author Contributions: Since the study was produced from a master's thesis, the first author contributed to all

parts of the study, and the second author, who is the academic advisor, contributed by guiding the process and made suggestions for additions and / or corrections to all parts if necessary.

Conflict of Interest: The authors declare that they have no conflict of interest.

Uluslararası Türk Eğitim Bilimleri Dergisi,
2023, 11 (21), 989-1031
https://doi.org/10.46778/goputeb.1312865

ISSN: 2148-2314

Ortalama İşlem Etkisi Kestiriminde Seçim Yanlılığını Gidermek İçin Eğilim Puanı Ağırlıklandırma Metotlarının Karşılaştırılması *

Sungur Gürel[a] Walter Lana Leite [b]

[a] *Dr. Öğr. Üyesi., Siirt Üniversitesi, Siirt, Türkiye, s.gurel@siirt.edu.tr*

[b] *Prof., Florida Üniversitesi, Florida, Amerika, walter.leite@coe.ufl.edu*

ÖZET

Bu Monte Carlo simulasyon çalışmasında, ters olasılık ağırlıklandırması, kesilmiş ters olasılık ağırlıklandırması, eğilim puanı tabakalandırması, eğilim puanı tabakalandırması üzerinden marjinal ortalama ağırlıklandırması, optimal tam eğilim puanı eşleştirmesi ve optimal tam eğilim puanı eşleştirmesi üzerinden marjinal ortalama ağırlıklandırması olmak üzere bireylerin ağırlıklandırmalarına dayalı altı farklı eğilim puanı metodu uygulamasının performansı araştırılmıştır. Bu metotlar gözlemsel çalışmalarda kestirilen ortalama işlem etkisinde bulunan seçim yanlılığını düşürmeyi amaçlar. Ağırlıklandırma ile ortalama işlem etkisinin standart hatası kestiriminde ağırlıklandırılmış en küçük kareler, Taylor serileri doğrusallaştırma ve jackknife metotları kullanılmıştır. Araştırma sonucunda, simule edilen bütün durumlarda kovaryans düzeltmesi ilaveli ağırlıklandırmaya dayalı eğilim puanı metotlarının Taylor serileri doğrusallaştırma ve jackknife standart hata kestirim metotları ile birlikte kullanılması ile seçim yanlılığının uygun bir şekilde ortadan kaldırdığı ve doğru standart hataların kestirildiği bulunmuştur.

MAKALE BİLGİSİ

Makale Türü
Araştırma

Makale Geçmişi
Gönderim tarihi:
13.06.2023
Kabul tarihi:
22.09.2023

Anahtar Kelimeler
Gözlemsel Çalışma,
Eğilim Puanı Metodları,
Ortalama İşlem Etkisi,
Standart Hata

Atıf Bilgisi: Gürel, S. ve Leite, W L. (2023). Ortalama işlem etkisi kestiriminde seçim yanlılığını gidermek için eğilim puanı ağırlıklandırma metotlarının karşılaştırılması. *Uluslararası Türk Eğitim Bilimleri Dergisi, 11* (21), 989-1031.

Sorumlu yazar: Sungur Gürel, e-posta: s.gurel@siirt.edu.tr

* Bu çalışma 2012 yılında Florida Üniversitesinde kabul edilen "The performance of propensity score methods to estimate the average treatment effect in observational studies with selection bias: A Monte Carlo simulation study" başlıklı tezde üretilmiştir. Ayrıca 2012 Yılında Amerikan Eğitimsel Araştırma Derneği'nin Yıllık Toplantısında kısmen "Comparison of Inverse Probability of Treatment Weighting and Optimal Full Matching Methods to Estimate the Average Treatment Effect: A Monte Carlo Simulation Study" Başlıklı sözlü sunumda sunulmuştur.

Giriş

Ulusal Eğitim İstatistikleri Merkezi (NCES, İng. National Center of Educational Statistics) ve Ulusal Bilim Vakfı (NSF, İng. National Science Foundation) gibi kurumlar tarafından toplanan ve ulusal düzeyde temsil gücü olan çeşitli veri tabanlarının mevcudiyeti nedeniyle, eğitimsel işlemlerin etkilerinin ikincil veriler kullanılarak kestirilmesi eğitim araştırmalarında yaygın hale gelmiştir (Strayhorn, 2009). Ancak, bu ulusal çalışmalarda katılımcıların işlem gruplarına atanması rastgele olmadığından, işlemlerin etkilerine ilişkin kestirimler hem gözlenen hem de gözlenmeyen ortak değişkenler nedeniyle seçim yanlılığına açıktır (Shadish vd., 2002). Son kırk yılda, işlem koşullarına rastgele atama yapılmayan çalışmalarda işlem etkilerini kestirmek ve seçim yanlılığıyla başa çıkmak için çeşitli metotlar ortaya çıkmıştır (Heckman, 1978; Rosenbaum ve Rubin, 1983; Abadie ve Imbens, 2006; Heckman vd., 1997) ve bunlar toplu olarak gözlemsel çalışmalar olarak adlandırılmaktadır. Eğilim puanı metotları, sosyal bilim araştırmalarında gözlemsel çalışmaların analizinde en sık kullanılan metotlar arasındadır (Thoemmes & Kim, 2011). Eğilim puanı metotları, işlem etkisi kestirimlerindeki seçim yanlılığıyla başa çıkmak için, işleme tabi olan ve olmayan katılımcılar arasında gözlemlenen ortak değişkenler üzerinde önceden var olan farklılıkları dengelemeye çalışır. Rosenbaum ve Rubin (1983) eğilim puanı (PS, İng. Propensity Score) terimini ilk kez kullanmış ve gözlenen ortak değişkenler göz önüne alındığında işleme tabi olma olasılığı kestirimi olarak tanımlamıştır. İşleme tabi olmanın gözlenen ortak değişkenlere bağlı olması durumunda, bir eğilim puanı düzeyinde işleme tabi olan ve olmayan katılımcılar arasındaki gözlenen farkın, o düzeydeki ortalama işleme tabi olma etkisinin (ATE, İng. Average Treatment Effect) yansız bir kestirimi olduğunu kanıtlamışlardır (Rosenbaum & Rubin, 1983). Eğilim puanları, gözlemleri PS'deki benzerliklerine göre eşleştirerek, gözlemleri PS'nin tersi ile ağırlıklandırarak veya gözlemleri PS'ye göre homojen gruplara ayırarak ATE kestirimindeki seçim yanlılığını azaltmak için kullanılabilir (Stuart, 2010). Bire bir eşleştirme dışında, tüm PS metotları gözlem ağırlıkları üretir ve bu nedenle örnekleme ağırlıklandırmaya dayalı kestirim metotlarına benzer şekilde işlem etkisi kestiriminde PS ağırlıklandırma uygulanabilir (Leite, 2016). Lunceford ve Davidian (2004), örnekleme literatüründe (örneğin, Lohr, 1999) tartışılan kestiricilerle eşdeğer olarak eğilim puanı ağırlıklandırma ve tabakalandırma ile elde edilen işlem etkisi kestiricilerini göstermektedir. Ayrıca, karmaşık anket verileriyle standart hata kestiriminde kullanılan bootstrapping, jackknife ve Taylor serileri doğrusallaştırma gibi metotlar (Stapleton, 2008) eğilim puanı ağırlıklarına sahip verilere de uygulanabilir.

Literatürde çeşitli işlem etkileri tanımlanmış olsa da (Guo & Fraser, 2010), sosyal bilimler literatüründe en yaygın olarak bulunan kestirimler ortalama işlem etkisi (ATE, İng. Average Treatment Effect) ve işleme tabi olanlar üzerindeki ortalama işlem etkisidir (ATT, İng. Average Treatment Effect on Treated) (Thoemmes & Kim, 2011). Bir PS metodunun spesifik uygulaması, ATE veya ATT'nin ilgi alanına girip girmediğine bağlı olarak farklılık göstermektedir. ATT'yi kestirmek etmek için PS metotlarının uygulamalarını karşılaştıran birkaç çalışma yapılmıştır (Gu ve Rosenbaum, 1993; Cepeda vd., 2003; Austin, 2010b; Harder vd., 2010), ancak ATE'nin kestirimi için başlıca PS metotlarını karşılaştıran bir çalışma yapılmamıştır. Bu nedenle, bu çalışmanın ilk amacı, ters olasılık ağırlıklandırması (IPTW, İng. Inverse Probability of Treatment Weighting), kesilmiş olasılık ağırlıklandırması (TIPTW, İng. Inverse Probability of Treatment Weighting), eğilim puanı tabakalandırması (PSS, İng. Propensity Score Stratification), eğilim puanı tabakalandırması üzerinden marjinal ortalama ağırlıklandırması (MMWS, İng. Marginal Mean Weighting through Propensity Score Stratification), optimal tam eğilim puanı eşlemesi (OFM, İng. Optimal Full

Propensity Score Matching), optimal tam eğilim puanı eşleştirmesi üzerinden marjinal ortalama ağırlıklandırması (MMWFM, İng. Marginal Mean Weighting through Optimal Full Propensity Score Matching) ve bunların kovaryans düzeltmesi uzantılarını ATE kestirimlerinde seçim yanlılığını azaltma yetenekleri açısından karşılaştırmaktır. PS metotlarını karşılaştıran çalışmaların çoğu işlem etkisi kestirimine odaklandığından ve standart hataların kestirimine değinmediğinden, bu çalışmanın ikinci amacı, her bir PS metoduyla elde edilen ATE kestiriminin standart hatalarını kestirmek etmek için ağırlıklı en küçük kareler regresyonu (WLS, İng. Weighted Least Squares), Taylor serisi doğrusallaştırma (TSL, İng. Taylor Series Linearization) ve jackknife (JK, İng. Jackknife) metotlarını karşılaştırmaktır.

Potansiyel Çıktılar Teorik Çerçevesi

Rubin'in potansiyel çıktılar teorik çerçevesi (1974), gözlemsel çalışmalarda seçim yanlılığı sorununu ele almak için yaygın olarak kullanılmaktadır. Temel ilkesi, işleme tabi olan ve olmayan bireylerin hem işlemin varlığı hem de yokluğunda potansiyel çıktılara sahip olmasıdır. Örneğin, işleme tabi olan bir *i* katılımcısının gözlemlenen çıktısı Y_{iT}^{t} olsun, Y_{iT}^{c} ise bu katılımcının kontrol grubunda yer alması durumunda ortaya çıkacak potansiyel çıktısı olsun. Benzer şekilde Y_{iC}^{c} kontrol grubu katılımcısı *i* için gözlemlenen çıktıdır ve Y_{iC}^{t} ise aynı katılımcının işleme tabi olması durumunda ortaya çıkacak potansiyel çıktıdır. Başka bir deyişle, kontrol grubu katılımcısı işleme tabi olması koşulu altında potansiyel bir çıktıya sahiptir. Tersine, işleme tabi olan bir katılımcı ise kontrol koşulu altında potansiyel bir çıktıya sahiptir. ATE, işleme tabi olmaları durumunda tüm bireyler için potansiyel çıktıları ile işleme tabi olmamaları durumunda tüm bireyler için potansiyel çıktılar arasındaki fark olan $E[Y_i^t] - E[Y_i^c]$'dır (Winship & Morgan, 1999). Randomize çalışmalarda ATE, $E[Y_{iT}^{t}]$ - $E[Y_{iC}^{c}]$'ye eşittir ancak gözlemlenen çalışmalarda $E[Y_{iT}^{c}] = E[Y_{iC}^{c}]$ ve $E[Y_{iC}^{t}] = E[Y_{iT}^{t}]$ koşulları sağlanmadığı için bu eşitlik sağlanmaz. Ayrıca randomize çalışmalarda $E[Y_{iT}^{t}] - E[Y_{iT}^{c}]$ olarak tanımlanan ATT, ATE'ye eşittir ancak gözlemsel çalışmalarda bu eşitlik her zaman doğru değildir. Bu nedenle ATE'nin kestiricisi olan $\overline{Y}_{iT}^{t} - \overline{Y}_{iC}^{c}$'nin (yani işleme tabi olan ve işleme tabi olmayan bireylerin gözlemlenen çıktıları ortalamaları arasındaki farkın) yansız olması için işleme tabi olma ve olmama durumlarına atanmanın potansiyel çıktılardan bağımsız olması gereklidir. Daha açık bir ifadeyle, Y_i^t'nin işleme tabi olma durumunda elde edilen potansiyel çıktı, Y_i^c'nin işleme tabi olmama durumunda elde edilen potansiyel çıktı, X'in bütün ortak değişkenler ve T'nin işleme tabi olma durumu değişkeni olması halinde $T \perp \left\{Y_i^t, Y_i^c\right\} \mid X$ koşulu sağlanmalıdır. Bu koşul, işleme tabi olmanın güçlü ihmal edilebilirliği (SITA, İng. Strong Ingorability of Treatment Assigmnment) olarak bilinir (Rubin, 1974). Ayrıca, bir birimin potansiyel çıktısının belirli bir işleme tabi olma durumuna atanmasından veya diğer birimlerin potansiyel çıktılarından etkilenmemesini gerektiren sabit birim işleme tabi olma değeri varsayımının (SUTVA, İng., Stable Unit Treatment Value Assumption) karşılanması da gereklidir (Rubin, 2007). Gözlemsel çalışmalarda, hem SITA hem de SUTVA ihlal edilebilir, bu da ATE'nin yanlı kestirilmesine ve çalışmanın iç geçerliliğinin zayıf olmasına yol açar (Shadish, 2002). Gözlemsel çalışmalarda, SUTVA'nın geçerli olduğu varsayımı altında, işleme tabi olma ile olmama grupları arasında gözlenen ortak değişkenlerin dağılımlarını dengeleyerek işleme tabi olma grubuna atanmanın güçlü ihmal edilebilirliğini sağlamaya çalışmak için PS metotları kullanılır. Bir ebeveynin bir öğrenciyi bir eğitimsel işleme tabi olmasının kararının öğrencinin sınıf arkadaşlarının işleme tabi olma durumundan etkilenmesi gibi SUTVA ihlalleri, eğilim puanlarının

kestiriminde ve PS metotlarının uygulanmasında başka yerlerde tartışılan özel hususları gerektirir (Arpino ve Mealli, 2011; Hong ve Hong, 2008; Thoemmes ve West, 2011; Leite vd., 2015).

ATE'de Seçim Yanlılığını Azaltmak için Eğilim Puanı Metotları

Herhangi bir PS metodunun kullanılması, hem işleme tabi olma koşullarına seçim hem de çıktılarla ilgili olan gözlenen ortak değişkenlerin seçilmesiyle başlayan çok adımlı bir süreç gerektirir (değişken seçimine ilişkin bir tartışma için bkz. Brookhart vd., 2006). İkinci adım, en yaygın olarak lojistik regresyon ile gerçekleştirilen eğilim puanlarını kestirmektir, ancak diğer parametrik modeller veya parametrik olmayan modeller de (McCaffrey vd., 2004) kullanılabilir. Üçüncü adım, kestirilen eğilim puanlarının ortak destek alanını değerlendirmektir; bu, hem işleme tabi olan hem de olmayan gruplar için değerlerin bulunduğu eğilim puanı dağılımının alanıdır (Guo & Fraser, 2010). Eğilim puanı dağılımının belirli bir alanı için ortak desteğin olmaması, kestirimlerin genellenebilirliğini yalnızca ortak desteğin olduğu alt popülasyona kısıtlar. Dördüncü adım, tercih edilen eğilim puanı metodu göz önüne alındığında ortak değişkenlerin dağılımının dengesini doğrulamaktır. Beşinci adım, PS metodu ile parametrik veya parametrik olmayan kestiricilerle birleştirerek ATE ve standart hatasının kestirilmesi ile ATE'nin istatistiksel anlamlılığı hakkında sonuçlara ulaşmaktır. Son adım ise sonuçların önemli ortak değişkenlerin olası ihmaline karşı duyarlılığını değerlendirmektir (Rosenbaum, 2010).

Eğilim puanlarının ATE'nin parametrik, model tabanlı kestirilmesi birleştirilmesi (Ho vd., 2006), ATE kestirimlerinin yanlılığını azaltma avantajına sahiptir çünkü araştırmacının doğrusal modeller, genelleştirilmiş doğrusal modeller, karışık etkili modeller ve yapısal denklem modelleriyle sonuç hakkındaki karmaşık hipotezleri değerlendirmesine olanak tanır. Eğilim puanı metotları, sonlu anket örneklemlerinden model kestiriminde örnekleme ağırlıklarına benzer şekilde kullanılan ağırlıklar oluşturularak, model tabanlı kestirime dâhil edilebilir (Leite, 2016). Aşağıda yaygın olarak kullanılan eğilim puanı metotları ve her bir metotla ağırlıkların hesaplanması açıklanmaktadır.

Ters Olasılık Ağırlıklandırması (IPTW)

Olasılık Ağırlıklandırması, 20. yüzyılın ortalarında Horvitz ve Thompson (1952) tarafından anket kestirimlerde örnekleme deseninin etkisini hesaba katmak için ortaya atılmıştır. Robins ve diğerleri (2000), gözlemsel çalışmalarda seçim yanlılığını kontrol etmek için bu kavramı IPTW olarak genişletmiştir. IPTW'nin arkasındaki fikir, katılımcıları gerçekte içinde bulundukları grupta olma koşullu olasılığının tersi ile ağırlıklandırmaktır. $T_i = 1$, işleme tabi olan bir katılımcıyı, $T_i = 0$ ise işleme tabi olmayan bir katılımcıyı temsil etmek üzere; T_i işleme tabi olma durumunun gösterge değişkeni olsun. $\hat{e}_i$ 'de kestirilmiş eğilim puanı olsun. ATE'yi kestirmek için *i* katılımcısının ağırlığı w_i (Stuart, 2010) denklem 1'de sunulmuştur.

$$w_i = \frac{T_i}{\hat{e}_i} + \frac{1 - T_i}{1 - \hat{e}_i} \qquad (1)$$

IPTW, gözlemlerin ağırlıklara göre çoğaltıldığı sahte bir evren yaratır, böylece katılımcılar sadece kendilerini değil, aynı zamanda diğer gruptaki benzer özelliklere sahip olanları da hesaba katar (Hernan vd., 2004). Neugebauer ve van der Laan (2005) IPTW'nin performansının, tüm ağırlıkların sıfırdan farklı olmasını gerektiren deneysel işleme tabi olma ataması varsayımına bağlı olduğunu bulmuştur. Ayrıca, herhangi bir işleme tabi olma olasılığının sıfıra yakın olması durumunda, yeni

ağırlıklı örneklemin hedef evreni temsil etmeyebileceğini bulmuşlardır.

Kesilmiş Ters Olasılık Ağırlıklandırması (TIPTW)

IPTW metodu, ağırlıklar aşırı olduğunda veya eğilim puanları aşırı olduğunda gösterdiği performans açısından eleştirilmiştir (Freedman ve Berk, 2008). Aşırı ağırlıklar aşırı etkili gözlemler yaratmakta ve kestirimlerin örnekleme çeşitliliğini artırmaktadır. Birçok araştırmacı bu sorunu çözmek için farklı çözümler üretmiştir. Bembom ve van der Laan (2008) IPTW kestiricileri için veriye uyarlanabilir bir kesme seviyesi seçimi geliştirmiştir. Kestirimlerin ortalama karesel hatasında *%7*'ye kadar verimlilik elde etmeyi başarmışlardır. Freedman ve Berk (2008) 20'den büyük olan ağırlıkları 20 ile değiştirmiş böylelikle 20'den büyük gözlemleri kesmiştir. Ancak, her iki metodun da seçim yanlılığını azaltamadığı sonucuna varmışlardır. Sturmer ve diğerleri (2010), 2,5. ve 97,5. yüzdelik diliminden daha uç olan eğilim puanlarına kadar kesmenin, hiçbir gözlemi kesmemeye ve daha fazla gözlemi kesmeye kıyasla seçim yanlılığını azalttığını bulmuştur. Crump ve diğerleri (2009), ATE kestirimine ek yanlılık getirmeden işlem etkisi kestiriminin asimptotik varyansını azaltmak için yalnızca optimal seçim kuralını karşılayan veri alt örneklemi kullanarak ATE'yi kestirmeyi etmeyi önermiştir. Optimal seçim kuralı, aşırı eğilim puanları olmayan gözlemleri tutmak için basitçe üst ve alt sınırlar koyar. Araştırmacılar ayrıca, eğilim puanları çeşitli beta dağılımlarını takip ettiğinde *[.1, .9]* temel seçim kuralının optimal seçim kuralı kadar iyi çalıştığı sonucuna varmıştır. Aşırı eğilim puanlarının neden olabileceği olası aşırı ağırlıklarla uğraşmak yerine, aşırı ağırlıkların istenmeyen sonuçlarıyla başa çıkmak için daha basit bir yaklaşım öneriyoruz. IPTW ağırlıkları hesaplandıktan sonra, IPTW ağırlıklarının *99.* yüzdelik diliminden daha büyük olan ağırlıklar *99.* yüzdelik dilimi temsil eden ağırlık değerine kesilerek TIPTW ağırlıkları elde edilir.

Eğilim Puanı Tabakalandırması (PSS)

PSS, eğilim puanları açısından benzer olan bireyleri içeren tabakalar oluşturmaktan ibarettir (Stuart, 2010); burada her tabaka en az bir işleme tabi olan ve bir işleme tabi olmayan birey içermelidir. PSS genellikle eğilim puanlarının dağılımını eşit büyüklükteki aralıklara bölerek gerçekleştirilir. Seçim yanlılığını azaltmak için tek bir ortak değişkene dayalı tabakalandırma Cochran (1968) tarafından önerilmiştir, ancak Rosenbaum ve Rubin (1984) eğilim puanlarına dayalı beş tabakaya ayırmanın seçim yanlılığının yaklaşık %90'ını azalttığını göstermiştir. Uygulamalı sosyal bilim araştırmalarında, Thoemmes ve Kim (2011) çoğu çalışmanın 5 ila 10 tabaka kullandığını bulmuştur. ATE'yi kestirmek için tabakaların elde edilmesi, örneklemin tüm üyelerinin bir tabakaya yerleştirilmesini gerektirirken, yalnızca işleme tabi olmayan gözlemleri içeren tabakalar ATT kestiriminde çıkarılabilir. Ayrıca, ATE'nin kestirimesi, aşağıdaki formüle göre her bir tabakadaki birey sayısına göz önüne alınarak katılımcıların ağırlıklandırılmasını gerektirmektedir;

$$w_i = 0.5\left[T_i \frac{n^s}{n_s^t} + (1-T_i)\frac{n^s}{n_s^c} \right] \qquad (2)$$

Burada *s* simgesi tabaka üyeliğini, *t* ve *c* işleme tabi olma ve olmama durumunu göstermektedir. n^s , tabaka s için tabaka büyüklüğünü, n_s^t tabaka *s*'deki işleme tabi olan katılımcı sayısını, n_s^c ise tabaka *s*'deki işleme tabi olmayan katılımcı sayısını gösterir. Buna karşın ATT ağırlıkları her bir tabakadaki işleme tabi olan katılımcı sayısına dayalı olarak hesaplanır.

Optimal Tam Eğilim Puanı Eşleştirmesi (OFM)

Tam eşleştirme, alt sınıfların sayısının gözlemlenen verilere göre şekillendirildiği bir tabakalandırma metodudur (Leite, 2016). Aynı alt sınıftan rastgele bir işleme tabi olan ve işleme tabi olmayan birim seçildiğinde, belirli bir mesafe ölçüsüne göre bu ikisi arasındaki beklenen fark Δ 'ıir. OFM algoritması, tüm örneklem için minimum maliyetli bir akış bularak eşleşen kümeler içinde Δ 'yı en aza indirmek için ağ akışı teorisine (İng., Network Flow Theory) dayalı olarak oluşturulmuştur (Rosenbaum, 1989:1991). İşleme tabi olan *i* katılımcısının eğilim puanı ile işleme tabi olmayan *j* katılımcısının eğilim puanı arasındaki mesafe olarak $\Delta_{ij} = \left|\hat{e}_{it} - \hat{e}_{jc}\right|$ ile OFM en sık eğilim puanları ile gerçekleştirilir. Rosenbaum (1991), *T*'nin işleme tabi olan katılımcı sayısı, *C*'nin ise işleme tabi olmayan katılımcı sayısı olduğu durumda her zaman $\sum_{i,j}^{T,C} \Delta_{ij}$ 'nin minimize edildiği optimal bir eşleştirme olduğunu bulmuştur. OFM ile ATE'yi kestirmek, PSS ile aynı şekilde hesaplanan ağırlıklar gerektirir (bkz. Denklem 2).

Eğilim Puanı Tabakalandırması Üzerinden Marjinal Ortalama Ağırlıklandırması (MMWS)

Hong (2012), örneklemdeki her bir işleme tabi olan veya olmayan katılımcının, her bir tabaka içerisinde adeta işleme tabi olma ve olmama durumuna rastgele atanmış gibi ayarlanması için MMWS'yi önermiştir. İşleme tabi olmanın güçlü ihmal edilebilirliği varsayımı ile birlikte ağırlıklandırılmış veriler İşleme tabi olma atamasının güçlü ihmal edilebilirliği varsayımı ile birlikte ağırlıklandırılmış veriler, ortak değişkenlere göre işleme tabi olan ve işleme tabi olmayan grupların eşdeğer olduğu sahte bir evreni temsil eder. Aşağıdaki denklem, ağırlıkların nasıl hesaplandığını göstermektedir:

$$w_i = T_i \Pr(\mathrm{T}=1)\frac{n_s}{n_s^t} + (1-T_i)\Pr(\mathrm{T}=0)\frac{n_s}{n_s^c} \qquad (3)$$

Burada sırasıyla $\Pr(\mathrm{T}=1)$ ve $\Pr(\mathrm{T}=0)$ bütün örneklemde işleme tabi olan ve işleme tabi olmayanların oranlarıdır, diğer terimler daha önce tanımlandığı gibidir.

Optimal Tam Eğilim Puanı Eşleştirmesi Üzerinden Marjinal Ortalama Ağırlıklandırması (MMWFM)

MMWS'nin doğrudan bir uzantısı olarak MMWFM'yi öneriyoruz. Bu metot, Hong (2012)'nin yaklaşımını kullanarak gözlemleri ağırlıklandırmayı, ancak PSS metodu yerine OFM metoduna dayalı tabakalar elde etmeyi gerektirir. Tabakalar OFM ile şekillendirildikten sonra, ağırlıklar Denklem (2) yerine Denklem (3) kullanılarak hesaplanır.

PSS ile MMWS'yi ve OFM ile MMWFM'yi karşılaştırdığımızda, orijinal ağırlıkları veya marjinal ortalama ağırlıkları kullanmanın aynı ATE'yi ve ATE'nin standart hatasını kestirdiğini. Bunun nedeni, PSS için Denklem 2'deki ağırlık ile MMWTS-PSS için Denklem 3'teki ağırlık arasında yalnızca tek bir terimin farklılık göstermesidir: Denklem 2'deki 0,5 sabiti, seçim yanlılığı olmaması durumunda örneklemin yarısının her koşulu alacağını belirtmektedir. Öte yandan, Denklem 3'teki marjinal ortalama ağırlıklar, işleme tabi olan $\Pr(\mathrm{T}=1)$ ve işleme tabi olmayanların $\Pr(\mathrm{T}=0)$ marjinal oranlarını içermektedir. Hem PSS ağırlıklarının hem de MMWS ağırlıklarının tabakalar arasında tam olarak aynı ilişkiye sahip olduğunu ve aynı durumun MMWFM ile karşılaştırıldığında OFM için de geçerli olduğunu gördük. Bu nedenle, MMWS ve MMWFM daha ileri analizlerden

hariç tutulmuştur ve PSS ve OFM'den elde edilen sonuçlar sırasıyla MMWS ve MMWFM'ye genellenebilir. Öte yandan, eğilim puanı koşullandırması, tabakalar arasındaki eğilim puanı farklılıkları nedeniyle PSS ve MMWS ile OFM ve MMWFM arasında bir fark yaratmaktadır.

Kovaryans Düzeltmesi (CA)

Eğilim puanı kullanarak kovaryans düzeltmesine (CA) duyulan sezgisel ihtiyaç, çifte sağlamlık özelliğinden kaynaklanmaktadır: Eğilim puanı kestirim modeline dahil edilen gözlemlenen ortak değişkenler, ATE'nin parametrik veya parametrik olmayan kestiricilerine de eklenebilir (Robins ve Rotnitzky, 2001; Funk, vd., 2011); hem işleme tabi olma hem de çıktı modellerinde aynı ortak değişkenlerin kontrol edilmesi, eğilim puanı modelinin veya çıktı modelinin (ancak her ikisinin aynı anda değil) yanlış belirlenmesine karşı koruma sağlar. Eğilim puanını kullanan CA, çıktı regresyon modeline işleme tabi olma göstergesini ve eğilim puanını dâhil ederek tıbbi araştırmalarda yaygın olarak kullanılmaktadır. (Weitzen vd., 2004). What Works Clearninghouse standartlarında (U.S. Department of Education vd., 2013) yarı deneysel eğitim çalışmaları için katılımcıların başlangıçtaki ortak değişkenlerin değerlerine göre eşleştirilmesine ek olarak CA önerilmektedir. CA'nın herhangi bir eğilim puanı metodu uygulamadan tek başına eğilim puanını kullanması önerilmemektedir çünkü bu, eğilim puanının çıktı doğrusal olarak ilişkili olduğu anlamına gelir ve işleme tabi olma göstergesinin kestirim katsayısı ATE olarak yorumlanamaz (Schafer & Kang, 2008). Ayrıca, Austin ve diğerleri (2007) eğilim puanı ile birlikte sadece CA kullanılmasının işlem etkisini yanlı kestirilmesine neden olabileceği sonucuna varmıştır. Eğilim puanı ile birlikte CA'yı, çalışma boyunca incelediğimiz PS metodlarıyla birlikte kullanıyoruz. Bu kombinasyonda CA'nın, PS metodu tarafından giderilmeyen işleme tabi olmayan ve işleme tabi olmayan gruplar arasındaki küçük artık eş değişken dengesizliğini gidererek kestirimleri iyileştirmesi beklenmektedir (Stuart, 2010). Orijinal eğilim puanı metodlarının IPTW-CA, TIPTW-CA, PSS-CA, MMWS-CA, OFM-CA ve MMWFM-CA gibi kovaryans düzeltme uzantılarını belirtmek için eğilim puanı metodunun kısaltmasına "-CA" ekliyoruz.

Daha önce tartıştığımız gibi, Denklem 2 ve 3 ağırlık denkleminde tek bir terim ile farklılık göstermektedir ve aynı tabaka ve işleme tabi olma koşulundaki bireyler aynı sayı ile ağırlıklandırılacaktır, dolayısıyla tabakalar arasındaki ilişkiler sabit kalacaktır. Dolayısıyla, ağırlıkları hesaplamak için Denklem 2 veya 3 kullanılırsa ATE ve standart hata kestirimleri sabit kalacaktır. Ancak, PSS-CA ve MMWS-CA veya OFM-CA ve MMWFM-CA eşdeğer şekilde davranmaz çünkü bireylerin bir tabakada aynı veya doğrusal olarak dönüştürülmüş eğilim puanına sahip olması gerekmez.

İşlem Etkilerinin Kestirimi

Ağırlıklar Denklem 1, 2 veya 3 ile hesaplandıktan sonra, ATE ağırlıklı ortalamalar arasındaki fark olarak kestirilebilir (Schafer ve Kang, 2008; Lunceford ve Davidian, 2004):

$$\Delta = \frac{\sum_{i=1}^{nt} w_{it} y_{it}}{\sum_{i=1}^{n_t} w_{it}} - \frac{\sum_{i=1}^{n_c} w_{jc} y_{jc}}{\sum_{j=1}^{n_c} w_{jc}} \qquad (4)$$

Burada w_{it} , w_{jc} ile y_{it} , y_{jc} sırasıyla işleme tabi olan ve işleme tabi olmayan katılımcılar için

ağırlıklar ve çıktılardır. Alternatif olarak işlem etkisi, Y_i'nin çıktı olduğu, β_0'nin sabit olduğu, β_1'in ATE kestirimi olduğu ve e_i'nin artık olduğu $Y_i = \beta_0 + \beta_1 T_i + e_i$ ağırlıklı regresyon modeli ile elde edilebilir.

Standart Hata Kestirimi

Ağırlıklı En Küçük Kareler Regresyonu (WLS)

WLS, yukarıda sunulan tüm PS metotlarıyla ATE ve standart hata kestirimlerini elde etmek için kullanılabilir (Schafer & Kang, 2008). WLS ile standart hatalar aşağıdaki formülle kestirilir (Fox, 2008):

$$SE(\hat{B}_1) = \sqrt{\frac{\sum_{i=1}^{n}\left(e_i / w_i\right)^2}{n}} \tag{5}$$

Taylor Serisi Doğrusallaştırma (TSL)

PS metotları ile ATE kestirimlerinin standart hatalarını elde etmek için Taylor Serisi Doğrusallaştırma, Jackknife ve Bootstrapping gibi başka metotlar da kullanılabilir (Rodgers, 1999). Bu metotlar sonlu örneklem araştırmalarında yaygın olarak kullanılmaktadır ancak PS metotları ile kapsamlı olarak araştırılmamıştır. TSL, kestiriciyi gözlemlerin doğrusal bir fonksiyonu ile yaklaştırarak bir istatistiğin varyansını elde etmek için kullanılabilir (Wolter, 2007). Açıkça $U_i(ATE)$ gözlem *i* ve ATE'nin bir fonksiyonu olsun ve gerçek evren ATE^*'si sıradaki denklemi sağlasın:

$$\sum_{i=1}^{N} U_i(ATE^*) = 0 \tag{6}$$

Daha sonra, karmaşık bir örneklemde $\hat{ATE}$'yi aşağıdaki ağırlıklandırılmış örneklem denkleminin çözümü olarak tanımlayabiliriz:

$$\sum_{i=1}^{N} \hat{U}_i(ATE) = 0 \tag{7}$$

ATE'nin varyansı, delta metodunu uygulanarak aşağıdaki gibi tanımlanabilir (Binder, 1983):

$$\hat{\text{var}}[\hat{ATE}] \approx \left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE}\right)^{-1} \text{cov}\left[\sum_{i=1}^{n} \hat{U}_i(\hat{ATE})\right]\left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE}\right)^{-1} \tag{8}$$

i gözlemi için $\hat{B}_0$'in sabit, $\hat{B}_1$'in eğim (yani kestirilen ATE), w_i'nin ağırlık ve $\hat{\bar{x}}$'nın *x*'in ortalaması olduğu çıktı regresyon denkleminde TSL kullanılarak ATE'nin standart hatası aşağıdaki şekilde tanımlanır (Lohr, 1999):

$$SE(\hat{B}_1) = \sqrt{\frac{\hat{V}(\sum_{i=1}^{n} w_i(y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})}{\left[\sum_{i=1}^{n} w_i x_i^2 - \frac{\left(\sum_{i=1}^{n} w_i x_i\right)^2}{\sum_{i=1}^{n} w_i}\right]}} \quad (9)$$

Jackknife (JK)

JK ve bootstrapping'in her ikisi de orijinal verilerden yeniden örneklemeye dayanmaktadır. Jackknife'ın en yaygın uygulaması, her tekrarda örneklemin bir üyesinin rastgele çıkarıldığı ve ilgilenilen parametrelerin, gözlem çıkarıldıktan sonra yeniden hesaplanan çoğaltılmış ağırlıklar kullanılarak kestirildiği sil-1 JK'dir. Sil-1 jackknife için w_i bir i gözlemi için başlangıç ağırlığı ve n örneklem büyüklüğü olsun. Seçilen PS metoduna bağlı olarak, w_i Denklem 1'den elde edilen IPTW ağırlığı veya kesilmiş ağırlıklar olabilir:

$$\boldsymbol{w_{ik}} = \begin{cases} \mathbf{0} & \textbf{Eğer gözlem i tekrar k'da silinmişse} \\ \frac{\boldsymbol{n}}{\boldsymbol{n-1}}\boldsymbol{w_i} & \textbf{Eğer gözlem i tekrar k'da silinmemişs} \end{cases} \quad (10)$$

Her tekrarda, ilgilenilen parametre olan ATE ($\hat{B}_{1k}$) tekrar hesaplanacaktır. Standart hata aşağıdaki gibi olacaktır (Lohr, 1999):

$$SE(\hat{B}_1) = \sqrt{\frac{n-1}{n}\sum_{k=1}^{k}(\hat{B}_{1k} - \hat{B}_1)^2} \quad (11)$$

Eğilim Puanı Metotlarının Karşılaştırılması

Gu ve Rosenbaum (1993) ve Cepeda ve diğerleri (2003) optimal eşleştirmenin, PS eşleştirmesi için en yaygın kullanılan algoritma olan açgözlü algoritma (İng. Greedy Algorithm) ile eşleştirmeden sürekli olarak daha iyi performans gösterdiğini bulmuştur. Austin (2009a), belirli bir kumpas dâhilindeki (İng. Caliper) eğilim puanı ve IPTW metotlarıyla eşleştirmenin, gruplar arasındaki sistematik farklılıkları PSS ve kovaryans düzeltmesinden daha fazla ortadan kaldırdığını bulmuştur. Austin (2010a) ayrıca IPTW-CA metodunun yanlılık, varyans kestirimi, güven aralıklarının kapsamı, ortalama karesel hata ve Tip I hata oranları açısından PSS, eğilim puanı eşleştirme, IPTW ve kovaryans düzeltme metotlarından daha iyi çalıştığını bulmuştur. Ancak, Austin bu metotları yalnızca ikili çıktı koşulu altında karşılaştırmıştır. Ayrıca, Austin (2009b) eğilim puanı eşleştirmesi için standart hata kestirim metotlarını değerlendirmiş ve verilerin eşleştirilmiş doğasını dikkate alan metotların daha küçük standart hata yanlılığı ve nominal Tip I hata oranına daha yakın gerçek Tip I hata oranları ile sonuçlandığını bulmuştur. Leite ve diğerleri (2019), birden fazla işlem seçeneği olduğu durumlarda işlem etkilerinin kestirimi için IPTW, OFM ve MMWS'yi karşılaştırmış ve IPTW'nin en düşük düzeyde yanlılık ürettiğini, bunu OFM ve MMWS'nin izlediğini bulmuştur. Ayrıca, işlem etkilerinin standart hatalarının IPTW ile yansız olduğunu, ancak bunları kestirmek için TSL, JK veya bootstrapping kullanılıp kullanılmadığına bakılmaksızın OFM ve MMWS ile fazla kestirildiğini bulmuşlardır. Yukarıda özetlendiği üzere, literatürde çok sayıda PS ve SE kestirim metodunun göreli performansını karşılaştıran çeşitli çalışmalar yapılmıştır. Ancak, her bir spesifik araştırma sadece birkaç metodu içermektedir. IPTW, PSS, MMWS ve OFM ile karşılaştırıldığında TIPTW ve MMWFM'nin performansı hakkında çok az şey bilinmektedir. Bu PS metotlarının

özellikle -CA uzantılarının SE kestirim metotlarıyla birlikte kullanıldığı çok az araştırma vardır. Bu çalışma literatürdeki boşluğu üç şekilde dolduracaktır. İlk olarak, yaygın olarak kullanılmayan iki PS metodunun (yani TIPTW ve MMWFM) performansı, daha yaygın olarak kullanılan PS metotlarıyla (IPTW, PSS, MMWS, OFM) karşılaştırılacaktır. İkinci olarak, PS metotlarının -CA uzantılarının performansı değerlendirilecektir. Son olarak, PS metotlarıyla ile birlikte çeşitli SE kestirim metotlarının performansı büyük ölçekte değerlendirilecektir. Bu PS metotları arasındaki karşılaştırmaların azlığı göz önüne alındığında, mevcut çalışmada aşağıdaki araştırma soruları ele alınacaktır:

1. Hangi eğilim puanı metodu (IPTW, TIPTW, PSS, MMWS, OFM ve MMWFM) farklı örneklem büyüklüğü ve işleme tabi olan oranına sahip koşullar altında ATE'nin yansız kestirimi açısından en iyi performansı gösterir?

2. Kestirilen eğilim puanları kullanılarak yapılan kovaryans düzeltmesi, ATE kestirimi açısından eğilim puanı metotlarının (IPTW-CA, TIPTW-CA, PSS-CA, MMWS-CA, OFM- CA ve MMWFM-CA) performansını iyileştiriyor mu?

3. Hangi standart hata kestirim metodu (WLS, TSL ve JK) farklı eğilim puanı metotlarıyla birlikte kullanıldığında en doğru standart hataları üretir?

4. Hangi eğilim puanı metodu ve standart hata kestiricisi kombinasyonları istatistiksel olarak anlamlı ATE'yi tespit etmek için en fazla istatistiksel gücü sağlar?

Yöntem

Veri Simülasyonu

Araştırma sorularını yanıtlamak için R programı kullanılarak bir Monte Carlo simülasyon çalışması gerçekleştirilmiştir (R Development Core Team, 2011). Verileri simüle etmek üzere gerçekçi evren parametreleri elde etmek için 2007-2008 Okul Suç ve Güvenlik Anketi (SSOCS, İng. School Survey on Crime and Safety) anket sonuçlarından kestirimler alınmıştır (Ulusal Eğitim İstatistikleri Merkezi, 2010). İşleme tabi olma değişkeni okul dışında bir disiplin planının mevcut olup olmadığı, çıktı değişkeni ise belirtilen suçlara karışan toplam öğrenci sayısıdır. Ortak değişkenler ise başka bir okuldan transfer edilen öğrenci sayısı, tipik sınıf değişikliği sayısı, standart testlerde *15.* yüzdelik dilimin altında kalan öğrenci yüzdesi ve özel okullara yapılan toplam transfer sayısıdır.

R'deki *MASS* paketini (Venables & Ripley, 2002) kullanarak simülasyon çalışması için çok değişkenli-normal dağılımlı ortak değişkenler oluşturuldu. Veri simülasyonunun ilk adımı olarak normal dağılan $X_{1i}, X_{2i}, X_{3i}, X_{4i}$ ortak değişkenleri, evren ortalamaları 0 ve kovaryans matrisi aşağıda sunulan evren kovaryans matrisine (SSOCS veri setinden elde edilmiştir) eşit olacak şekilde oluşturulmuştur.

$$\begin{bmatrix} 1.00 & .145 & -.004 & .125 \\ .145 & 1.00 & .001 & .467 \\ -.004 & .001 & 1.00 & .061 \\ .125 & .467 & .061 & 1.00 \end{bmatrix} \quad (12)$$

İkinci adımda, çıktı regresyon artıklarını simüle ettik. Artıklar, ortalaması 0 ve standart sapması *166.*278 olan normal bir dağılımdan simüle edilmiştir. Artıkların evren standart sapması, çıktı regresyonu için evren R^2=*.211* olacak şekilde tanımlanmıştır. Ortak değişkenler ve çıktı artıkları simüle edildikten sonra, aşağıdaki denklemlere dayanarak örneklemdeki tüm bireyler için potansiyel işleme tabi olmama çıktıları Y_C ve potansiyel işleme tabi olma çıktıları Y_T elde edilmiştir.

$$\begin{aligned} Y_{Ci} &= \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i \\ Y_{Ti} &= Y_{Ci} + ATE \end{aligned} \qquad (13)$$

β_0 , β_1 , β_2 , β_3 ve β_4 katsayılarının evren değerleri *0*, *16.221*, *58.642*, *15.704* ve *33.601'dir*. ATE'nin evren değeri *20'dir; bu da* Cohen'in etki büyüklüğünün *.085* olduğunu ve bunun küçük bir etki olduğunu göstermektedir. Bir sonraki adım, simüle edilen örneklerdeki hangi bireylerin işleme tabi olacağını maruz kaldığını belirlemekti. İşleme tabi olma ataması için evren modeli şöyledir:

$$\text{logit}(\text{P}(T_i = 1 | X_{1i}, X_{2i}, X_{3i}, X_{4i})) = \log(rt/(1-rt)) + \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i} \qquad (14)$$

Burada γ_0, $\gamma_1, \gamma_2, \gamma_3$ ve γ_4 evren değerleri *0*, *16.221*, *58.642*, *15.704* ve *33.601'dir* ve *rt*, işleme tabi olma oranıdır. Seçim yanlılığının gücü McKelvey & Zavonia sahte R^2(McKelvey & Zavoina, 1975) temel alınarak tanımlanmıştır ve simule edilen veriler için evren değeri *.028*'dir. İşleme tabi olan grupta olma olasılığını $rt/(1-rt)$ ile işleme tabi olma ataması modeline dâhil ettik ve böylece işleme tabi olma oranını kontrol edebildik.

Sadece örneklem büyüklüğünü ve işleme tabi olan bireylerin oranını (yani toplam örneklem büyüklüğü ve işleme tabi olanlar grubunun toplam örneklem büyüklüğüne göre büyüklüğünü) manipüle ettik çünkü bu iki özellik nicel bir araştırma deseni açısından birçok araştırmacıya yol gösterir. Gu ve Rosenbaum (1993), Freedman ve Berk (2008) ve Austin (2009a) tarafından PS metotlarını karşılaştırmak için yapılan Monte Carlo simülasyon çalışmalarında tek örneklem büyüklüğü olarak 1000 kullanılmıştır. Örneklem büyüklüğünü manipüle ederek, ATE'yi test etme gücü açısından PS metotları arasında fark olup olmadığını belirleyebildik. Örneklem büyüklükleri 500, 1000 ve 2000'e eşit olan verileri simüle ettik.

İşleme tabi olan örneklem oranının *1/10*, *1/7*, *1/4*, *1/3* ve *1/2* olarak belirlendiği veriler ürettik. Bu koşullar, Gu ve Rosenbaum'un (1993) sadece *1/7*, *1/4* ve *1/3* oranlarını inceleyen çalışmasının bir uzantısıdır.

Ortak değişkenlerin sayısını değiştirmedik çünkü Gu ve Rosenbaum (1993) işleme tabi olma mekanizması tam olarak modellendiği sürece, ortak değişkenlerin sayısının, ortak değişkenlerin çoklu doğrusallığı ve yakınsama sorunları gibi potansiyel sorunlar haricinde, eğilim puanı metodunun performansını etkilemediği sonucuna varmıştır. Bu simülasyon çalışmasında hem çıktıyla hem de işleme tabi olma atamasıyla ilişkili olan dört sürekli ortak değişken kullandık. Verileri dört ortak değişkene dayalı olarak simüle ettiğimiz ve dört ortak değişkeni de kullanarak eğilim puanlarını kestirdiğimiz için, işleme tabi olma atamasının tamamen modellendiği varsayımı çalışmadaki tüm koşullar için karşılanmıştır. Veri simülasyonunda manipüle edilen koşullar Tablo 1'de özetlenmiştir.

Tablo 1

Veri Simülasyonunda Manipüle Edilen Koşulların Özeti

Koşul	Seviyeler
Örneklem büyüklüğü	*500*, *1000* ve *2000*
İşleme tabi olan bireylerin oranı	*1/10*, *1/7*, *1/4*, *1/3* ve *1/2*

ATE ve Standart Hata Kestirimi

Koşul başına *1000 veri kümesi* simüle edilmiş ve simüle edilen her bir veri kümesi ayrı ayrı dört adımda analiz edilmiştir: 1) Lojistik regresyon kullanarak her birey için PS'yi kestirmek; 2) Ortak destek alanını hesaplamak; 3) Ağırlıklandırılmamış bir regresyon modeli ile seçim yanlılığını da kapsayan ATE ve standart hatayı kestirerek herhangi bir PS metodu uygulanmadığında elde edilecek temel sonucu belirlemek; 4) Eğilim puanı metotlarını uygulamak. Denklem 16'yı kullanarak IPTW ile ATE'yi kestirilmiştir. IPTW'nin uygulanmasından sonra, TIPTW ağırlıklarını oluşturmak için IPTW'nin *99.* yüzdelik diliminden büyük olan ağırlıkları *99.* yüzdelik dilimi ile değiştirilmiştir. PSS'yi uygulamak için, R'deki *MatchIt* paketi (Ho, Imai, King, Stuart, 2007) kullanılarak PS'lerdeki benzerliğe dayalı olarak işleme tabi olan ve işleme tabi olmayan bireyler beş tabakaya ayrılmıştır. 5 tabaka kullanılmıştır çünkü sosyal bilimlerdeki PS metotları uygulamalarında en yaygın kullanılan tabaka sayısı 5'dir (Thoemmes & Kim, 2011). MMWS'yi uygulamak için ağırlıklar Denklem 3 kullanılarak hesaplanmıştır. OFM metodu için, R'de *optmach* paketinde (Hansen & Klopfer, 2006) uygulanan OFM algoritması kullanılarak işleme tabi olan ve işleme tabi olmayan bireylerin benzerliğe dayalı olarak veri tanımlı sayıda tabakaya ayrılması sağlanmıştır. MMWFM'yi uygulamak için OFM ile elde edilen tabakalar kullanılmıştır, ancak ağırlıklar Denklem 11 kullanılarak hesaplanmıştır.

ATE kestirimi β_1'in ATE olduğu $Y_i = \beta_0 + \beta_1 T_i + e_i$ ağırlıklandırılmış regresyon denklemi ile gerçekleştirilmiştir. Kovaryans düzeltmesini eklemek için ATE, $Y_i = \beta_0 + \beta_1 T_i + \beta_2 PS + e_i$ ağırlıklandırılmış regresyon denklemi ile gerçekleştirilmiştir. Burada PS, bir ortak değişken olarak modele dahil edilen eğilim puanıdır. ATE kestirimlerinin WLS standart hataları R'deki *lm* fonksiyonu kullanılarak elde edilirken, TLS standart hataları ve sil-1 JK standart hataları R'deki *survey* paketinin (Lumley, 2011) svyglm fonksiyonu ile elde edilmiştir.

Analiz

Eğilim puanlarının ortak destek alanını ölçmek için dağılımların örtüşme oranını yansıtacak şekilde Cohen'in (1988) U_1 fonksioynuna benzer bir örtüşme ölçütü kullanılmıştır. A ve C, eğilim puanlarının logit dönüşümünün alt ve üst uçlarındaki örtüşmeme alanı, B'de örtüşme alanı olsun. O zaman:

$$U_1 = \frac{A + C}{A + B + C} \tag{15}$$

Burada U_1'deki artışlar, ortak destek alanındaki azalmalara karşılık gelmektedir. Simüle edilen koşullarda ortalama U_1 *.001* ile *.604* arasında değişmekte olup ortalama ve standart sapma sırasıyla *.190* ve *.091*'dir. Örneklem büyüklüğü arttıkça veya işleme tabi olanların oranı yükseldikçe örtüşme de artmıştır.

PS metotları ATE kestirimlerinin göreli yanlılığı ve yanlılık azaltma yüzdesi açısından

karşılaştırılmıştır. ATE'nin göreli yanlılığı $\bar{\hat{\theta}}$'nin bir koşul için bütün tekrarlarda ATE kestirimlerinin ortalaması ve θ'nın evren ATE'si olduğu durumda $B(\hat{\theta}) = (\bar{\hat{\theta}} - \theta) / \theta$ formülü ile hesaplanmıştır. Eğer hesaplanan $B(\hat{\theta})$ mutlak değerde .05'den büyükse yanlılık kabul edilemez olarak değerlendirilmiştir(Hoogland & Boomsma, 1998). Ayrıca yanlılık azaltma yüzdesi(PBR, İng. Proportion Bias Reduction) aşağıdaki formül ile değerlendirilmiştir:

$$\boldsymbol{PBR(\hat{\theta}) = \frac{B(\hat{\theta})_{temel} - B(\hat{\theta})_{metot}}{B(\hat{\theta})_{temel}} x100} \quad (16)$$

Burada $B(\hat{\theta})_{metot}$ belirli bir PS metodu kullanmanın ortalama göreli yanlılığı ve $B(\hat{\theta})_{temel}$ herhangi bir PS metodu kullanmadan başlangıçtaki yanlılıktır (Cochran ve Rubin, 1973; Steiner vd., 2010). Cochran ve Rubin (1973) tarafından sunulan sınır değeriyle tutarlı olarak, bir PS metodunun başarılı sayılabilmesi için başlangıçtaki temel yanlılığın en az %90'ını gidermesi beklenmiştir.

Ayrıca standart hata kestirim metotlarını standart hataların göreli yanlılığı açısından da karşılaştırılmıştır. Standart hataların göreli yanlılığı $B(S_{\hat{\theta}}) = [\bar{S}_{\hat{\theta}} - SD(\hat{\theta})] / SD(\hat{\theta})$ olup, $\bar{S}_{\hat{\theta}}$ kestirilen ATE'lerin standart hatalarının ortalaması ve $SD(\hat{\theta})$ ampirik standart hatadır. Ampirik standart hata, herhangi bir koşulda bir PS metodu ile kestirilen ATE'lerin standart sapmasıdır. Hesaplanan $B(S_{\hat{\theta}})$ mutlak değerde *.10*'dan büyükse, yanlılık kabul edilemez olarak değerlendirilmiştir. Yansız ATE ve standart hata kestirimleri sağlayan metotlar için, her koşulda $\alpha = .05$ seviyesinde istatistiksel olarak anlamlı bulunan ATE oranı hesaplanarak söz konusu metotların istatistiksel olarak anlamlı ATE'yi tespit etme istatistiksel gücü hesaplanmıştır.

Bulgular

Eğilim Puanı Metotlarının Karşılaştırılması

Altı eğilim puanı metodunun her biri için göreli yanlılık ve yanlılık azaltma yüzdesi hesaplanmıştır. Kestirilen ATE'nin göreli yanlılığının ve yanlılık azaltma yüzdesinin sonuç, eğilim puanı metotlarının gruplar içi faktör, örneklem büyüklüğü ve işleme tabi olan katılımcıların oranının ise gruplar arası faktörler olduğu iki karmaşık ANOVA yürütülmüştür. Tüm olası etkileşimler de modele dâhil edilmiştir. Manipüle edilmiş koşulları karşılaştırmak için etki büyüklüğünün bir ölçüsü olarak genelleştirilmiş eta kare (η^2) kullanılmıştır. Eğilim puanı metodu faktörünün ana etkisi $\eta^2 = .023$ ve diğer tüm etki büyüklüklerinin *.01'den* küçük olması, PS metodunun göreceli yanlılık ve yanlılık azaltma yüzdesi performansının, herhangi bir PS metodu için simüle edilmiş örneklem büyüklüklerinde ve işleme tabi olan oranlarında çok benzer olduğunu göstermektedir. Bu nedenle, göreli yanlılık sonuçlarını örneklem büyüklüğü ve işleme tabi olanların oranına göre daralttık ve Tablo 2'de göreli yanlılığı ve yanlılık azaltma yüzdesini gösterdik.

Tablo 2

PS Metoduna Göre ATE Kestirimlerinin Göreli Yanlılığı ve Yanlılık Azaltma Yüzdeleri

	Kovaryans düzeltmesiz çıktı modeli		Kovaryans düzeltmeli çıktı modeli	
ATE Kestirim Metodu	Göreli Yanlılık	Yanlılık Azaltma Yüzdesi	Göreli Yanlılık	Yanlılık Azaltma Yüzdesi
IPTW	0.001	99.97%	-0.001	100.16%
TIPTW	0.086	92.54%	0.017	98.59%
PSS	0.116	89.96%	-0.002	100.24%
MMWS	0.116	89.96%	-0.004	100.37%
OFM	0.009	99.30%	0.002	99.90%
MMWFM	0.009	99.30%	0.002	99.90%

Not. Tüm Tekrarlardaki temel yanlılık 1.153'tür.

TIPTW, PSS ve MMWS metotlarının sırasıyla *0,086*, 0,*116* ve 0,*116* göreli yanlılık ile *.05* göreli yanlılık kriterini aştığı bulunmuştur. Ayrıca, PSS ve MMWS *%90'a* yakın yanlılığı ortadan kaldırırken, IPTW, OFM ve MMWFM *%99*'a yakın yanlılığı ortadan kaldırmıştır. Kesme işlemi yanlılığın giderilmesini azaltmış, TIPTW yanlılığın yaklaşık *%92*'sini gidermiştir. Dolayısıyla, IPTW, TIPTW, OFM ve MMWFM ile elde edilen hem göreli yanlılık hem de yanlılık azaltma yüzdesi kabul edilebilir seviyelerdedir. Ayrıca, eğilim puanı ile kovaryans düzeltmesinin, incelenen tüm PS metotlarında ATE kestirimlerinin doğruluğunu artırdığını gözlemlenmiştir; bu nedenle, kovaryans düzeltmesi ile birlikte tüm eğilim puanı metotları, ATE kestirimlerinde kabul edilebilir düzeyde göreli yanlılık ve yanlılık yüzdesi azaltımı sağlamıştır.

Standart Hata Kestirim Metotlarının Karşılaştırılması

Standart hata kestirimlerini incelerken, bazı standart hata kestirimlerinin yansız olduğunu, bazılarının ise olmadığını gözlemledik. Standart hata kestirimlerinin göreli yanlılığını etkileyen faktörleri araştırmak için, sonucun standart hata kestirimlerinin göreli yanlılığı olduğu, gruplar arası faktörlerin örneklem büyüklüğü ve işleme tabi olan oranı olduğu ve grup içi faktörlerin PS metodu, standart hata kestirim metodu ve kovaryans düzeltmesi olduğu bir karmaşık ANOVA gerçekleştirdik. Koşulları η^2etki büyüklüğü ölçüsüne göre karşılaştırdık (Olejnik & Algina, 2003). Tablo 3, standart hata kestirimlerinin yanlılığını etkileyen faktörler için özet ANOVA tablosunu göstermektedir.

Tablo 3

Standart Hata Kestirimlerinin Göreli Yanlılığı için Etki Büyüklüğü Özeti

Kaynak	η^2
Gruplar Arası Etkiler	
İşleme tabi olan oranı	0.109
Grup İçi Etkiler	
PS metodu	0.035
PS metodu * İşleme tabi olan oranı	0.017
Kovaryans düzeltmesi	0.047
Standard hata kestirim metodu	0.327
Standard hata kestirim metodu * İşleme tabi olan oranı	0.122
PS metodu * Standard hata kestirim metodu	0.062
PS metodu* Standard hata kestirim metodu * İşleme tabi olan oranı	0.030

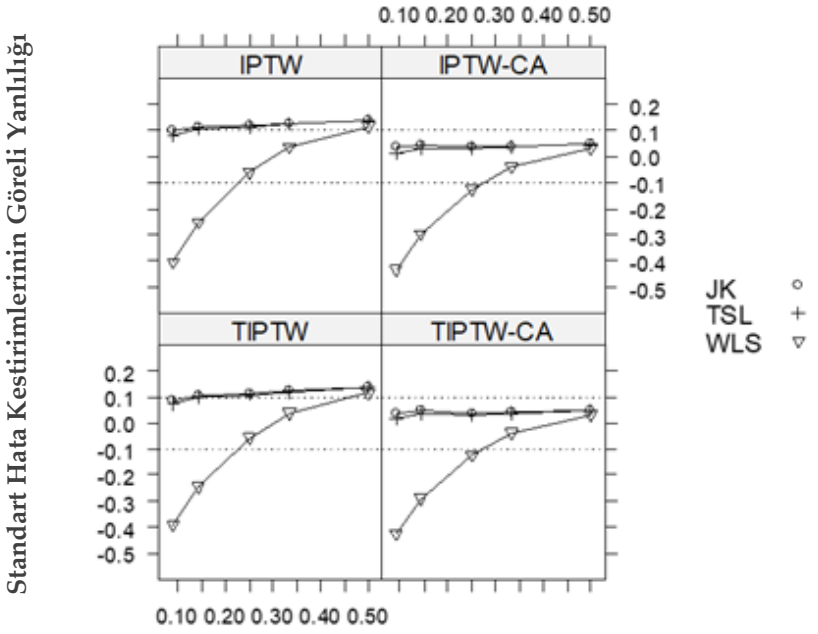
Not. .01'den daha küçük etki büyüklüklerine sahip kaynaklar raporlanmamıştır.

Sonuçlar, standart hata kestirim metodunun standart hata kestirimlerinin göreli yanlılığı üzerinde

$\eta^{2=}.327$ ile büyük bir etkiye sahip olduğunu göstermiştir. Bunula birlikte, standart hata kestirim metodu ile işleme tabi olanların iki yönlü orta düzeyde bir etkileşim vardır ve η^2=.122'dir. İşleme tabi olan oranının etki büyüklüğü η^2= .109'dur. Bu etkilere ek olarak çeşitli küçük etkiler de gözlemlenmiştir: PS metodu ve standart hata kestirim metoduna ilişkin iki yönlü etkileşimin etki büyüklüğü η^2= .062, PS metodu, standart hata kestirim metodu ve işleme tabi olan oranının üç yönlü etkileşiminin etki büyüklüğü η^2=.030, PS metodu ve kovaryans düzeltmesinin etki büyüklüğü sırasıyla η^2=.035 ve η^2=.047'dir. Örneklem büyüklüğünü içeren ana etki ve etkileşimlerin ihmal edilebilir olduğu göz önüne alındığında, standart hataların göreli yanlılığı örneklem büyüklüğü üzerinden daraltılmıştır.

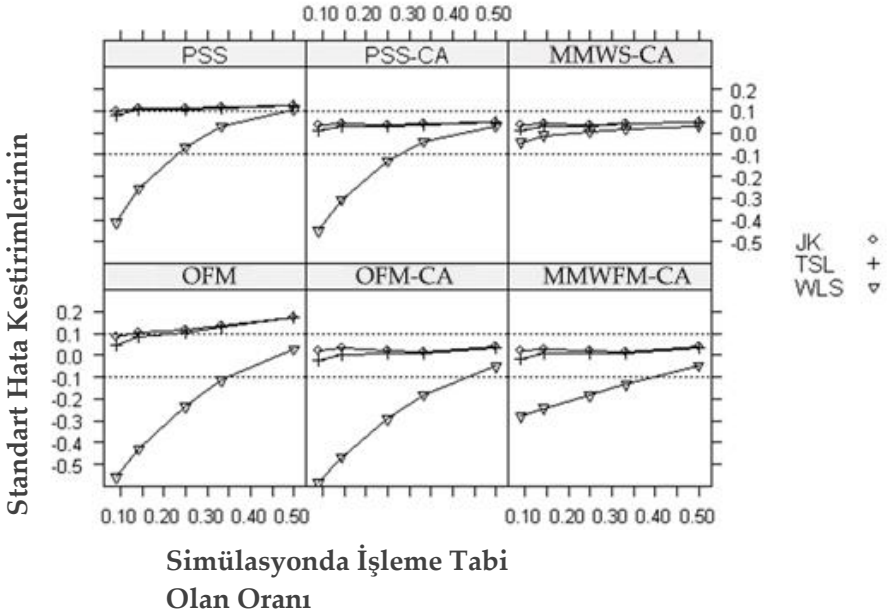
Şekil 1

IPTW ve TIPTW ile Simulasyonda İşleme Tabi Olan Oranına Göre Standart Hata Kestirimlerinin Göreli Yanlılığı



Şekil 2

IPTW ve TIPTW ile Simulasyonda İşleme Tabi Olan Oranına Göre Standart Hata Kestirimlerinin Göreli Yanlılığı



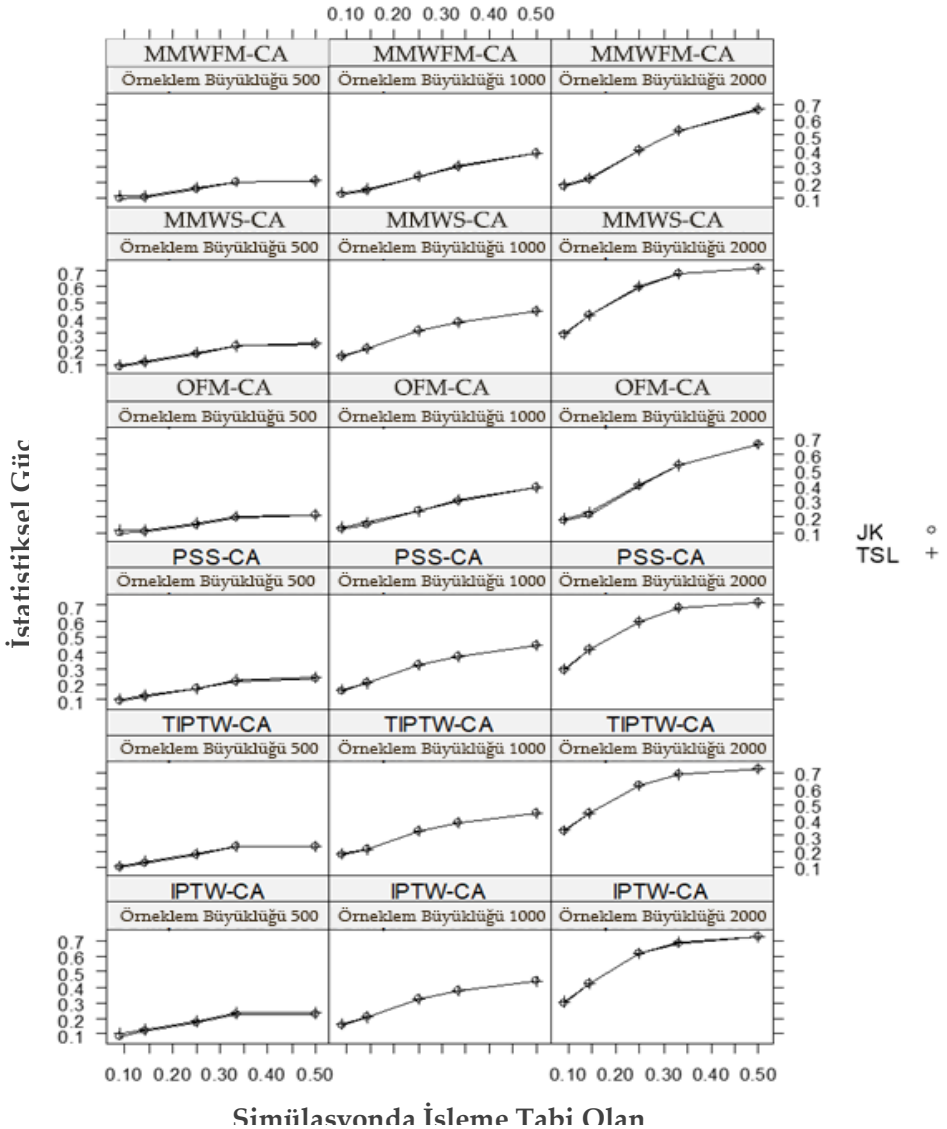
Standart hata kestirimlerinin göreceli yanlılıkları Şekil 1 ve Şekil 2'de özetlenmiştir. TSL ve JK'nın simüle edilen tüm koşullarda benzer standart hatalar sağladığı görülmüştür. İşleme tabi olan oranına göre herhangi bir PS metodu için TSL ve JK metotları ile elde edilen standart hataların göreli yanlılığı *-.024* ile *.177* arasında değişmekte olup ortalaması *.058*'dir. Standart hata kestirimlerinin bu göreli sapmalarının çoğu *(-.1,.1)* aralığını aşmamıştır, ancak işleme tabi olanların oranı *0.1*'den *0.5*'e yükseldikçe, TSL ve JK'nın standart hataları marjinal ila orta derecede arasında fazla kestirme eğiliminde olduğu görülmüştür. Bununla birlikte, her bir PS metodu için çıktı regresyon modeline kovaryans düzeltmesi eklenmesi, standart hata tahminlerinin yanlılığını kabul edilebilir seviyelere indirmiştir. Genel olarak, TSL, JK'yy göre biraz daha az yanlı standart hatalarla sonuçlanmıştır. WLS standart hataları *-.585* ile *.117* arasında değişmekte olup ortalaması *-.146*'dır. WLS, simüle edilen işleme tabi olanların oranı *.25* olduğunda MMWS-CA ve bazı PS metotları haricinde en düşük üç işleme tabi olma oranında standart hataları olduğundan düşük kestirmiştir. MMWS-CA metodu, üç standart hata kestirim metoduyla birlikte simüle edilen her koşulda kabul edilebilir standart hatalar sağlamıştır. İşleme tabi olanların oranı *0.1'*den *0.5'*e yükseldikçe, kestirilen standart hataların göreli yanlılığında açısından ikinci dereceden bir iyileşme görülür. WLS kestirimi ile birlikte çıktı modelinde kovaryans düzenlemesi ile ağırlıklandırmaya dayalı PS metotları, kovaryans düzenlemesi olmayan ağırlıklandırmaya dayalı PS metotlarına kıyasla marjinal olarak daha yanlı standart hatalar kestirmiştir. Bu ilişki, eşleştirmeye dayalı PS metotlarında tersine dönmüş ve MMWT metotları, MMWT olmayan metotlara kıyasla daha az yanlı tahminler sağlamıştır. İşleme tabi olanların oranı 0.5 olduğunda, WLS, simüle edilen tüm koşullarda TSL ve JK'den daha az yanlı standart hatalar sağlamıştır.

Bazı PS metotları ve standart hata kestirim metotları kombinasyonlarının simüle edilen çeşitli koşullar için yanlı standart hatalar sağladığını göz önünde bulundurarak, güç incelemesini TLI ve JK ile standart hata tahmini ile birleştirilmiş kovaryans düzeltmeli çıktı modelli PS metotlarıyla sınırlandırılmıştır çünkü bunlar başlangıçtaki yanlılığı ortadan kaldırmış ve simüle edilen tüm

koşullarda yansız standart hatalar sağlamıştır.

Şekil 3

İşleme Tabi Olan Oranı, Örneklem Büyüklüğü ve Eğilim Puanı Metodunun Bir Fonksiyonu Olarak ATE'yi Tespit Etme Gücü



Şekil 3'te sunulan sonuçlara dayanarak, örneklem büyüklüğü ve işleme tabi olanların oranının istatistiksel gücü etkileyen en önemli iki faktör olduğu gözlemlenmiştir. İşleme tabi olanların oranı olmayanların oranına yaklaştıkça veya örneklem büyüklüğü arttıkça, TSL ve JK standart hata kestirimi ile eşleştirilmiş kovaryans düzeltmeli PS metotlarının tümü artan güç sağlamıştır. Ayrıca, işleme tabi olanların oranı olmayanların oranına yaklaştıkça güçteki artışın daha yüksek örneklem büyüklükleri için daha fazla olduğu gözlemlenmiştir. TSL ve JK'nın istatistiksel gücünü çıktı modelinde kovaryans düzeltmeli altı PS metotlarıyla birlikte karşılaştırmak için, istatistiksel gücü işleme tabi olanların oranı ve örneklem büyüklüğü ile daralttık. Farklı PS metotları ile TSL ve JK standart hata kestirim metotlarının güçteki genel farklılıkların TSL lehine *.01*'den az olduğu bulunmuştur. Son olarak, PS metotları arasındaki genel farklılıkları gözlemlemek için işleme tabi olan oranı, örneklem büyüklüğü ve standart hata kestirim metotlarına daraltarak istatistiksel gücü elde ettik. Sonuçlar, MMWFM-CA ve OFM-CA'nın sırasıyla *.266* ve *.267* ortalama istatistiksel güç ile en düşük istatistiksel güce sahip olduğunu göstermiştir. Kalan dört PS metodu, *.339* ile *.349*

arasında değişen genel istatistiksel güç seviyeleri ile daha yüksek istatistiksel güce sahiptir.

Tartışma ve Sonuç

ATE kestirimlerinin uygunluğunu (yani, göreli yanlılık ve yanlılık azaltma yüzdesi), ATE kestirimi için standart hataların uygunluğunu (yani, göreli yanlılık) ve farklı ATE kestirim ve standart hata kestirim metotlarının kombinasyonlarının istatistiksel gücü araştırılmıştır. İlk araştırma sorumuz, farklı PS metotları kullanılarak yapılan ATE kestirimlerinin uygunluğuyla ilgiliydi. IPTW ile başlangıçtaki yanlılığın tamamının giderildiği gözlemlendi. IPTW ağırlıklarının *99.* yüzdelik diliminden büyük olan uç değerlerin kırpılması (yani TIPTW) kabul edilebilir ancak IPTW'den daha kötü performans göstermiştir. Bu bulgu Austin'in (2009a; 2010a) bulgularıyla tutarlıdır ve Freedman ve Berk'in (2008) ve Sturmer ve diğerlerinin (2010) bulgularının tersidir. Hem Freedman ve Berk (2008) hem de Sturmer ve diğerlerinin (2010) çalışmalarında veriler, uç ağırlıkların işleme tabi olma kestirimi üzerinde oldukça etkili olacağı şekilde simüle edilmiştir. Bizim çalışmamızda, Freedman ve Berk (2008) tarafından incelenen kırpma kuralı olan 20'den büyük ağırlıklar, simüle edilen iterasyonların %1'inden azında meydana gelmiştir. Ayrıca, ağırlık dağılımının sadece üst ucu kırpılmışır. TIPTW'nin IPTW'ye kıyasla daha düşük performans göstermesi, işleme tabi olmayanlar grubunda büyük ağırlığa sahip bireylerin, ortak değişken dağılımları açısından işleme tabi olanlar grubuna benzer vakalar olmasından kaynaklanmaktadır. Bu nedenle, büyük ağırlığa sahip vakalar ortak değişken dağılımlarının dengelenmesinde en önemli unsurdur. Bu yeni sonuçlar ve mevcut literatür göz önüne alındığında, uygulamalı araştırmacıların aşırı ağırlıkların oluşup oluşmadığını ve ne kadar aşırı olduklarını belirlemek için ağırlıklarını incelemeleri ve aşırı ağırlıkları kesen ve kesmeyen ATE kestirim metotlarını karşılaştırmaları önerilmektedir.

Gu ve Rosenbaum (1993) ve Cepeda ve diğerlerinin (2003) bulgularına paralel olarak, OFM'nin ATE kestirimlerindeki başlangıç yanlılığının neredeyse tamamını ortadan kaldırdığı gözlemlenmiştir. Cochan (1968) ve Rosenbaum ve Rubin'in (1984) bulgularına benzer şekilde, beş tabakalı PSS başlangıçtaki yanlılığın yaklaşık %90'ını ortadan kaldırmıştır. Austin (2009a; 2010a) ve Lunceford ve Davidian (2004) IPTW'nin PSS'ye kıyasla daha fazla sistematik farklılığı ortadan kaldırdığı sonucuna varmıştır ve bu durum bizim çalışmamızda da gözlemlenmiştir. OFM beklendiği gibi PSS'den daha iyi performans göstermiştir çünkü OFM en az bir işleme tabi olan ve bir işleme tabi olmayan birey içeren maksimum sayıda tabakaya ayırma türüdür. Marjinal ortalama ağırlıkların tanımına dayanarak beklediğimiz gibi, OFM ve MMWFM'den elde edilen sonuçların yanı sıra PSS ve MMWS'den elde edilen sonuçlar da aynıdır. Austin'in (Austin 2010a) bulgularına ve Stuart'ın (2010) iddialarına benzer şekilde, bir eğilim puanı metodu uygulandıktan sonra sonuç modelinde eğilim puanı için kovaryans düzeltmesi yapılması, daha düşük performansa sahip metotlar (yani TIPTW ve PSS) için daha sistematik farklılıkları ortadan kaldırmıştır. Ancak, kovaryans düzeltmesi eklemenin başlangıçtaki yanlılığın neredeyse tamamını ortadan kaldıran metotlar (yani IPTW ve OFM) için hiçbir etkisi olmamıştır. Bu nedenle, OFM-CA ile MMWFM-CA arasında ve PSS ile MMWS-CA arasında çok az fark gözlemlenmiştir. Kovaryans düzeltmesi ile incelenen tüm PS metotları, ATE'nin başlangıçtaki yanlılığının neredeyse tamamını ortadan kaldırmıştır.

İkinci araştırma sorumuz, PS metotlarının farklı kombinasyonlarını kullanarak doğru standart hata kestirimleri elde etmek ve işleme tabi olan oranı ve toplam örneklem büyüklüğünün farklı kombinasyonları için standart hata kestirim metotlarıyla ilgiliydi. ANOVA sonuçlarımız, PS metodunun, standart hata metodunun, kovaryans düzeltmesinin ve işleme tabi olan oranın standart

hata kestirimlerinin doğruluğunu etkilediğini göstermiştir. Öncelikle, MMWS-CA hariç tüm PS metodları için simüle edilen en küçük iki işleme tabi olan oranında WLS'nin standart hataları ciddi şekilde düşük kestirdiği edildiğini gözlemlenmiştir. İşleme tabi olan oranı 0,5'e yaklaştıkça, WLS ile kestirilen standart hataların yanlılığı azalmıştır. TSL ve JK karşılaştırıldığında, standart hataların birbirine benzer olduğu ancak TSL standart hatalarının JK standart hatalarına göre marjinal olarak daha az yanlı olduğu gözlemlenmiştir. Üçüncü olarak, PS metotlarından herhangi biriyle birlikte işleme tabi olan oran 0,5 olduğunda TSL ve JK kullanılarak standart hataların marjinal veya orta derecede fazla kestirildiği edildiğini gözlemlenmiştir. Kestirim aşırı değildir ancak işlem etkisinin anlamlılık testini yapay olarak muhafazakâr hale getirecek kadar büyüktür. Eğilim puanının ATE kestirimi çıktı regresyon modeline bir ortak değişken olarak dahil edildiği durumda, PS metodu, standart hata kestirim metodu ve işleme tabi olan oranı kombinasyonundan bağımsız olarak standart hataların doğruluğunda iyileşme gözlemlenmiştir. İyileşme, WLS standart hatalarının yansız olması için yeterince büyük değildir, ancak tüm TSL ve JK standart hataları, simüle edilen tüm koşullar altında kovaryans düzeltmesi kullanıldığında yansızdır. Bu nedenle, WLS standart hatalarının PS ağırlıkları ile kullanılmaması gerektiği sonucuna varılmıştır. Ancak, R yazılım paketindeki *lm* fonksiyonu gibi bazı yazılımlarda ağırlıklar sağlandığında WLS varsayılandır, bu nedenle uygulamalı araştırmacılar standart hata kestirimlerinde ağırlıkların nasıl ele alındığı konusunda dikkatli olmalıdır. Bu çalışmada standart hataları kestirmek için bootstrapping metotları incelenmemiştir, ancak JK ile benzerliği nedeniyle benzer şekilde performans göstermesini beklemek mantıklıdır.

Bu çalışmanın sonuçları göz önüne alındığında, ATE'nin PS ağırlıkları ve ek eğilim puanı kovaryans düzeltmesi ile kestirilmesi modeller öneriyoruz, çünkü bu sadece ATE kestirimlerinde ek yanlılık giderimi sağlamakla kalmamış, aynı zamanda standart hata kestirimlerini de iyileştirmiştir. Ancak, bu simülasyon çalışmasında kovaryans ayarlaması için kullanılan basit sonuç modelinin bir sınırlaması, eğilim puanının sonuçla doğrusal olarak ilişkili olduğunu varsaymasıdır. Uygulamada, eğilim puanının karesini ve küpünü de modele dahil ederek veya eğilim puanının kukla kodlu katmanlarını dahil ederek bu varsayımdan kaçınmak daha güvenlidir. Ayrıca, artık değişkenliği azaltacağı ve gücü artıracağı için çıktı modeli, işlem öncesi sonuç ölçümü gibi çıktı değişkeniyle güçlü bir şekilde ilişkili olan ortak değişkenleri içerecek şekilde genişletilmelidir. Temel ortak değişkenlerle genişletilmiş bir çıktı modeli aynı zamanda çift sağlamlık özelliğine de sahiptir; burada yanlılık, ortak değişkenlerin hem PS modelinde hem de çıktı modelinde doğru şekilde modellenmesiyle ortadan kaldırılabilir. PS ağırlıklandırma metotlarının bir avantajı, sözde maksimum olabilirlik (İng. Pseudo-Maximum Likelihood) tahmini (Asparouhov, 2006) gibi ağırlıkları içeren bir kestirici mevcut olduğu sürece herhangi bir çıktı modeliyle birlikte kullanılabilmesidir.

PS ağırlıklandırmasına ek olarak sonuç modelindeki eğilim puanının veya ortak değişkenlerin doğrudan ayarlanması, uygulaması kolay ve uygulamalı araştırmacılar için tanıdık olduğu için önerilmektedir. Bununla birlikte, Bang ve Robins (2005), Kang ve Schafer (2007) ve Lunceford ve Davidian'da (2004) açıklanan ATE'nin diğer çift sağlam kestiricileri de mevcuttur. Gelecekteki araştırmalar, bu iki kat sağlam kestiricileri yanlılığın azaltılması ve modelin yanlış belirlenmesine karşı sağlamlık açısından karşılaştırabilir.

Bu çalışma, yalnızca SUTVA'nın karşılandığı koşullara baktığı için de sınırlıdır. SUTVA ihlallerinin nasıl ele alınacağına ilişkin araştırmalar henüz başlangıç aşamasındadır ve Hong & Raudenbush (2006) SUTVA'nın doğrudan ele alındığı nadir bir örnektir. Ayrıca, eğitim araştırmalarındaki veri setleri genellikle kümelenmiş bir yapıya sahiptir, ancak Arpino ve Mealli (2011), Leite ve diğerleri

(2015) ve Thoemmes ve West (2011) gibi çok düzeyli yapıya sahip veriler için eğilim puanı ağırlıklandırma metotları üzerine az sayıda çalışma yapılmıştır, bu nedenle gelecekteki araştırmalar bu alanı genişletmelidir. Gelecekteki araştırmalar, farklı kovaryans matrisine ve farklı ortak değişken dağılımlarına sahip farklı veri yapılarını içerebilir.

Etik Kurul Onayı: Çalışma bir simülasyon çalışması olarak yürütülmüştür. Ayrıca 2020 yılından önceki yüksek lisans/doktora tezlerinden üretilen çalışmalarda etik kurul raporu gerekmemektedir.

Araştırmacıların Katkı Oranı: Çalışma bir yüksek lisans tezinden üretildiği için birinci yazar çalışmanın bütün bölümlerine katkı sağlamış, danışman olan ikinci yazar ise sürece rehberlik ederek katkı sağlayarak yeri geldiğinde bütün bölümlere ekleme ve / veya düzeltme önerilerinde bulunmuştur.

Çatışma Beyanı: Yazarlar potansiyel bir çıkar çatışması olmadığını beyan ederler.

References

Abadie, A., & Imbens, G. W. (2006). Large Sample properties of matching estimators for average treatment effects. *Econometrica, 74*, 235-2667. https://doi.org/10.1111/j.1468-0262.2006.00655.x

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*, 1770-1780. https://doi.org/10.1016/j.csda.2010.11.008

Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. *Communications in Statistics: Theory and Methods, 35*(3), 439-460. https://doi.org/10.1080/03610920500476598

Austin, P. C. (2009a). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making, 29*, 661-677. https://doi.org/10.1177/0272989X09341755

Austin, P. C. (2009b). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics, 5*(1), Art. 13. https://doi.org/10.2202/1557-4679.1146

Austin, P. C. (2010a). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine, 29*, 2137-2148. https://doi.org/10.1002/sim.3854

Austin, P. C. (2010b). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on propensity score. *Practice of Epidemiology, 172*(9), 1092-1097. https://doi.org/10.1093/aje/kwq224

Austin P.C., Grootendorst P., & Anderson G.M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*, 734–753. https://doi.org/10.1002/sim.2580

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics, 61*(4), 962-973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bembom, O., & van der Laan M. J. (2008). *Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators*. U.C. Berkeley Division of Biostatistics Working Paper Series. Paper 230.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156. https://doi.org/10.1093/aje/kwj149

Cepeda M. S., Boston, R., Farrar, J. T., & Strom, B. L., (2003). Optimal matching with a variable number of controls vs. a fixed number of controls for a cohort study: trade-offs. Journal of *Clinical Epidemiology, 56*, 230-237. https://doi.org/10.1016/S0895-4356(02)00583-8

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313. https://doi.org/10.2307/2528036

Cochran, W.G., & Rubin, D. B. (1973). Controlling bias in observational studies: a review. Sankhya: *The Indian Journal of Statistics, Series A 35*(4), 417-446.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press. https://doi.org/10.4324/9780203771587

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*, 187-199. https://doi.org/10.1093/biomet/asn055

Freedman, D. A. & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review, 32*(4), 392-409. https://doi.org/10.1177/0193841X08317586

Funk M. J., Westreich D., Wiesen C., Sturmer T., Brookhart M. A., & Davidian M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761-767. https://doi.org/10.1093/aje/kwq439

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 4*, 405-420. https://doi.org/10.2307/1390693

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: statistical methods and applications*. Sage.

Hansen, B.B., & Klopfer, S.O. (2006) Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*, 609-627. https://doi.org/10.1198/106186006X137047

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-349. https://doi.org/10.1037/a0019623

Heckman, J. J. (1978). Dummy endogenous variables in simultaneous equations system. *Econometrica, 47*, 931-960. https://doi.org/10.2307/1909757

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 65*, 261-294. https://doi.org/10.2307/2971733

Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 82*, 387-394. https://doi.org/10.1097/01.ede.0000135174.63482.43

Ho, D., Imai, K., King, G., & Stuart, A. E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis. 15*(3), 199-236. https://doi.org/10.1093/pan/mpl013

Hong, G. (2012). Marginal mean weighting through stratification: a generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological methods, 17*, 44-60. https://doi.org/10.1037/a0024918

Hong, G., & Hong, Y. (2008). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31*, 54-81. https://doi.org/10.3102/0162373708328259

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*, 901-910. https://doi.org/10.1198/016214506000000447

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research, 26*, 523-539. https://doi.org/10.1177/0049124198026003003

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association, 47*, 663-685. https://doi.org/10.2307/2280784

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 523-539. https://doi.org/10.1214/07-STS227

Leite, W. L. (2016). *Practical propensity score methods using R*. Sage.

Leite, W. L., Aydin, B., & Gurel, S. (2019). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *Journal of Experimental Education, 87*(1), 75-88. https://doi.org/10.1080/00220973.2017.1409179

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies. *Multivariate Behavioral Research, 50*, 265-284. https://doi.org/10.1080/00273171.2014.991018

Lohr, S. L. (1999). *Sampling: design and analysis*. Duxbury Press.

Lumley, T. (2011). "survey: analysis of complex survey samples". R package version 3.62.1

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*, 2937-2960. https://doi.org/10.1002/sim.1903

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*, 403-425. https://doi.org/10.1037/1082-989X.9.4.403

McKelvey, R. D., & Zavoina,W. (1975).A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120. https://doi.org/10.1080/0022250X.1975.9989847

National Center for Education Statistics. (2010). School survey on crime and safety. Retrieved from http://nces.ed.gov/surveys/ssocs on June 1 2011.

Neugebauer, R., & van der Laan, M. (2005). Why prefer double robust estimates in causal inference?. *Journal of Statistical Planning and Inference, 129*, 405-426. https://doi.org/10.1016/j.jspi.2004.06.060

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. https://doi.org/10.1037/1082-989X.8.4.434

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrived from http://www.Rproject.org.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*, 550-560. https://doi.org/10.1097/00001648-200009000-00011

Robins, J. M., & Rotnitzky A. (2001). Comment on the Peter J. Bickel and Jaimyoung Kwon, 'Inference for semiparametric models: Some questions and an answer'. Statistica Sinica, *11*, 920-936

Rosenbaum, R. P. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 408*, 1024-1032. https://doi.org/10.2307/2290079

Rosenbaum, R. P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistics Society, 53*, 597-610. https://doi.org/10.1111/j.2517-6161.1991.tb01848.x

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica, 70*, 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524. https://doi.org/10.2307/2288398

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34*, 441-456. https://doi.org/10.1207/S15327906MBR3404_2

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701. https://doi.org/10.1037/h0037350

Rubin, D. B. (2007). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. *Handbook of Statistics, 27*, 28-63. https://doi.org/10.1016/S0169-7161(07)27002-6

Schafer, J. L. & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Method, 13*(4), 279-313. https://doi.org/10.1037/a0014268

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Stapleton, L. (2008). *Chapter18: Analysis of data from complex surveys*. In: E. D. de Leeuw, J. J. Hox & D. A. Dillman. International handbook of survey methodology. Psychology Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-276. https://doi.org/10.1037/a0018719

Strayhorn, T. L. (2009). *Accessing and analyzing national databases.* In T. J. Kowalski & T. J. Lasley II (Eds.), Handbook of data-based decision making in education (pp. 105-122). NY: Routledge.

Sturmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution-a simulation study. *Practice of Epidemiology, 172*(7), 842-854. https://doi.org/10.1093/aje/kwq198

Stuart, E. A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science, 25*(1), 1-21. https://doi.org/10.1214/09-STS313

Thoemmes, F. J. & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*, 90-118. https://doi.org/10.1080/00273171.2011.540475

Thoemmes, F. J. & West, S. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*, 514-543. https://doi.org/10.1080/00273171.2011.569395

U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse. (2013). What Works Clearinghouse: Procedures and Standards Handbook (Version 3.0). Retrieved from Washington, DC: http://whatworks.ed.gov

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer.

Weitzen S., Lapane K. L., Toledano A. Y., Hume A. L., & Mor V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety, 13*, 841–853. https://doi.org/10.1002/pds.969

Winship, C. & Morgan, S. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-706.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer.