Iranian Journal of Economic Studies, 9(1) 2020, 181-212

Shiraz University

Iranian Journal of Economic Studies

Journal homepage: ijes.shirazu.ac.ir

Comparing Different Methods of Estimating the Variance of Propensity Score Matching Estimator

Alireza Kamalian, Seyed Komail Tayebi*, Alimorad Sharifi, Hadi Amiri

Department of Economics, University of Isfahan, Isfahan, Iran.

Article History

Received date: 05 August 2020
Revised date: 12 December 2020
Accepted date: 24 February 2021
Available online: 16 April 2021

JEL Classification
C14
C15
F02

Keywords
Matching
Propensity Score
Monte Carlo
Resampling

Abstract

Propensity score matching is extensively utilized in estimating the effects of policy interventions and programs for data observations. This method compares two treatment and control groups to make statistical inferences about the significance of the effects of these policies on target variables. Therefore, when using propensity score matching, it is significant to obtain the standard error to estimate the treatment effect. The precise estimations of variance and standard deviation facilitate more efficient statistical testing and more accurate confidence intervals. However, there is no agreement in the literature on the estimation method of standard error; some methods rely on resampling, while others do not. This study compares these methods using Monte Carlo simulation and calculating the Mean Squared Errors (MSE) of these estimators. Our results indicate that Jackknife and standard methods are superior to Abadie and Imbens (2006) bootstrap, and subsampling ones in terms of accuracy. Finally, reviewing Tayyebi et al. (2019) indicated that different methods of estimating variance in the matching estimator led to different statistical inferences in terms of statistical significance.

Highlights

- Propensity score matching is abundantly utilized in estimating the effect of policy interventions and programs for data observations.
- There is no agreement in the literature upon the estimation method of standard error.
- Our results indicate that Jackknife and standard methods are superior to Abadie and Imbens (2006) bootstrap, and subsampling ones in terms of accuracy.

* komail32@gmail.com

DOI: 10.22099/ijes.2021.38054.1692

1. Introduction

Any study in the social sciences that aims to analyze the impact of a policy intervention requires the use of statistical tests to examine and analyze the results of its outputs. Since the units that are intervened may have different characteristics from people who are in control and out of the intervention in the implementation of statistical tests, the differences among these groups should be controlled to obtain a virtually neutral estimate of the effects desired. For instance, in the study of estimating the impact of changing the subsidy method, the impact of high-income deciles may be very different from that of the low-income deciles which, due to a wide range of characteristics such as socioeconomic status and academic performance, are not even able to provide basic needs of life. It is vital to differentiate the effects of changing the subsidy method from the effects of previous differences between the "treatment" and "control" groups because without separating these concepts, it is not possible to have a valid estimate of the results of changing the subsidy method. These methods are known as program evaluation methods or policy evaluation.

The method of evaluating the program determines the success or failure of policy interventions and the extent to which they affect the researcher can influence the future decisions of economic policymakers. One of the latest methods of program evaluation and analysis of causal effects is the matching method widely used in experiments based on observational studies[1]. The matching methods include a growing set of techniques that attempt to simulate random experiments when using observational data. In this method, the matching in question is repeated in a quasi-random experiment. This way, at first, sub-samples are selected from the treated and control groups that are only randomly different from each other in all variables observed. In using the matching method, for statistical inference and testing of the matching estimator, it is necessary to pinpoint the variance of this estimator. Due to the non-parametric nature of this method and the lack of a mathematically closed solution for the variance of the estimator, different methods have been proposed. These methods are divided into two groups of re-sampling and non-sampling: the first group includes Bootstrap, subsampling and Jackknife resampling methods, and the second group consists of standard methods and the Abadie-Imbens (AI) estimator.

[1] In economic research, an observational study is a policy evaluation research in which the selection and allocation of individuals, firms and objects to be studied between the two groups of program (treatment) and control (control) is done without the intervention of the researcher. This method is different from the randomized experimental study method in which the researcher himself randomly classifies individuals between the experimental and control groups. In a completely randomized empirical study, a policy or corrective action is applied to a number of economic factors, households or individuals, and the effect of this policy action is observed on the units exposed to the program. In a randomized experiment, the researcher selects the individuals or entities being tested using a completely random method, such as tossing a coin, while in an observational study, the researcher observes the subjects or measures the variable without any intervention. It directly examines the subjects and observes and measures only as a third person and observer. The reason for using observational study versus experimental experiments is that in some cases it is not possible to retest an experiment. However, the common feature of both methods is that there are two groups of treatment and control in both of them (Keshavarz, 2018).

These methods have been proposed by econometrics researchers to calculate the variance of the matching estimator. For example, the AI estimator was emphasized by Abadie and Imbens (2006) and the Wild bootstrap by Otsu and Rai (2017). The nonparametric bootstrap was also considered by Lee (2005) and Cameron and Trivedi (2005) and the subset method by Abadie and Imbens (2008). Despite the introduction of different methods for estimating the variance of the matching estimator, no comprehensive research has been conducted to compare such methods. The purpose of this study is to draw a comparison through the examination of the characteristics of estimators. Therefore, the subject of this study is to evaluate the desirable properties of the variance estimator in each of the above-mentioned methods and then compare these characteristics across different methods. None of these methods would seem superior to the others; whether one specific method has the most desirable characteristics in each case depends on the situation and different matching methods. To evaluate the advantages of each sampling method (Bootstrap, subsampling, or Jackknife resampling methods) and non-sampling (standard and AI estimators), firstly, the desirable characteristics of the variance estimator must be identified. Secondly, it must be determined which characteristics are involved in this superiority, then, the method of discovering these characteristics is examined, and finally, different methods are compared with each other.

In this study, we first review the literature and the theoretical foundations of these methods to introduce a matching estimator as well as the various methods used for estimating its standard deviation. Next, we run a data simulation to compare them. Finally, the results of the simulation are discussed.

2. Literature Review

The matching method has been used since the early twentieth century, but its theoretical foundations did not develop until the 1970s when Cochran and Rubin (1973) and Rubin (1992) investigated this method in terms of the treatment effects on the data in which there was only one covariate.

Althauser and Rubin (1970) indicated that a larger control group would result in a better matching. Dealing with multiple explanatory variables (covariates) was a challenge that arose from the data problem and its comparison as well as the fact that it made it difficult to find matches that had overlapping values. For instance, Chapin (1947) reported that in a set of 671 people in the treatment group and 523 people in the control group, only 23 pairs in six explanatory variables matched completely. Tremendous progress occurred in 1983 with defining propensity score, i.e. the probability of participating in the program (Rosenbaum & Rubin, 1983). One of the most topics discussed in the field of matching is the estimation of the variance of the matching estimator. There is ongoing debate among researchers on how the uncertainty of the propensity score should be included in the estimation of variance (Stuart, 2010). Some researchers, such as Ho et al. (2007), proposed a method similar to randomized experiments, while others believe that the uncertainty considerations should be taken into account in the

calculations. In other words, some researchers, such as Rubin and Stuart (2006), show that using the estimated propensity scores instead of their real values can result in overestimation, which can, in turn, make the confidence intervals much larger than the actual values. This way, null-hypothesis will be accepted too often. Rubin and Thomas (1992) provided an analytical explanation of bias and variance reduction. Rubin and Thomas simulations and the examples of Hill et al. (2000) have confirmed these results. Due to the ineffectiveness of the standard methods used for variance estimation, the resampling methods, such as bootstrap were proposed (Lechner, 2002)

Abadie and Imbens (2006) examined the variance of the matching estimator, where they used the nearest neighbor to study the variance of the treatment effect. They demonstrated that the variance of the average treatment effect estimator is typically estimated using the sample variance of the within-pair differences, but the variance of the average treatment effect estimator can be considerably smaller for this matching estimator. Therefore, they presented a method for estimating the variance of the treatment effect using the pairs-of-pairs method. In this method, each unit is matched with another one within its group, and then the average treatment is obtained with a series of variance transformations. Although their method for estimating variance did not estimate the compatibility of variance, the mean of this variance was consistent with the conditional variance of the treatment effect in order to lead to a valid confidence interval.

Furthermore, the authors claimed that the results of the Monte Carlo simulation exhibited that their method was accurate even for small samples. First, they focused on the conditional variance of treatment effect rather than the variance of treatment effect, and secondly, they relied on the Monte Carlo simulation to prove their claim that the method worked even for small samples. In other words, this method involved two matches: one match to obtain treatment effect and another within-group match to obtain the variance of treatment effect.

In another study by Abadie and Imbens (2008), the authors examined the usage of the bootstrap method to estimate the variance of the matching estimator. By introducing the matching method, they first used the nearest neighbor method with a fixed number of matches with replacement. Next, they used the bootstrap method to estimate the standard errors in an example and then revealed that this estimator is an inconsistent estimator for the real standard error of the matching estimator (τ). This indicates the failure of bootstrap as considering the ratio of the people in the treatment group to the people in the control group ($\alpha = \frac{N_1}{N_0}$); this method is an overestimation in a range of α and underestimation in another range. This is since each observation in the bootstrap method may be used more than once, and this is the reason for the consistency of the bootstrap method; whereas, if α decreases, i.e. the ratio of people in the treatment group to the people in the control group decreases, this bias will decrease. The alternative method introduced in this paper is the AI or subsampling method.

Otsu and Rai (2017) introduced another method for estimating the variance of treatment effect. Considering the failure of bootstrap mentioned by Abadie and

Imbens (2008), the authors proposed wild bootstrap, originally introduced by Wu (1986) and Mammen (1993). The difference between this type of bootstrap and the one mentioned by Abadie and Imbens is that this method implements semi-parametric bootstrap rather than its non-parametric counterpart as well as two defined values of u_1^* and u_2^* instead of u^* of the bootstrap method. Their results indicated a consistent estimate of the variance of treatment effect which caused $\hat{\tau}$ to tend to the standard normal in the asymptotic distribution. The results obtained were confirmed by a simulation, which showed that this type of bootstrap could be used instead of its standard counterpart.

Austin and Small (2014) examined using bootstrap to obtain the variance of treatment effect. Their matching method involved propensity scores and was without replacement. They proposed two bootstrap methods: first, bootstrapping and resampling the pairs matched through their propensity score (the pairs that were matched using propensity score were exposed to resampling) and second, bootstrapping and resampling on the original data, which was done once every time on the original data, and then these steps were further repeated using the propensity score. Their simulations showed that using bootstrap would lead to results very similar to the parametric approach in calculating the variance of treatment effect, and the simple estimation of the matching variance will lead to the largest overestimation.

Pingel (2018) examined and analyzed several variance estimators of the matching estimator of the propensity score matching to estimate the average treatment effect (ATE) in which the role of smoothing parameters on the estimator variance of the matching is discussed. For this purpose, the criterion of mean squared error is used. His results indicated that the variance estimator proposed by Abadie and Imbens (2012) is effective in large samples. However, there are some caveats in using this estimator including what R-package software introduced in the Psmatch packages in Stata software, special settings must be considered (Sekhon, 2007). Pingel (2018) examined from 5 to 15 matches in his review. Subsequently, through changing, there is no difference in his results. In addition, his findings showed that in small samples, the probability of the bias is higher in the variance estimator, i.e. the higher the sample size, the less likely there would be bias and problems with the confidence interval.

Austin and Cafri (2020) investigated the different variance estimators of the mating coefficient estimator using Monte Carlo simulations. They studied cases where the matching is a placement and the data is survival or time-to-event (TTE) outcomes. Their simulation results demonstrated that the matching estimator shows the results without bias of the average treatment effect. However, in examining the variance of this estimator in the usual method, there is a bias in some cases. Although they estimated the size of this bias to be less than 30%, estimating the variance is still a problem for over-estimating. When the prevalence of treatment is relatively low (30%), one should use a robust estimator that calculates intra-pair clustering, as this leads to the most accurate estimate of sampling variability is the log odds ratio. While, in case the prevalence of

treatment is high (e.g., 50%), the proposed variance should be preferred. It should be noted, however, that the first set in a low-prevalence setting is likely to be the one in which matching without placement is well performed and placement matching may not be necessary.

Reviewing the literature reveals that despite the introduction of methods for calculating the variance of the matching estimator, comparisons between these methods have received less attention. This study aims to gather different methods used to obtain the variance of the matching estimator and then compare the robustness of the methods.

3. SETUP

3.1 Matching Method

This study addresses the standard model. The goal is to evaluate the effect of a treatment method based on the data retrieved from the results, treatments, and auxiliary variables for the treated and untreated units. In general, the matching method determines the average effect of a binary treatment variable on an outcome variable (Y). For each unit $i = 1, \dots, n$, there are two values of Y: $Y_i(1)$ when the unit participates in the program and $Y_i(0)$ when it does not. The variable $w_i \in \{0,1\}$, which indicates the participation of the unit in the program is defined as Equation 1:

$$Y_i = \begin{cases} Y_i(0) & w_i = 0 \\ Y_i(1) & w_i = 1 \end{cases} \quad (1)$$

If $E(Y_1) - E(Y_0) > 0$, the program performed on the unit is effective. Equation 2 is used to determine the treatment effect on the entire population.

$$E(Y_1 - Y_0) = E(Y_1) - E(Y_0) \quad (2)$$

If individuals are distributed into treated and control groups randomly, this effect can be presented via Equation 3:

$$\tau = E(Y|w = 1) - E(Y|w = 0) = E(Y_1) - E(Y_0) \quad (3)$$

If individuals are not distributed randomly, other observable variables that affect the outcome variable (Y) should be identified and their effects should be controlled. To this purpose, one can choose to compare the groups where the variables have overlapping values, arising from k covariates of X:

$$E(Y|x, w = 1) - E(Y|x, w = 0) = E(Y_1|x, w = 0) - E(Y_0|x, w = 0) =$$
$$= E(Y_1|x) - E(Y_0|x) = E(Y_1 - Y_0|x) \quad (4)$$

In this case, conditional on given X, a random distribution, which is called selection on observables, is carried out (Lee, 2005).

Assuming that participation in the program is independent of the outcome variable and that the probability of participating in the program for each given X is in the range of 0 and 1 (Hackman et al., 1998), the treatment effect can be presented via Equation 5:

$$\tau(x) = E[Y(1) - Y(0)|\boldsymbol{X} = \boldsymbol{x}]$$
$$= E[Y|W = 1, \boldsymbol{X} = \boldsymbol{x}] - E[Y|W = 0, \boldsymbol{X} = \boldsymbol{x}] \quad (5)$$

In these conditions, the difference between the variables on the right-hand side of this equation for each X is detected. As a result, the average treatment effect can be determined by calculating $\mathbb{E}[Y|W = 1.X = x] - \mathbb{E}[Y|W = 0.X = x]$ for all X values (Lee, 2005). The average treatment effect for a treated group is defined via Equation 6:

$$\tau = E[\tau(X)] = E\big[E[Y|\boldsymbol{W} = 1, \boldsymbol{X} = \boldsymbol{x}] - E[Y|W = 0, \boldsymbol{X} = \boldsymbol{x}]\big] \quad (6)$$

One of the main challenges of program evaluation is whether only one of the variables of $Y_i(1)$ and $Y_i(0)$ (for the treatment and control groups, respectively) would be observable (Holland, 1986). In this regard, the potential unobservable outcome of each sample should be estimated (Stuart, 2010), in which the potential outcome of the program ($Y_i(0)$) in case of a relationship with the covariate X is considered as a variable exposed to the treatment. If the participation in treatment for units with the same covariates is completely random, the outcome variable of the control group can be used to estimate the potential outcome in case of non-participation in the treatment program provided that the covariates overlap. This is the basis of the matching method (Keshavarz, 2018).

As the matching of a large number of control variables leads to the decrease of the data overlapping and reduces the possibility of comparing the control and treatment groups, it is necessary to devise alternative methods for matching, one of which is using a propensity score that is drawn from control variables (Rosenbaum & Rubin, 1983). In propensity score matching, when the vector dimensions of the control variables are so large that the assumption of the data overlap is difficult, a logit or probit estimation is first fitted to the vector of the control variables, and then the probability of participating in the program for each unit is obtained. Finally, data matching is carried out using this propensity score (Cameron & Trivedi, 2005). In the propensity score approach, matching can be performed using four methods, namely nearest neighbor, caliper and radius, stratification and interval, and kernel (Becker & Ichino, 2002). In other words, we have:

$$y_0, y_1 \perp W|x \;\rightarrow\; y_0, y_1 \perp W|p(x) \quad (7)$$

The assumption of conditional independence under given x also includes the assumption of conditional independence under given p(x). In this regard, we have:

$$\begin{aligned} &Pr[W = 1|y_0, y_1, p(x)] = E[W|y_0, y_1, p(x)] \\ &= E[E[W|y_0, y_1, p(x), x]|y_0, y_1, p(x)] \\ &= E[E[W|y_0, y_1, x]|y_0, y_1, p(x)] \\ &= E[E[W|x]|y_0, y_1, p(x)] \\ &= p(x) \end{aligned} \quad (8)$$

The method of the nearest neighbor is commonly implemented in the matching method using observable variables. To achieve this, each unit *i* in the treatment group is matched with unit j of the control group provided that the values of the explanatory variables of these two units are nearest to each other. In this case, their subtraction and average are assigned to the outcome variable and treatment effects, respectively. If each data can be matched only once, the

matching is without replacement, and if it can be matched more than once, it is called "matching with replacement" (Abadie & Imbens, 2011).

In matching the nearest neighbor for each unit *i* in the treatment group, D is the distance between the value of the explanatory variables of the i[th] unit and the value of the explanatory variables of the closest match in the control group.

$$D_i = min\left\|X_i - X_j\right\| \qquad j = 1, \dots, n: W_j = 0 \tag{9}$$

In case of n-matched units:

$$\zeta(i) = \left\{j \in \{1,2, \dots, n\}: W_j = 0, \left\|X_i - X_j\right\| = D_j\right\} \tag{10}$$

ζ (i) is the set of the matched units for the i-th unit (when *i* is in the control group, the set ζ(i) will be empty). Thus:

$$\hat{Y}_i(0) = \frac{1}{\#\zeta(i)} \sum_{j \epsilon \zeta(i)} Y_i \tag{11}$$

In this case, the matching estimator of the treatment effect is defined as Equation 12:

$$\hat{\tau} = \frac{1}{n_1} \sum_{i:W_i=1} \left[Y_i - \hat{Y}_i(0)\right] \tag{12}$$

To test the treatment effect, it is necessary to obtain variance after estimating the average within-paired effects. The following methods are suggested for this purpose:

3.2 Methods of Variance Estimation

1) Methods not based on resampling

a) Common Standard Error: This is obtained using the usual method of calculating the mean difference (between the treatment and control groups) with the weight observed by the weights on consistent data matched. Note that the standard error presented in this method does not take into account the uncertainty of the matching method.

b) AI method (Abadie & Imbens, 2006): In this method, an estimator is proposed for variance in Equation 13:

$$\hat{V}^{AI} = \frac{1}{N_2^1} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i(0) - \hat{\tau}\right)^2 + \frac{1}{N_2^1} \sum_{i=1}^{N} \left(K_i^2 - K_{sq.i}\right) \hat{\sigma}^2(X_i . W_i) \tag{13}$$

where $\hat{\sigma}^2(X_i . W_i)$ is the conditional variance estimator of the outcome variable Y_i in given X_i and W_i based on matching where each variable will be matched with the closest explanatory variables within the group:

$$K_i = \begin{cases} 0 & W_i = 1 \\ \sum_{W_j=1} 1\{i \in \zeta(j)\} \frac{1}{\zeta(j)} & W_i = 0 \end{cases} \tag{14}$$

$$K_{sq.i} = \begin{cases} 0 & W_i = 1 \\ \sum_{W_j=1} 1\{i \in \zeta(j)\} \frac{1}{\zeta(j)}^2 & W_i = 0 \end{cases} \tag{15}$$

2) Methods based on Resampling:

a) Bootstrap: The bootstrap algorithm is similar to the Monte Carlo simulation, except that in the latter, the random sample is extracted from a distribution given with known parameters, such as a normal distribution, but in the bootstrap, random samples are derived from the empirical distribution

function (EDF). Another difference is that this method is based on the principle of replacement, according to which, the empirical distribution function based on an observed sample is the best estimation of the theoretical distribution function in a non-parametric analysis.

Efron (1992) originally proposed the basis of the bootstrap method and stated that the observed data set is a random sample with size N, derived from the theoretical distribution function; in other words, the empirical distribution function of data is the best estimate of the theoretical distribution function of the data. The empirical distribution function is defined as a discrete distribution in which the probability of occurrence of each of the observed values is equal to 1/n. Hence, the empirical distribution is formed through a random variable rather than a predetermined distribution such as the normal distribution.

Thus, a bootstrap sample is a random sample $X_1^*, X_2^*, \dots, X_n^*$ which is obtained through replacement and placing the probability of 1/n for each of the values observed. The steps of using the bootstrap method are as follows: 1) From the given sample $w_1,\dots,w_n$, a bootstrap sample with size n is derived. This new sample is denoted by $W_1^*, W_2^*, \dots, W_n^*$. 2) A suitable statistic that uses a bootstrap sample is calculated. This step includes a) estimation of $\hat{\theta}^*$ fromθ, b) the standard error of $s_{\hat{\theta}^*}$ from $\hat{\theta}^*$, c) statistic $t^* = \left(\hat{\theta}^* - \hat{\theta}\right)/s_{\hat{\theta}^*}$ that is distributed around the main estimate $\hat{\theta}$. Here, $\hat{\theta}^*$ and $s_{\hat{\theta}^*}$ are calculated in the usual way, but using the bootstrap sample rather than the original one. 3) Independent repetition of steps 1 and 2 for B times, where B is a large number of the bootstrap iterations of the desired statistic, such as $\hat{\theta}^*{}_1 \dots \hat{\theta}^*{}_B$ or $t_1^* \dots t_B^*$. 4). This repetition for B times is used to calculate the distribution of the bootstrap statistics, such as τ. The simplest way to bootstrap is to use an empirical data distribution that considers the sample as the entire population. Then, $W_1^*, W_2^*, \dots, W_n^*$ are obtained through sampling with the replacement of $w1,\dots,w_n$. In each of the bootstrap samples obtained, some original data are repeated several times, while some data do not exist at all. This method is called the Empirical Distribution Function (EDF) or non-parametric bootstrap.

$$\widehat{se}_{bootstrap}\left[\hat{\theta}\right] = \left[\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_{(bootstrap)} - \overline{\widehat{\theta_{bootstrap}}}\right)^2\right]^{1/2} \quad (16)$$

b) Subsampling Method: The subsampling method implements subsamples with size m that are much smaller than the original sample (N). These subsamples can be selected with or without replacement (Politis & Romano, 1994).

The subsample with replacement provides subsets which are random samples of the population rather than using a random sample distribution estimate, and this is similar to that of the bootstrap. Subset bootstrap is used when a complete bootstrap sample is invalid or when it is used as a validation of a complete bootstrap sample. Results vary by selecting different sizes of subsets (Cameron & Trivedi, 2005).

$$\widehat{se}_{subsample}\left[\hat{\theta}\right] = \left[\frac{1}{SB-1}\sum_{i=1}^{SB}\left(\hat{\theta}_{(\zeta(i))} - \overline{\widehat{\theta_{subsample}}}\right)^2\right]^{1/2} \quad (17)$$

c) Jackknife Method: An alternative method to bootstrap resampling is the Jackknife method discussed earlier. Jackknife is derived from N deterministic subsets from the original data with size N-1, which are used by separating each of the n observations and recalculating the estimators. Consider the estimator $\hat{\theta}$ from the parameter vector θ that is based on a sample with size N. For i = 1, ... , N, sequentially delete the i-th observation and perform N Jackknife repetitions of estimator $\hat{\theta}$ from N resampling (sample size of N-1). The unbiasedness estimation of Jackknife from $\hat{\theta}$ is equal to $(N-1)\left(\bar{\hat{\theta}}-\hat{\theta}\right)$, where $\bar{\hat{\theta}}=N^{-1}\sum_i \hat{\theta}_{(-i)}$ is called the average n repetition of Jackknife from $\hat{\theta}_{(-i)}$. The unbiasedness appears to be large as it is multiplied by n-1, but the difference $(\hat{\theta}_{(-i)}-\hat{\theta})$ is much smaller than that of the bootstrap, because in Jackknife resampling, the difference between a new sample and the original one is just one observation.

Jackknife is considered as a solution for a wide range of statistics. In particular, Jackknife's estimate of the standard deviation of estimator $\hat{\theta}$ is equal to Equation 18:

$$\widehat{se}_{Jack}[\hat{\theta}]=\left[\frac{N-1}{N}\sum_{i=1}^{n}\left(\hat{\theta}_{(-i)}-\bar{\hat{\theta}}\right)^2\right]^{1/2} \quad (18)$$

It seems that Jackknife is a linear approximation of bootstrap (Tibshirani & Efron, 1993). In small samples where N<B[2], Jackknife involves fewer calculations, but when $B \to \infty$, the performance of the bootstrap will improve.

Each of these methods has been proposed by econometricians to calculate the variance of the matching estimator. For example, the AI method and wild bootstrap were introduced by Abadie and Imbens (2006) and Otsu and Rai (2017), respectively. Furthermore, non-parametric bootstrap was introduced by Lee (2005), Cameron and Trivedi (2005), while the subset method was proposed by Abadie and Imbens (2008). According to these studies, however, no comprehensive research has been conducted to compare these methods. This study thus aims to carry out such comparison has and opt for mean squared error (MSE) as its criterion.

4. Methodology

4.1 An Introduction to Simulation

There are two methods to explore the validity of statistical tests: 1) theoretical proof, 2) performing simulations. For statistics that are based on resampling, simulation methods are often preferred due to the complexity of theoretical proofs (in some cases they can also lead to no results). Simulation is a branch of science that uses artificial experiments which are very similar to their real-world counterparts to answer real-world questions. This method is used in the modelling of physics, cosmology, chemistry, meteorology, biology, economics and social and engineering sciences. In terms of models, computer

[2] B is the repetition number of the bootstrap method.

simulations can be classified based on different perspectives, namely stochastic or deterministic[3], static or dynamic[4], continuous or discrete[5], and local or distributed. In sum, simulation denotes the visualization of something about which sensory information is not available, and it uses a fictitious environment and related theoretical model to estimate the behavior of a real-world system.

One of the most widely utilized simulation methods is the Monte Carlo simulation method or the Monte Carlo algorithm, which is a class of computational algorithms based on the repetition of random samples. This method is substantially used for solving non-random equations of physical, mathematical and statistical systems as well as confirming the solution of analytical methods.

Simulation is a virtual representation of reality[6]. The Monte Carlo method is reliable for solving mathematical and statistical problems. The Monte Carlo simulation implements repetitive sampling to determine the characteristics of some phenomena. Despite some variations, this method usually involves the following steps: 1) Defining the domain of possible inputs; 2) Random generating inputs according to a probability distribution over the domain; 3) Applying theoretical and deterministic operations on inputs; and 4) Collecting the results.

4.2 Monte Carlo Simulation

In this study, the Monte Carlo simulation is utilized to obtain the mean asymptotic standard error of various methods used for variance estimation to estimate the variations of treatment effect through propensity score matching. Its outputs are considered to be continuous.

4.3 Data Generation Process

The data are based on an environment of 10 covariates ($X_1,\ldots,X_{10}$), which are simulated from the independent standard normal distribution. Some of these variables affect the participation probability; while some affect the outcome

[3] A deterministic simulation model consists only of deterministic (not stochastic) components, in which, all the mathematical and logical relations between items (variables) are fixed and not exposed to uncertainty. A typical example of a deterministic simulation model is a complicated (and analytically intractable) system of differential equations describing a chemical reaction. In contrast, a model with at least one random input variable is a random model. Most queuing and inventory systems are modelled stochastically.

[4] Static models are not exposed to change over time, and therefore, do not convey the passage of time, while a dynamic simulation model represents a system as it evolves over time (for example, the performance of a traffic light).

[5] In discrete simulations, the state variables change instantaneously at separated points in time. The mathematical models used for calculating the numerical answers for a set of differential equations is an example of a continuous simulation, while the queuing models are examples of discrete simulations.

[6] The Monte Carlo method can also be explained according to the definition of Banks and Carson: Monte Carlo method uses random numbers to solve non-stochastic problems or some other stochastic ones where time has no major role. Random numbers in this definition refers to independent random variables with a uniform statistical distribution in the range of [0,1]. According to this definition, Monte Carlo method is considered as a static method rather than a dynamic one. This does not apply to the Monte Carlo simulation whose only similarity to the Monte Carlo method is the use of random numbers. In fact, the time factor is involved in the simulation; i.e. the simulation is a dynamic method.

variable. These variables influence the program selection or outcome variable to different degrees: weak, moderate, strong, or very strong.

For each item, the probability of program selection is determined by the following logistic model:

$$\log it(p_i) = \alpha_{0,treat} + \alpha_w x_1 + \alpha_M x_2 + \alpha_S x_3 + \alpha_w x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS} x_7 \quad (19)$$

The constant term for the model selection ($\alpha_{0,treat}$) is an item proportional to the treatment effect in the simulated samples of the program. The regression coefficients $\alpha_{VS}, \alpha_S, \alpha_M, \alpha_W$, is the set of log(0.01), log(0.25), log(0.5), log(0.8). These numbers are chosen such that the selected effect will be weak, medium, strong and very strong. For each item, the program state is generated from the Bernoulli distribution with the parameter p_i $Z_i \sim Be(p_i)$. Finally, the outcome variable is produced by a linear relationship among the variables that affect the outcome variable and using the treatment effect obtained with Bernoulli's output. In the simulation of the outcome variable, the error term with normal distribution is used. The process starts with 100 data, and 30 data are added to the simulation in every series. The bootstrap, Jackknife and subsampling methods are also repeated 1000 times unless the sample size is small, in each case, the maximum possible number of repetitions has been performed.

4.4 Simulation Results

Considering the asymptotic investigation of mean squared error in this study, the sample size starts with 130 and ends in 1000. The periods are presented in Table 1.

Table 1. Results of Mean Squared Error

N=1000	*N=850*	*N=700*	*N=550*	*N=130*	
			M=1		
359.3028633	276.232176	320.8182815	251.6042997	500.8100998	MSE(AI)
0.016929846	0.020581401	0.016482093	0.028412348	0.486505001	MSE(NORM)
1794.136113	347.2546649	448.0167965	222.7023287	701.0831262	MSE(bootstrap)
0.012201001	0.002689886	0.000772705	0.026731482	36.53862128	MSE(jackknife)
2085.182374	1570.961438	1917.263219	1785.628421	807.5742217	MSE(moutofn)
			M=2		
68.67815456	114.9749764	130.0815033	153.6618816	483.5759725	MSE(AI)
0.008739926	0.022771289	0.024879217	0.041212902	2.537876126	MSE(NORM)
17.34399815	68.39050707	153.1516715	108.5377164	474.7539165	MSE(bootstrap)
0.000300102	0.000260086	0.000216443	0.005465036	22.98551485	MSE(jackknife)
373.7873398	512.1124617	363.0121975	524.5128277	280.1948388	MSE(moutofn)

Table 1 (Continued). Results of Mean Squared Error

N=1000	*N=850*	*N=700*	*N=550*	*N=130*	
			M=3		
153.8657165	54.6881585	45.08985479	123.2876588	248.5297026	MSE(AI)
0.035104951	0.019458713	0.021911503	0.067272708	3.431057022	MSE(NORM)
227.5139889	186.4969325	116.2256058	132.8650925	336.4494912	MSE(bootstrap)
0.000208132	0.034295835	0.001420443	0.288799861	1.108353293	MSE(jackknife)
277.8710203	225.571591	266.7967234	180.0236714	141.8771906	MSE(moutofn)
			M=4		
31.00458324	44.89214356	26.27894386	37.88330743	70.22433792	MSE(AI)
0.012669206	0.027771789	0.02011709	0.034149579	1.185585967	MSE(NORM)
22.53560794	104.136727	15.69573467	20.25199681	72.86821334	MSE(bootstrap)
1.10E-05	9.89E-05	2.84946E-05	0.009779622	0.016678995	MSE(jackknife)
141.5917367	118.8029588	96.34481076	102.3407322	80.90357222	MSE(moutofn)
			M=5		
47.09603191	21.94513073	34.53372145	24.75062024	35.6585449	MSE(AI)
0.033099209	0.021262261	0.028486752	0.035276015	1.019225971	MSE(NORM)
122.6722605	23.44849666	71.80630725	36.05028064	92.45200067	MSE(bootstrap)
4.60E-05	6.06E-06	0.063169732	0.065037417	0.019011204	MSE(jackknife)
107.2703018	70.35170284	76.6647025	84.7591145	70.74899017	MSE(moutofn)

***Source**: Authors*

To facilitate comprehension, a simpler presentation of the results is depicted in Fig.

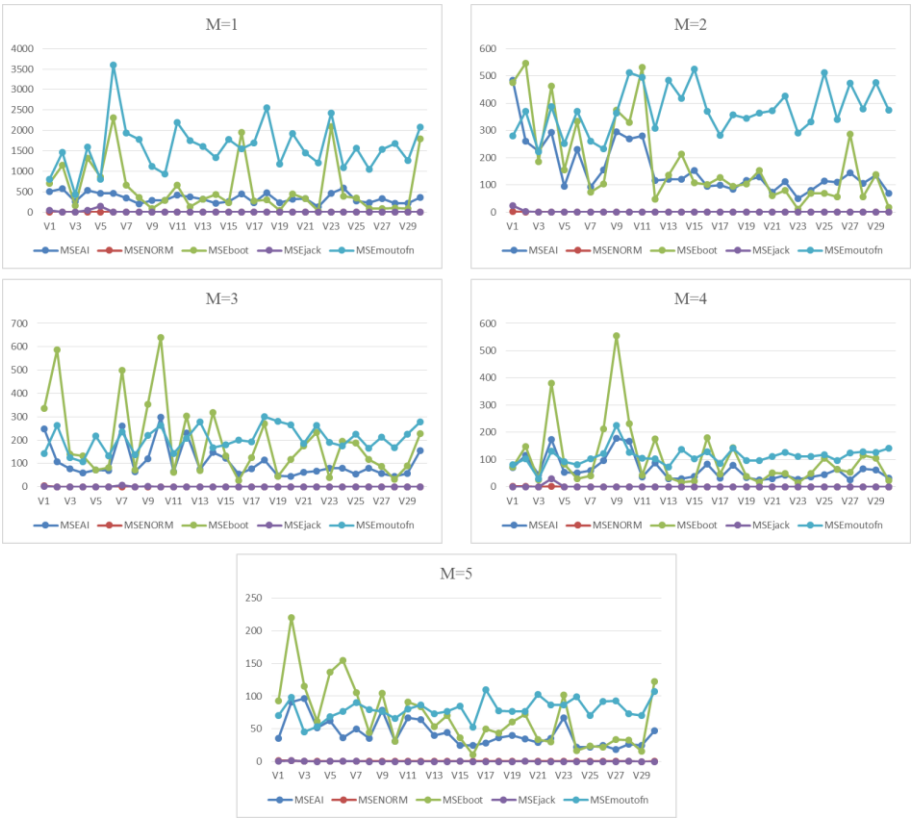

Figure 1. Results of Mean Squared Error

Table 1 and Figure 1 show the analyses of MSE values. For each of the five Standard and AI, Bootstrap, Jackknife and subsampling, the MSE values are investigated. This study is accompanied by an increase in the sample size from sample size (130) to 1000. In other words, MSE is examined asymptotically. Moreover, initially, the number of matches for each observation in the treatment group was one, which i increased to 5. This takes into account the effect of increasing the sample size along with the effect of increasing the number of matches on the average squares of the error with each other, hence, more robust results. Increasing the number of matches can lead to stronger conclusions when the number of people in the control group is much higher than the number of people in the treatment group. In all these cases, the average square error in the standard method and Jack Knife method was less than other three AI, Bootstrap and subsampling methods, which indicates that these two methods are more efficient than other methods. It was more efficient in that it was present at all levels of sample sizing. In other words, these two methods were superior to other methods in small samples, as well as in large samples, although the MSE in all methods decreases with increasing sample size. It indicates the compatibility of

these methods. However, despite such a reduction, the standard and Jackknife methods are still more efficient than the AI, subsampling and Bootstrap methods. Given that SME is the sum of the squared bias and variance of the estimator, it can be claimed about the Jackknife method that the variance of this method is less than others'. This reduction in variance is because this method is more effective than other types of representations. In other words, using the Jackknife method minimizes the estimator variance regarding the actual value of the population. This reduction in variance is also because new samples obtained from the sampling are not fundamentally different from the original sample. This advantage reduces the variance of the estimators calculated by the Jackknife method. In addition, the use of the Jackknife and subsampling methods in which there is no repetitive observation of the original sample will reduce the bias caused by repetition, which was addressed in the Abadi-Imbens study where these two methods are compared to the bootstrap method. However, since the deletion of each observation causes the loss of the information obtained from it, the superiority of the Jackknife over the subsampling method is also evident for the number of observations deleted in the Jackknife method as minimum.

An empirical example is provided. In the continuation of the discussion and to express the issue that the use of different methods of estimating variance in the matching estimator causes differences in the significance of treatment, an empirical example is given. In this section, only the significant difference has been considered using different methods of estimating variance, and from experimental examples, no reason can be deduced from the advantage of any method. Tayyebi et al. (2019)[7] investigated the impact of globalization on government budget deficit using the matching method for modelling. To this end, the government debt and participation in the Organization for Economic Co-operation and Development (OECD) were considered as an outcome variable and a treatment variable, respectively. The matching variables were per capita production based on purchasing power, inflation, unemployment and trade liberalization index. The standard deviation in this study was obtained using the Abadie-Imbens method. The results indicated that the treatment variable (globalization) had a significant effect on the budget deficit of governments and increased them. They used the data from this research to re-estimate the standard deviation (by the five aforementioned methods) as well as the treatment effect. The results are represented in Table 2.

[7] The aim of reviewing Tayyebi et al. (2019) was simply to point out that the use of different methods of estimating variance leads to different results in rejecting or accepting the null hypothesis. So, no reference was made on the appropriateness of any method (and its data) in the present paper.

Table 2. Results of hypothesis testing using various methods of variance estimation

	Treatment Effect	Standard Deviation	Statistic t	The Critical Value of 5%	Result
Standard method	31.2014	11.3137	2.7578	1.96	Null-hypothesis is rejected.
Abadie-Imbens method		20.3224	1.5353		Null-hypothesis is accepted.
Bootstrap method		12.8062	2.4364		Null-hypothesis is rejected.
Jackknife method		1.7146	18.1974		Null-hypothesis is rejected.
Subsampling method		31.9530	0.9764		Null-hypothesis is accepted.

***Source**: Authors*

The results of Table 2 indicate that different methods used to estimate the standard deviation results in a change in the significance of the treatment variable as bootstrap and Jackknife methods render the treatment variable significant, while the use of Abadie-Imbens and subsampling methods makes it insignificant.

5. Conclusion

Estimating variance is controversial in the matching literature. Authors are mostly uncertain in computing either a propensity score or a matching method. Some researchers opt for a method similar to random experiments where models run conditionally on the auxiliary variables that are given as exogenous. In such conditions, the uncertainty related to the matching process is not taken into consideration. Other researchers state that analysis should consider uncertainty in propensity score estimation. Nevertheless, in practice, using estimated propensity scores instead of their real counterparts can lead to the overestimation of variance. Among the proposed methods, researchers have mostly focused on the methods based on resampling due to their simplicity and lack of necessity to estimate the distribution of the sample and population. Due to the non-parametric distribution of treatment statistics, it is not possible to find out the efficiency of and comparison among these methods. Therefore, the use of the simulation approach can be an appropriate indicator for comparing these methods. This study thus conducted a comparison between simulation methods using Monte Carlo simulation method. To this purpose, we calculated the mean squared errors of the estimated statistic, the results of which indicated that the mean squared errors of the bootstrapping, Jackknife, normal and Abadie-Imbens methods were

asymptotically zero. It was revealed that the normal and Jackknife methods had distinctively smaller mean squared errors, and their superiority was more pronounced for the models with small volume samples.

References

Abadie, A., & Imbens, G. W. (2009). Matching on the estimated propensity score. No. w15301. *Cambridge, MA: National Bureau of Economic Research*. doi, 10, w15301.

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235-267.

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537-1557.

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1), 1-11.

Abadie, A., & Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498), 833-843.

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics*, 86(1), 180-194.

Althauser, R. P., & Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology*, 76(2), 325-346.

Austin, P. C., & Cafri, G. (2020). Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in Medicine*, 39(11), 1623-1640.

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation*, 38(6), 1228-1234.

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine*, 33(24), 4306-4319.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962-973.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2(4), 358-377.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge university press.

Chapin, F. S. (1947). *Experimental designs in sociological research Harper & Row*. New York.

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

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

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a

discussion of some questions. *Journal of the National Cancer Institute*, 22(1), 173-203.
Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151-161.
Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932-945.
Efron, B. (1992). *Bootstrap methods: Another look at the jackknife*. In Breakthroughs in Statistics (pp. 569-593). Springer, New York, NY.
Fabra, N., von der Fehr, N. H., & Harbord, D. (2002). Designing electricity auctions: Uniform, discriminatory, and Vickrey. *preprint*.
Federico, G., & Rahman, D. (2003). Bidding in an electricity pay-as-bid auction. *Journal of Regulatory Economics*, 24(2), 175-211.
Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 29-46.
Greenwood, E. (1945). *Experimental sociology: A study in method*. King's Crown Press.
Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, statistics in medicine.
Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2), 481-488.
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.
Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2), 261-294.
Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4), 605-654.
Hill, J. L., Rubin, D. B., & Thomas, N. (2000). The design of the New York school choice scholarship program evaluation. *Validity and Social Experimentation: Donald Campbell's legacy*, 1, 155-180.
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199-236.
Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2), 481-502.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4-29.

Keshavarz Haddad, Gh.R. (2018). Micro econometrics and policy evaluation. *Ney*.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(1), 59-82.

Lee, M. J. (2005). *Micro-econometrics for policy, program, and treatment effects*. Oxford University Press on Demand.

Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *The Annals of Statistics*, 255-285.

Otsu, T., & Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520), 1720-1732.

Pingel, R. (2018). Estimating the variance of a propensity score matching estimator for the average treatment effect. *Observational Studies*, 4, 71-96.

Politis, D. N., & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 2031-2050.

Ren, Y. (2001). A comparison of pool cost and consumer payment minimization in electricity markets (Doctoral dissertation, McGill University Libraries).

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

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20-36.

Rubin, D. B., & Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics*, 34(4), 1814-1826.

Rubin, D. B., & Thomas, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, 1079-1093.

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

Sekhon, J. S. (2007). Multivariate and propensity score matching software for causal inference.

Song, J., Belin, T. R., Lee, M. B., Gao, X., & Rotheram-Borus, M. J. (2001). Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services and Outcomes Research Methodology*, 2(3-4), 317-329.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2), 395.

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score

matching in the medical literature between 1996 and 2003' by Peter Austin, statistics in medicine. *Statistics in Medicine*, 27(12), 2062-2065.

Stuart, E. A., & Lalongo, N. S. (2010). Matching methods for the selection of participants for follow-up. *Multivariate Behavioral Research*, 45(4), 746-765.

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

Tahmasbi, R. and Rezaei, S. (2012). Statistical simulation, *Professor Hesabi*.

Tayyebi, S. K., Kamalian, A.R., Sarkhosh Sara, A., and Mobini_Dehkordi, M. (2019). Analyzing the effects of globalization on the government budget deficit: The matching approach. *Economics and Modelling*, 10 (1), 65-96.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57, 1-436.

Wacholder, S., & Weinberg, C. R. (1982). Paired versus two-sample design for a clinical trial of treatments with the dichotomous outcome: Power considerations. *Biometrics*, 801-812.

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(12), 841-853.

Wu, C. F. J. (1986). Jackknife, bootstrap, and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4), 1261-1295.

Yu, C. H. (2002). Resampling methods: Concepts, applications, and justification. *Practical Assessment, Research, and Evaluation*, 8(1), 19.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics*, 86(1), 91-107.

Appendices

This section focuses on different simulation parameters that were transformed to examine and compare the results of MSE in different cases. The results indicate that changing the simulation parameters did not affect the study results.

Simulating the data and calculating MSE asymptotically in the beta.low coefficient change mode:

In this section, by changing the beta.low rate, its effect on the efficiency of variance estimation methods are investigated and MSE is examined asymptotically. The sample size starts from 100 and ends in 1100. The following table presents the results:

Table 1. MSE results by changing the variance of the treatment variable

N=1000	N=850	N=700	N=550	N=100	
			beta.low=log(0.01)		
45.03238	167.638	104.8148	314.1532	851.7977	MSE(AI)
0.023863	0.098655	0.084299	0.510176	8.04999	MSE(NORM)
20.18511	213.3705	195.6385	301.4295	2422.222	MSE(bootstrap)
3.03E-05	0.001576	0.000607	6.513935	1.364211	MSE(jackknife)
188.1768	216.2965	280.2235	231.4278	745.0608	MSE(moutofn)
			beta.low=log(0.02)		
36.67987	33.33802	38.65057	45.43468	285.2486	MSE(AI)
0.017977	0.021943	0.031761	0.068533	2.065475	MSE(NORM)
57.85447	28.47822	57.64277	24.02514	327.8553	MSE(bootstrap)
0.127796	0.000833	0.001251	0.314528	4.305075	MSE(jackknife)
131.6835	153.0183	145.3705	111.1654	184.0575	MSE(moutofn)
			beta.low=log(0.05)		
18.93307	22.0278	32.80268	24.81357	65.85022	MSE(AI)
0.008345	0.012467	0.015974	0.085154	0.453846	MSE(NORM)
30.46188	18.85829	28.00153	31.14083	51.53285	MSE(bootstrap)
5.08E-05	0.000141	0.003814	0.000298	0.240732	MSE(jackknife)
85.09032	103.0268	89.44665	71.23848	49.95104	MSE(moutofn)
			beta.low=log(0.10)		
15.99651	5.921596	10.34922	13.23577	33.84705	MSE(AI)
0.009209	0.008623	0.009921	0.022379	0.233514	MSE(NORM)
42.48825	7.180093	8.401672	15.85387	40.59844	MSE(bootstrap)
0.000119	3.26E-05	1.51E-05	0.010328	0.396901	MSE(jackknife)
57.25824	44.86089	53.49309	62.06096	44.42195	MSE(moutofn)

Table 1 (Continued). SME results by changing the variance of the treatment variable

beta.low=log(0.20)					
3.390019	4.815908	5.982962	5.479099	23.82313	MSE(AI)
0.002842	0.005899	0.005177	0.015145	0.119289	MSE(NORM)
4.886883	8.27036	9.401162	5.603212	24.69075	MSE(bootstrap)
0.004882	4.57E-05	4.28E-05	6.38E-05	0.105251	MSE(jackknife)
44.16853	37.49696	31.04247	40.39463	30.41826	MSE(moutofn)

***Source**: Research findings*

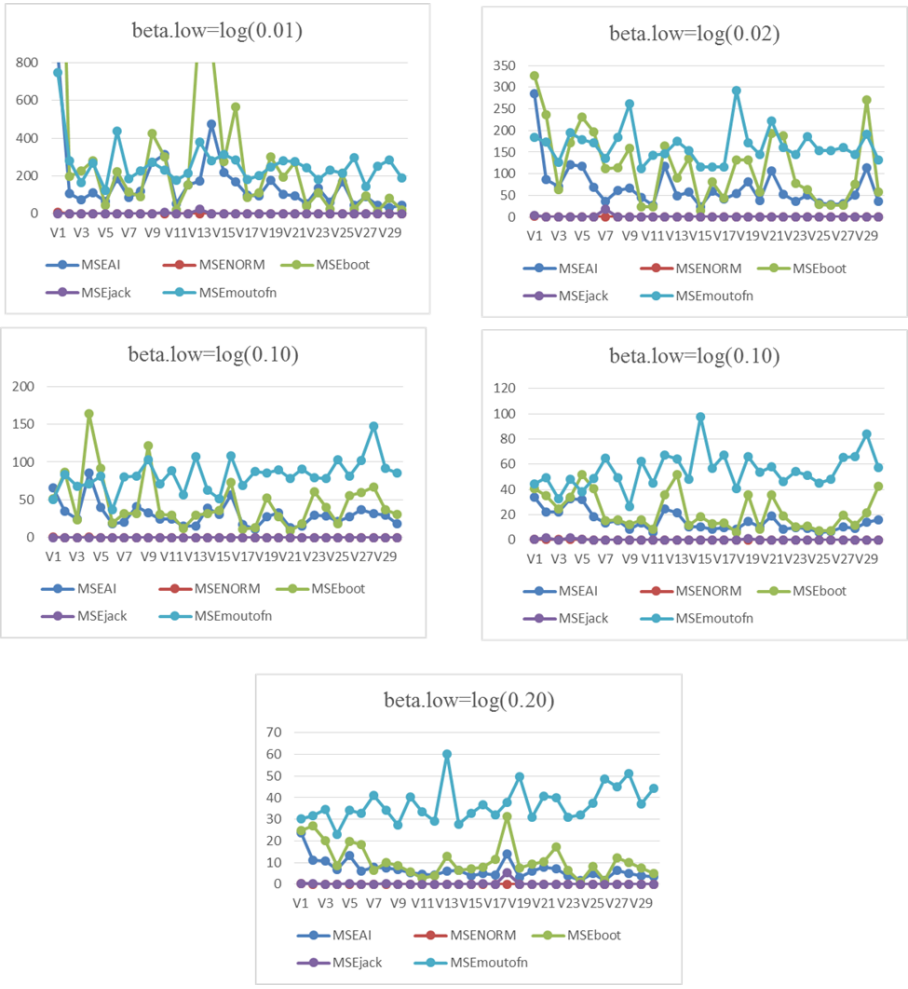

Figure 1. Graphs of the MSE results with the variance of the treatment variable

As the results in the above table and diagram shows, the MSE in the Bootstrap, Standard, AI, Jackknife, and Subsampling methods are decreasing asymptotically. Among these methods, the MSE for the Standard and the Jackknife method are lower, indicating their advantage over other methods.

Simulating the data and calculating MSE asymptotically in case of the variance of treatment variance

In this section, by changing the true variance of the treatment variable, its effect on the efficiency of variance estimation methods is investigated and MSE is asymptotically examined. The sample size starts from 50 and ends at 1100. The following table presents the results:

Table 2. MSE results by changing the variance of the treatment variable

N=1100	*N=650*	*N=450*	*N=250*	*N=50*	
			Var=1		
152.9716	171.2781	128.879	333.5097	377.0785	MSE(AI)
0.012475	0.018519	0.039859	0.26015	2.009331	MSE(NORM)
39.69625	194.4715	92.82319	693.1792	450.4136	MSE(bootstrap)
0.001067	0.013243	0.003224	0.608237	1.056548	MSE(jackknife)
1157.497	999.9874	1388.192	890.7482	0.001222	MSE(moutofn)
			Var=3		
353.3888	303.2482	207.9898	508.6827	485.6425	MSE(AI)
0.02288	0.022891	0.042301	0.148902	2.155842	MSE(NORM)
91.86069	154.2267	129.7413	958.2828	1805.274	MSE(bootstrap)
0.003128	0.007423	0.030791	0.471797	7.441926	MSE(jackknife)
945.0743	1751.966	636.2644	1515.68	0.032443	MSE(moutofn)
			Var=5		
822.8534	347.5849	802.0943	1446.203	1247.2	MSE(AI)
0.051176	0.07356	0.108614	0.37527	4.371456	MSE(NORM)
899.2683	299.3553	725.8821	4312.976	2308.827	MSE(bootstrap)
0.021144	0.001845	91.32544	560.8169	24.59164	MSE(jackknife)
1924.086	3388.587	2563.797	4607.964	0.030104	MSE(moutofn)
			Var=8		
1149.558	2737.275	1079.18	1750.589	2331.255	MSE(AI)
0.089751	0.20159	0.27377	0.389284	9.789377	MSE(NORM)
632.5505	1390.685	1174.216	1601.634	3128.942	MSE(bootstrap)
0.010748	0.12584	0.042169	0.15351	8.172749	MSE(jackknife)
6172.71	9851.1	10472.18	4904.871	0.087618	MSE(moutofn)
			Var=10		
2578.118	1890.971	3872.348	4053.145	8561.868	MSE(AI)
0.245034	0.315959	0.458182	1.229395	34.74394	MSE(NORM)
2880.444	1231.555	4291.435	4174.795	20226.44	MSE(bootstrap)
0.138363	0.059269	0.275863	1.064264	184.5844	MSE(jackknife)
11811.97	15405.78	12801.79	10725.46	0.224664	MSE(moutofn)

***Source**: Research findings*

For better understanding, the simulation results are also shown graphically as follows.

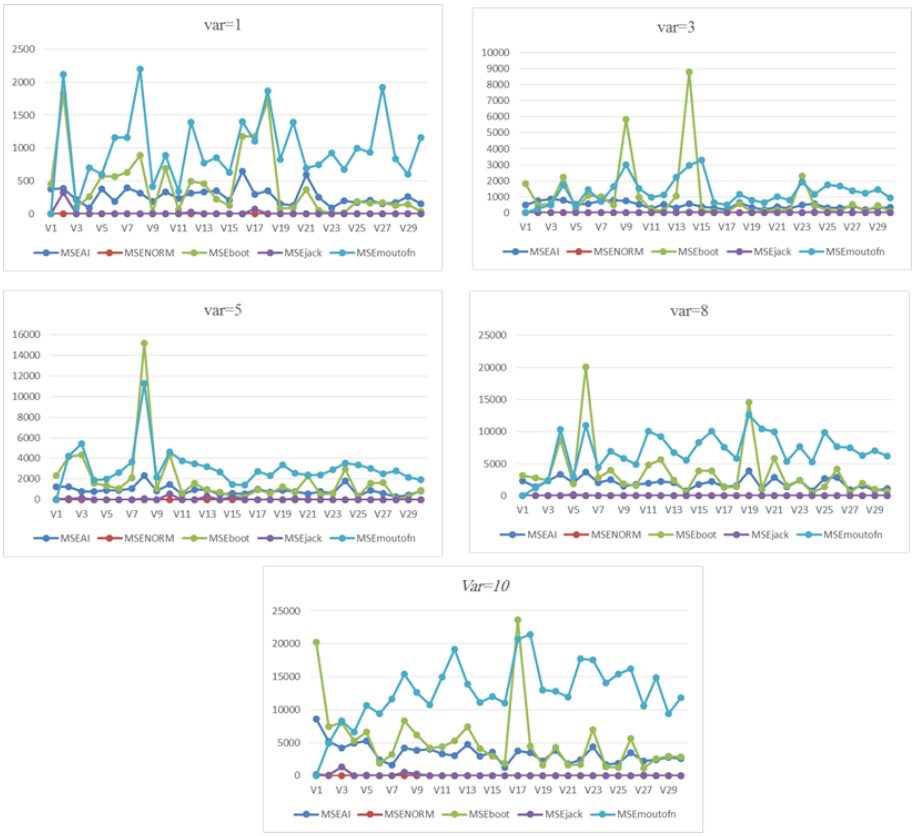

Figure 2. Diagrams of the results of the mean squares of error with the variance of the variable treatment

As can be seen from the results in the table and diagram above, MSE in the Bootstrap, Standard, AI, Jackknife, and Subsampling methods are decreasing asymptotically. Among these methods, MSE for the Standard and Jackknife methods are lower, indicating their advantage over other methods.

Simulating the data and calculating MSE asymptotically in case of changing the number of observations deleted in the following sample method

In this case, the first half of the data is deleted in the following sample method and the sample size starts from 100 observations and ends with 1000 observations. Then the number of deleted observations is reduced to one-third of the number of observations. This process will continue until the number of deleted observations reaches one-sixth of the total observations. MSE in this situation is reported in Table 3.

Table 3. MSE results by changing the number of observations deleted in the following sample method

N=1000	N=850	N=700	N=550	N=100	
			K=N/2		
28.65465	28.19384	36.65972	93.39186	58.78391	MSE(AI)
0.013195	0.019497	0.039933	0.155377	1.292855	MSE(NORM)
30.51024	37.41592	62.51748	205.018	296.2926	MSE(bootstrap)
8.01E-06	6.97E-06	2.93E-05	0.06227	0.860603	MSE(jackknife)
78.74072	79.64062	75.92326	84.01712	66.65849	MSE(moutofn)
			K=N/3		
30.50361	18.39525	36.7155	36.65011	78.83854	MSE(AI)
0.020355	0.018037	0.043026	0.045871	2.58793	MSE(NORM)
37.29474	26.62329	39.2228	49.4034	79.28877	MSE(bootstrap)
0.048309	0.000416	2.47E-05	0.003691	0.002617	MSE(jackknife)
127.8436	109.4234	112.4147	105.026	52.3588	MSE(moutofn)
			K=N/4		
18.28663	22.29392	23.72902	50.75837	42.88705	MSE(AI)
0.012609	0.019723	0.022909	0.077705	1.197849	MSE(NORM)
9.882499	19.07068	45.98614	52.18786	41.58316	MSE(bootstrap)
4.60E-06	8.46E-06	0.000229	0.000328	0.283022	MSE(jackknife)
138.8225	144.8462	135.6663	116.4668	69.59456	MSE(moutofn)
			K=N/5		
45.34659	20.25373	49.01397	43.5414	137.5781	MSE(AI)
0.021033	0.012599	0.04306	0.071808	4.271098	MSE(NORM)
59.82961	9.974066	61.52358	51.96502	330.7382	MSE(bootstrap)
3.63E-05	0.000475	0.000126	0.005871	0.088614	MSE(jackknife)
161.7997	192.5408	156.4408	186.3628	206.7466	MSE(moutofn)
			K=N/6		
38.94308	27.5658	36.55182	59.6292	45.55238	MSE(AI)
0.024246	0.021504	0.029613	0.059198	1.422849	MSE(NORM)
54.0015	24.03733	45.87355	135.4796	82.55372	MSE(bootstrap)
0.000116	1.38E-05	6.13E-05	0.191306	0.052862	MSE(jackknife)
241.4212	323.8279	226.2395	273.1131	151.55	MSE(moutofn)

***Source**: Research findings*

The results show that as in the previous case (change in the number of matches), MSE in the Bootstrap, Standard, AI, and Jackknife methods are

decreasing asymptotically, while this index was not decreasing in the following sample method. Among these methods, MSE for the Standard and Jackknife methods are lower. This indicates their advantage over other methods. To make it easier to view the results, the data in Table 3 and Figure 3 are shown graphically.

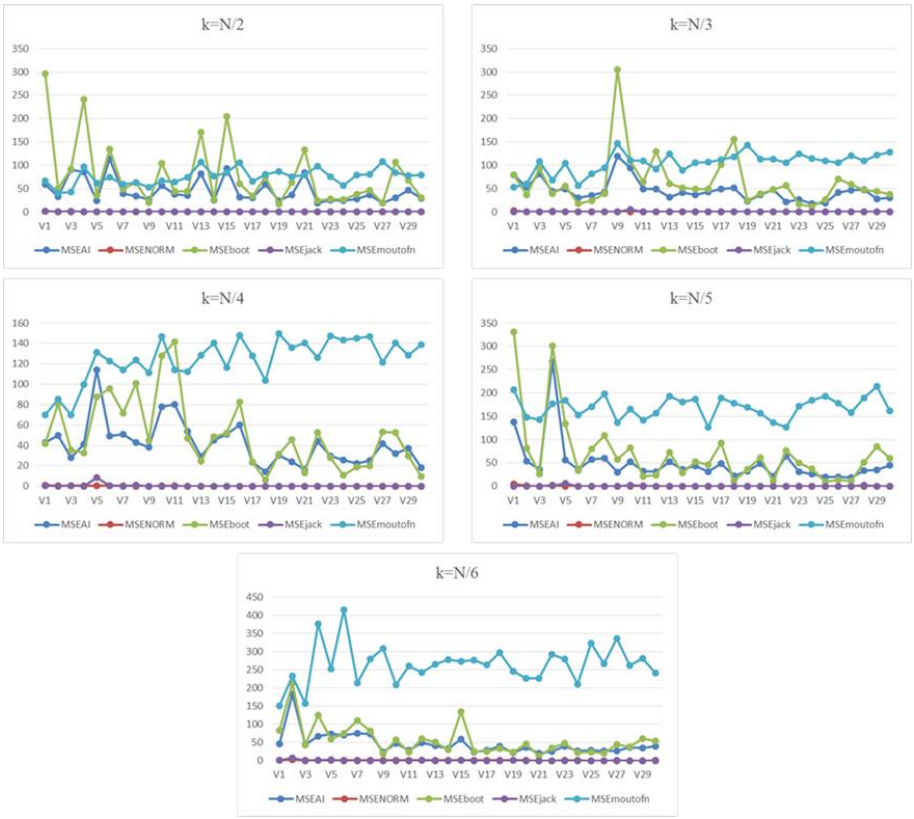

Figure 3. Graphs of the results of the mean squares of the error by changing the Deleted observation ratio of the following sample method

The results show that as in the previous case (change in the number of matches), MSE in the Bootstrap, Standard, AI, and Jackknife methods are decreasing asymptotically, while this index was not decreasing in the following sample method. Among these methods, the average error squares for the Standard and Jackknife method are lower, which indicates their advantage over other methods. To make it easier to view the results, the data in Table 4 and Figure 4 are shown graphically.

Table 4. MSE results by changing the number of repetition of the subsamples

N=1100	*N=650*	*N=450*	*N=250*	*N=50*	
			$B_{subsample}=50$		
248.7356	134.9739	107.9964	330.8653	529.6412	MSE(AI)
0.022047	0.037957	0.043224	0.087817	6.100182	MSE(NORM)
49.67163	17.01103	22.61767	47.64819	860.8127	MSE(bootstrap)
0.000429	9.92E-05	0.000404	0.002	2.073773	MSE(jackknife)
2045.959	1484.815	716.8665	473.4452	0.014953	MSE(moutofn)
			$B_{subsample}=100$		
178.9319	465.2119	792.3518	328.7763	482.7729	MSE(AI)
0.013606	0.04056	0.127905	0.225925	1.79902	MSE(NORM)
20.21442	438.1554	1248.376	275.9289	266.0861	MSE(bootstrap)
0.000345	0.000112	261.6429	0.125222	3.561777	MSE(jackknife)
958.0219	1304.061	1197.516	885.6246	0.011384	MSE(moutofn)
			$B_{subsample}=200$		
360.2601	414.3336	533.84	391.4492	560.8724	MSE(AI)
0.024227	0.025359	0.043455	0.125757	4.209544	MSE(NORM)
117.9272	380.7331	1175.746	178.7558	1517.231	MSE(bootstrap)
0.001051	0.008098	0.001217	0.014072	9.885017	MSE(jackknife)
1116.453	881.8515	1344.729	1288.892	0.013012	MSE(moutofn)
			$B_{subsample}=500$		
344.0046	320.3356	279.4204	833.2824	427.6695	MSE(AI)
0.026119	0.028842	0.050406	0.215308	2.9543	MSE(NORM)
821.8234	181.6626	110.0391	11581.5	589.4341	MSE(bootstrap)
0.025523	0.004476	0.004756	984.4382	0.161917	MSE(jackknife)
2573.406	1930.434	849.7428	4458.238	0.013819	MSE(moutofn)
			$B_{subsample}=1000$		
341.9951	462.0875	358.6715	323.2287	351.3017	MSE(AI)
0.019512	0.04287	0.044341	0.105567	1.446269	MSE(NORM)
317.2949	433.2199	79.74541	390.0648	439.5515	MSE(bootstrap)
0.00368	0.039303	0.000653	0.070336	0.894009	MSE(jackknife)
1323.612	1875.089	1471.801	825.8653	0.015939	MSE(moutofn)

Source: *Research findings*

For better understanding, the simulation results are also shown graphically as follows:

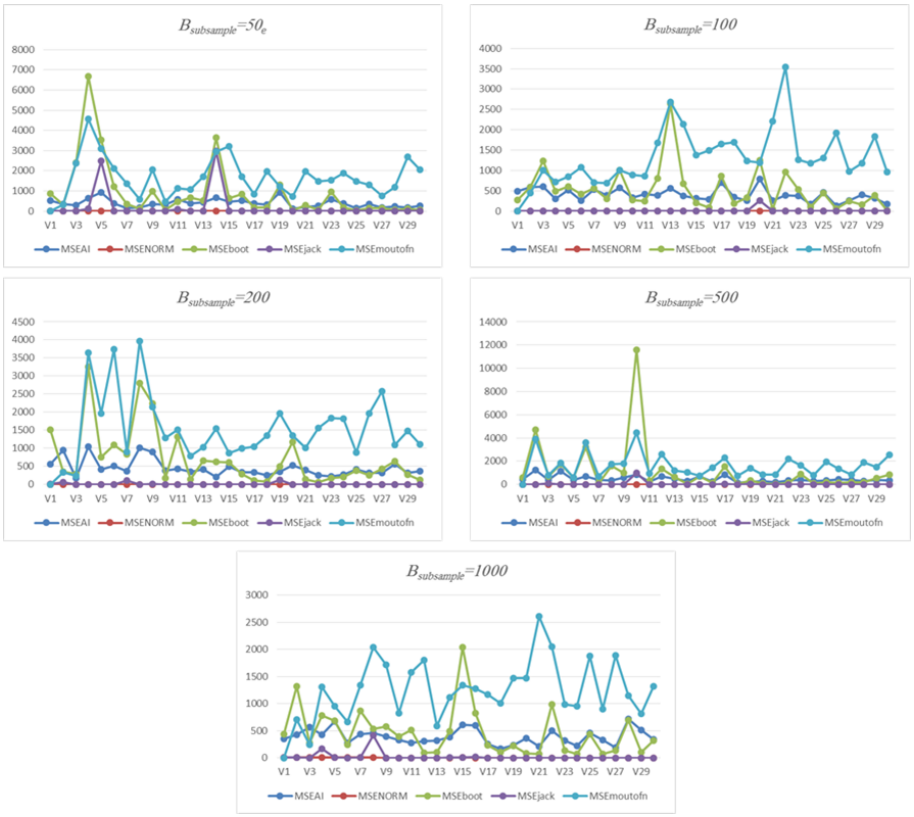

Figure 4. Graphs of the results of the mean squares of the error by changing the number of repetitions of the following sample method

As can be seen from the results of Table 4 and Figure 4, the average squares of error in the Bootstrap, Standard, AI, and Jackknife methods are decreasing asymptotically, while this index was not decreasing in the subsampling method. Among these methods, MSE for the normal method and the Jackknife method are lower, indicating their advantage over other methods.

Simulating the data and calculating MSE asymptotically in the case of changing the number of Bootstrap repetitions

MSE in this part of the simulation is checked by changing the Bootstrap repetitions and asymptotically. The sample size starts at 100 and ends at 1100. Table 5 presents the results:

Table 5. MSE results by changing the number of Bootstrap repetitions

N=1100	*N=650*	*N=450*	*N=250*	*N=50*	
			$B_{bootstrap}=500$		
477.0239	179.502	418.8944	228.8409	1112.115	MSE(AI)
0.033443	0.027131	0.09247	0.224875	3.648499	MSE(NORM)
2428.812	519.8813	751.6481	221.0139	1105.008	MSE(bootstrap)
0.007517	0.365821	0.261751	0.049491	298.802	MSE(jackknife)
3896.132	1659.63	2823.532	1111.028	0.011265	MSE(moutofn)
			$B_{bootstrap}=1000$		
236.6235	1058.463	673.8688	395.9643	445.5329	MSE(AI)
0.02287	0.095758	0.044192	0.187133	2.030951	MSE(NORM)
41.53934	705.7493	1507.854	653.0617	1166.436	MSE(bootstrap)
0.001451	143.2814	0.011571	219.4277	0.888042	MSE(jackknife)
919.0762	1548.031	2852.903	1705.596	0.011035	MSE(moutofn)
			$B_{bootstrap}=2000$		
341.8669	526.0844	687.2572	413.6001	693.246	MSE(AI)
0.015833	0.036965	0.105537	0.242459	3.077611	MSE(NORM)
113.6102	1519.746	3053.295	322.3889	1227.587	MSE(bootstrap)
0.000563	0.073821	796.3424	0.19781	0.276324	MSE(jackknife)
1384.015	3101.47	2182.227	768.2862	0.043175	MSE(moutofn)
			$B_{bootstrap}=6000$		
674.5495	258.2441	411.0071	515.3442	819.3827	MSE(AI)
0.051193	0.027934	0.047066	0.207452	2.639941	MSE(NORM)
574.0718	130.9901	44.15289	310.139	420.7778	MSE(bootstrap)
0.011961	0.003287	0.001366	0.020396	3.943456	MSE(jackknife)
1319.515	1502.019	718.3087	1517.604	0.011726	MSE(moutofn)
			$B_{bootstrap}=10000$		
391.4366	216.6119	313.5653	319.2639	582.3094	MSE(AI)
0.021078	0.032569	0.052911	0.117972	2.305478	MSE(NORM)
135.0861	104.3346	74.99605	916.1145	5910.211	MSE(bootstrap)
0.104926	0.002315	0.000672	205.7097	55.62754	MSE(jackknife)
1103.474	1338.889	730.6563	1623.122	0.020915	MSE(moutofn)

***Source**: Research findings*

For better understanding, the simulation results are also shown graphically as follows:

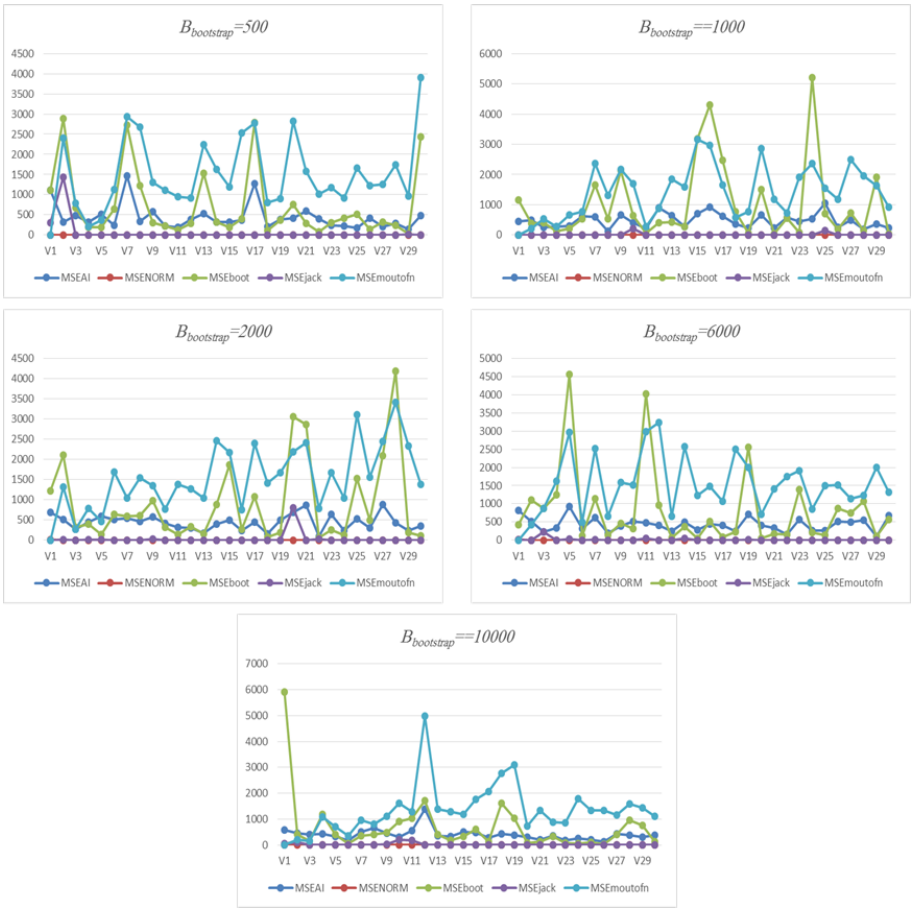

Figure 5. MSE results by changing the bootstrap repetitions

As the results in the above table and diagram shows, MSE in the Bootstrap, Standard, AI, Jackknife, and Subsampling methods are decreasing asymptotically. Among these methods, MSE for the Standard Jackknife methods is lower, indicating their advantage over other methods.

Simulating the data and calculating MSE asymptotically when the sample size is small.

In this section, small samples are examined and simulated, the efficiency of variance estimation methods is examined, and MSE is investigated asymptotically. The sample size starts at 50. Two observations are added each time until it will reach 110 observations. Tables 6 shows the results:

Table 6. MSE results in the small sample

N=110	N=100	N=90	N=80	N=70	N=60	N=50	
452.717	840.1775	547.3737	564.1866	274.6258	1335.324	1112.416	MSE(AI)
0.524775	0.912142	0.87638	1.700162	0.669477	3.052585	3.927309	MSE(NORM)
694.3893	11653.73	2367.416	1607.533	227.0527	358.2389	621.1525	MSE(bootstrap)
0.205842	2468.557	2.472419	40.2916	0.821282	2.310499	9.393599	MSE(jackknife)
1448.701	5491.874	3540.52	1750.472	400.8739	483.5897	470.9005	MSE(moutofn)

***Source**: Research findings*

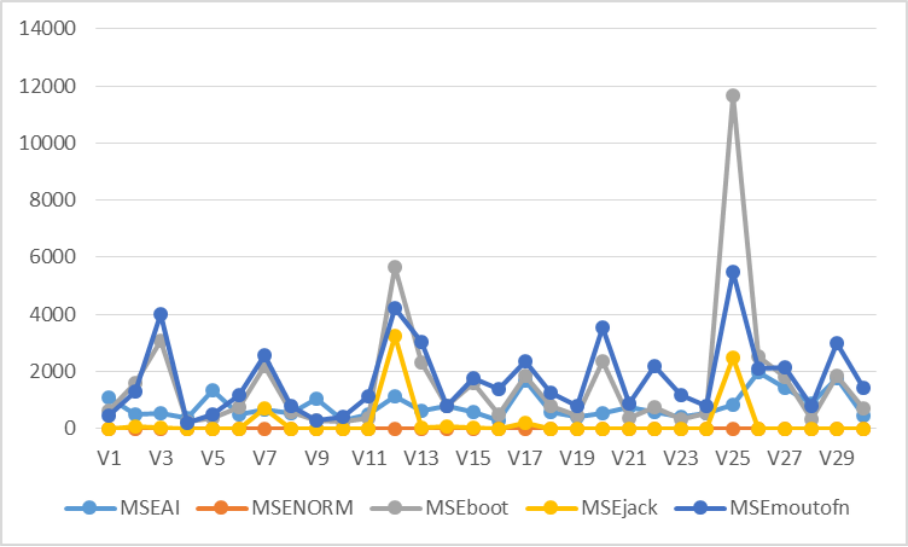

Figure 6. MSE results in the small sample

The results indicate that in small samples, the efficiency of Jackknife variance estimators and the Standard method is better than other methods.