Bootstrap Confidence Intervals for the Estimation of Average Treatment Effect on Propensity Score

Xia Peng

China University of Mining and Technology, Beijing 100083, China

E-mail: pengxia0715@yahoo.cn

Ping Jing

China University of Mining and Technology, Beijing 100083, China

E-mail: jping@cumtb.edu.cn

Zi'ou Feng

University of Exeter Business School, EX4 4PU, UNITED KINGDOM

E-mail: zf205@exeter.ac.uk

Funded by collage st. innovative projects

Received: January 21, 2011 Accepted: February 10, 2011 doi:10.5539/jmr.v3n3p52

Abstract

Causal inferences on the average treatment effect in observational studies are always difficult problems because the distributions of samples in the two treatment groups can not be observed at the same time, and the estimation of the treatment effect is often biased.In this paper, the propensity score and the propensity score subclassification, selected from several methods, are used to assess the treatment effect.The estimation of the average treatment effect give the Bootstrap confidence intervals. Simulation studies are inducted for the continuous samples in normal distribution and the mixed samples of discrete and continuous type.

Keywords: Average treatment effects, Propensity score, Propensity score subclassification, Maximum likelihood estimation, Bootstrap confidence intervals

1. Introduction

Causal inferences on the average treatment effect in observational studies are difficult problems because the effect could be confounded with the covariates whose distributions differ systematically in the two treatment groups, and a direct estimation of the treatment effect is often biased. Propensity score method has been shown to be an effective way to reduce this bias in the point estimation of the average treatment effect. However, we have not been well developed the inference procedures concerning this average treatment effect. A generally used approach is to stratify the data based on the estimated propensity scores and carry out the desired inferences as it is in a stratified random sample.

However, the validity of such procedures, is rather questionable. As the subclassification is based on the propensity scores estimated from a common logistic model, the responses within each subclass and between the subclasses are not likely to be independent. Meanwhile, the estimation of the unknown propensity scores also presents an addition source of variation, which will affect the variance estimate in the inference.

We introduce a Bootstrap confidence interval that takes into account the dependent structure of the propensity score stratified data in this paper, as well as the extra variation arisen from the propensity score estimation, under an assumption that the measured covariates can be balanced within all the subclasses based on estimated propensity scores.Different from the current methods, this procedure does not require an estimation of the variance quantity on the purpose of inference. Nor does it assume any specific distribution for the pivotal statistic used in the traditional confidence interval construction.

1.1 Proposition of the problem

Let P be a population from which we have a random sample consisting of N units.For each unit i in the sample, $i = 1, \cdots, N$, let Z_i be a binary treatment assignment variable so that unit i receives some control treatment if $Z_i = 0$, and unit i receives some treatment if $Z_i = 1$. For example, we may toss a coin and let unit i receives the treatment if the Head appears.The coin may or may not be biased.For an unbiased coin we have

$$pr(Z_i = 1) = pr(Z_i = 0) = \frac{1}{2}, i = 1, \cdots, N \qquad (1)$$

This is the case, for instance, when we randomly divide the N units into two groups, one for treatment and the other for control.For many problems in medicine and economics involving observational studies or evaluation studies, the above assumption on randomization is usually not realistic. In these cases, typically, the assignment probabilities vary with individuals.It is useful to think that they depend on some extraneous individual characteristics.

In this paper we shall consider the simple case where each unit i has a scalar outcome depending on the assignment variable Z_i.We let $Y_i(1)$ denote the outcome if $Z_i = 1$, that is, unit i is under treatment;and $Y_i(0)$ denote the outcome if $Z_i = 0$, or unit i is under control. For each unit i, the following quantities are assumed to be observable:

(1) a covariate vector $X_i = (X_{i1}, \cdots, X_{ip})^T$

(2) an assignment variable Z_i, which is correlated with correlated with X_i

(3) the outcome variable

$$Y_i = Y_i(0) + Z_i[Y_i(1) - Y_i(0)] \tag{2}$$

The quantity $Y_i(1) - Y_i(0)$ is known as the causal effect for unit i caused by the treatment.The primary focus of this paper concerns the construction of good confidence intervals for θ, the average treatment effect over population defined by

$$\theta = E[Y_i(1) - Y_i(0)] \tag{3}$$

We can also write

$$\theta = E[g(X_i)] \tag{4}$$

noting that using the conditional mean treatment effect

$$g(X_i) = E[Y_i(1) - Y_i(0)|X_i] \tag{5}$$

Actually, there exists a problem that we want $Y_i(1)$ and $Y_i(0)$ at the mean time when estimating the average treatment effect.Throughout this paper we shall make the following important assumptions on conditional independence(Rosenbaum and Rubin, 1983) to solve this difficulty.

ASSUMPTION 1.1 *The treatment assignment is said to be strongly ignorable if $Y_i(1)$ and $Y_i(0)$ are independent of Z_i conditional on*

(1) $Z \perp (Y_i(1), Y_i(0))|X$

(2) $0 < P(Z = 1|X = x) < 1$

Using the assumption we may express $g(X_i)$ as

$$g(X_i) = E[Y_i(1)|Z_i = 1, X_i] - E[Y_i(0)|Z_i = 0, X_i] \tag{6}$$

This last expression seems to suggest that $g(X_i)$ depends on the dimension of covariate X_i.Serious practical problems occur when the dimension of X_i is large, which is usually the case in many applications involving comparisons of treatment effects.To reduce the dimension of X_i, we shall use the idea of propensity score proposed by Rosenbaum and Rubin(1983).

DEFINITION 1.1 (PROPENSITY SCORE)*The propensity scores are probabilities for receiving the treatment conditional on the covariates, that is*

$$e(X_i) = pr(Z_i = 1|X_i), i = 1, \cdots, n \tag{7}$$

By the assumptions 1.1 and definition 1.1 we get

$$Z_i \perp (Y_i(1), Y_i(0))|e(X_i) \tag{8}$$

and then draw

$$\theta = E[E[Y_i(1) - Y_i(0)|e(X_i)]] = E[E[Y_i(1)|Z_i = 1, e(X_i)] - E[Y_i(0)|Z_i = 0, e(X_i)]] \tag{9}$$

It can be directly obtained the unbiased estimate of $E[Y_i(1)|Z_i = 1, e(X_i)]$ and $E[Y_i(0)|Z_i = 0, e(X_i)]$ if the propensity score is known, then we naturally obtain the effect estimate of the project, as well as the treatment group.We will apply the Bootstrap method to estimate the confidence interval of θ.

This paper is organized as follows:In section 2, we briefly introduce the maximum likelihood method to estimate propensity score. In Section 3 we describes two methods to estimate the average treatment effect using propensity score. In

Section 4 we report the several methods of estimation for Bootstrap confidence intervals. In Section 5, we illustrate the estimation of the average treatment effect for confidence intervals with simulations of different methods, and make a brief discussion about the potential utility of the method in practice.

2. Maximum likelihood estimation of propensity score

Since the propensity score $e(X)$ is rarely known, we usually estimate the unknown propensity scores via a logistic model

$$\log \frac{pr(Z=1|X=x)}{1-pr(Z=1|X=x)} = x^t\beta \tag{10}$$

As (Z, X) follows the logistic model, we drawn a random sample $(1, X_1), \cdots, (1, X_{n_1}), (0, X_{n_1+1}), \cdots, (0, X_{n_1+n_2})$, which contains $n_1 + n_2$ vectors from the population (Z, X). Then the logarithmic likelihood function is

$$l(\beta) = \sum_{i=1}^{n_1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}) - \sum_{i=1}^{n_1+n_2} ln(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq})) \tag{11}$$

Derivate to parameters like $\beta_0, \beta_1, \cdots, \beta_q$ respectively.Using Newton-Raphson iterative method to seek the maximum of $l(\beta)$.Maximum likelihood estimation $\hat{\beta}_{mle}$ of β is available and then estimate the propensity score $e(X)$ as

$$\hat{e}(x) = \frac{\exp(\hat{\beta_0} + \hat{\beta_1}x_1 + \cdots + \hat{\beta_q}x_q)}{1 + \exp(\hat{\beta_0} + \hat{\beta_1}x_1 + \cdots + \hat{\beta_q}x_q)} \tag{12}$$

3. Different methods in estimating average treatment effect

3.1 Propensity Score Method

As described in the introduction we can directly use estimated propensity score considering covariates were known and the estimation is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n_1}[\frac{Y_i Z_i}{\hat{e}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{e}(X_i)}] \tag{13}$$

Where $\hat{e}(X_i)$ is an estimate of the propensity score $e(X_i)$.

3.2 Propensity Score Subclassification

Using the estimated propensity scores we stratify all the subjects into K subclasses so that the estimated propensity scores have similar values within each subclass. Let J_{im} represent the mth subclass' characteristic function. One way of this method is to divide the unit into M blocks. The boundary of each block is $\frac{m}{M}, m = 1, \cdots, M-1$, so $J_{im} = I\{\frac{m-1}{M} < e(X_i) < \frac{m}{M}\}$ $m = 1, \cdots, M$ has N_{wm} observations in each subclass, where $N_{wm} = \sum_i I\{Z_i = z, J_{im} = 1\}$.The estimation of average treatment effect of a given group is

$$\hat{\theta}_m = \frac{1}{N_{1m}}\sum_1^N J_{im}Z_iY_i - \frac{1}{N_{0m}}\sum_1^N J_{im}(1-Z_i)Y_i \tag{14}$$

Then the population average treatment effect is

$$\hat{\theta}_{block} = \sum_{m=1}^{M}\hat{\theta}_m\frac{N_{1m}+N_{0m}}{N} \tag{15}$$

4. Bootstrap method in estimating confidence intervals

Bootstrap method is a better approximation method to estimate the interval structure, and we will introduce two common methods in Bootstrap interval estimation.

4.1 Percentile Method

Percentile method is also called Bootstrap-p method.Let $\theta = \theta(F)$ and $\hat{\theta} = \hat{\theta}(F_n)$, where $\hat{\theta}$ is the estimate of θ.We want to find the $1 - 2\alpha$ confidence interval of θ.If $P(\hat{\theta}(F_n) \leq X) = G$, then $P(G^{-1}_{1-\alpha} < \theta < G^{-1}_{\alpha}) = 1 - 2\alpha$. We can derive confidence interval directly by the formula if G is known.If G is unknown, the empirical distribution function G_n of G can be substituted. Algorithm is as follows:

1. Sample $X_1, X_2, \cdots, X_n$ from F,and calculate $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$

2. Draw Bootstrap data sets $X_1^*, X_2^*, \cdots, X_n^*$ from F_n generated by $X_1, X_2, \cdots, X_n$

3. Calculate $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*)$

4. Compute the Bootstrap replication $\hat{\theta}_b^* = \hat{\theta}(X_{b1}^*, X_{b2}^*, \cdots, X_{bn}^*)$, $b = 1, 2, \cdots, B$ as step 2 and 3,and obtain $\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq \cdots \leq \hat{\theta}_{(B)}^*$ in the ordered list of B replications of $\hat{\theta}_b^*$

5. The required confidence interval is $(\hat{\theta}_{[\alpha B]}, \hat{\theta}_{[(1-\alpha)B]})$

4.2 BCa Method

BCa method is the abbreviation standing for bias-corrected and accelerated method, which is defined more complicated than the percentile method.

1. Obtain $B\ \hat{\theta}_b^*,\ b = 1, 2, \cdots, B$ as described in the previous section

2. The bias-correction constant z_0 is computed as

$$z_0 = \Phi^{-1}(\frac{\sharp\{\hat{\theta}_b^* < \hat{\theta}\}}{B}) \tag{16}$$

where Φ^{-1} is the inverse function of the standard normal distribution.$\sharp\{\hat{\theta}_b^* < \hat{\theta}\}$ denotes the times of $\hat{\theta}_b^* < \hat{\theta},\ b = 1, 2, \cdots, B$

3. Compute the acceleration parameter a.There are various ways to compute the acceleration parameter a.The easiest to explain is given in terms of the jackknife values.$G_{(i)}$ denotes the sample with the ith observation removed form the original sample G, and $\hat{\theta}_{(i)} = \hat{\theta}(G_{(i)})$, then

$$a = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{\frac{3}{2}}} \tag{17}$$

where $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{(i)}}{n}$

4. The resulting $1 - 2\alpha$ confidence interval is defined as

$$BCa:\ (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}) \tag{18}$$

where $\alpha_1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(a)}}{1 - \hat{a}(\hat{z}_0 + z^{(a)})}),\ \ \alpha_2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-a)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-a)})})$, Here $\Phi(\cdot)$ is the standard normal cumulative distribution function and $z^{(a)}$ is the α percentile point of a standard normal distribution.

Notice that if $\hat{a}$ and $\hat{z}_0$ equal zero, then $\alpha_1 = \Phi(z^{(a)}) = \alpha$ and $\alpha_2 = \Phi(z^{(1-a)}) = 1 - \alpha$, so that the BCa interval is the same as the percentile interval.

We calculate the $100(1 - 2\alpha)\%$ BCa confidence interval of the average treatment effect following Hall and Martin.The bias-correction constant is computed as

$$\hat{d} = \Phi^{-1}(\frac{\sharp[\hat{\theta} < \hat{\theta}^b]}{B}) \tag{19}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The two-sample acceleration parameter is computed as

$$\hat{a} = \frac{1}{6}\hat{\sigma}^{-\frac{3}{2}}(n_t^{-2}\hat{r}_t - n_c^{-2}\hat{r}_c) \tag{20}$$

where $\hat{\sigma}^2 = \hat{\sigma}_{tjack}^2 n_t^{-1} + \hat{\sigma}_{cjack}^2 n_c^{-1}$; we use the jackknife variance estimates here from original sample $\hat{\sigma}_{tjack}^2$ and $\hat{\sigma}_{cjack}^2$ to estimate the unknown variances of the two treatment groups, according to the method of Efron and Tibshirani(1993);$\hat{r}_t$ and $\hat{r}_c$ are the sample skewnesses of the respective groups.

Sorting the Bootstrap treatment effect estimates into increasing order, $\hat{\theta}^{(1)} \leq \hat{\theta}^{(2)} \leq \cdots \leq \hat{\theta}^{(B)}$, the resulting $100(1 - 2\alpha)\%$ confidence interval is defined as $(\hat{\theta}^{[B(\hat{\beta}_{a(\alpha)})]}, \hat{\theta}^{[B(\hat{\beta}_{a(1-\alpha)})]})$,where $\hat{\beta}_{a(\alpha)} = \Phi\{\hat{d} + (\hat{d} + z_\alpha)[1 - \hat{a}(\hat{d} + z_\alpha)]^{-1}\}$ and $[x]$ is the largest integer less than or equals to x.

5. Simulation and discussion

We re-introduce the notation with the subscription reserved for the subject to better describe the proposed Bootstrap procedure: Let $n = n_t + n_c$ be the total number of subjects;(Y_i, X_i, Z_i) contain the response variable, the covariate vector for the true propensity model, and the treatment assignment for the ith subject, where $i = 1, \cdots, n$. Our procedure begins with the fitting of propensity model from using the original sample (X_i, Z_i) for $i = 1, \cdots, n$. We estimate the propensity score of each subject after fitting the model.

a) Obtain the average treatment effect using propensity score:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}[\frac{Y_i Z_i}{\hat{e}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{e}(X_i)}] \tag{21}$$

b) We stratify the subjects into K homogeneous subclasses based on these estimates. Then the post-stratification balance of each covariate is then examined.A point estimate for the average treatment effect is obtained as

$$\hat{\theta}_{block} = \sum_{m=1}^{M}\hat{\theta}_m \frac{N_{1m}+N_{0m}}{N} \tag{22}$$

We re-sample with replacement n_t treated and n_c control subjects separately from the treated and control subjects in the original sample for each bootstrap iteration. Let $(Y_{i'}^{(b)}, X_{i'}^{(b)}, Z_{i'}^{(b)})$ be the bth bootstrap, $i' = 1, \cdots, n$, $n = n_t + n_c$. By using the re-sampled data $(X_{i'}^{(b)}, Z_{i'}^{(b)})$, we re-fit the same logistic model and re-estimate the propensity score for each of the re-sample subjects,$\hat{e}_{i'}^{(b)}$. We then stratify the bootstrapped responses $Y_{i'}^{(b)}$, and compute the mean treatment effect, and denote it as $\hat{\theta}^b$.

We conducted two simulation studies to assess the finite sample performance of the proposed procedure.Firstly, we generate the covariates X.We consider a logistic regression propensity model with three covariates in our simulation:

1) There are continuous covariates X_1,X_2 and X_3.We compare the confidence intervals of percentile method and the BCa method simulated by propensity score.

To simulate different situations of the covariate distributions systematically in the two treatment groups, we generate the covariate deviates for the treatment and control groups separately: For the control group($Z = 0$), we assume that $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim N(0, \sigma_1^2)$ and $X_3 \sim N(0, \sigma_1^2)$;for the treatment group ($Z = 1$), we assume that $X_1 \sim N(d, \sigma_1^2)$, $X_2 \sim N(d, \sigma_1^2)$ and $X_3 \sim N(d, \sigma_1^2)$.

2) A continuous covariate X_1 and two binary covariates X_2 and X_3. We estimate the average treatment effect using percentile method and propensity score subclassification method and obtain the confidence interval by the two methods.

To simulate different situations of the covariate distributions systematically in the two treatment groups, we generate the covariate deviates for the treatment and control groups separately: For the control group($Z = 0$), we assume that $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim Bernoulli(p_{2c})$ and $X_3 \sim Bernoulli(p_{3c})$;for the treatment group ($Z = 1$), we assume that $X_1 \sim N(d, \sigma_1^2)$, $X_2 \sim Bernoulli(p_{2t})$ and $X_3 \sim Bernoulli(p_{3t})$.We would be able to simulate situations of varying level of differential covariate distributions by controlling d and the probabilities p_{2c}, p_{2t}, p_{3c} and p_{3t} in the Bernoulli distributions.

With the pseudo-random covariate deviates X and the treatment assignment Z, we obtain responses Y from a linear relationship $Y = Z\delta + X^t\beta + \varepsilon$ by giving values of δ and β and the independently generated normal errors $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Here, δ represents the true treatment effect. As covariates X have different distributions in the two treatment groups, the effect of the treatment δ can not be directly estimated from the response Y before the first adjusting for the effects of X. For simplicity, throughout the simulation we set the coeffcients of the covariates to be 1)$(\beta_1, \beta_2, \beta_3) = (0.5, 0.4, 0.4)$;2)$(\beta_1, \beta_2, \beta_3) = (0.5, 0.4, 1.5)$.

The other values of the parameters that we used in the simulation are listed in Table 1 and 2. It should be noted that several factors lead to the extend of confounding effects of X on Y, including: 1) numbers of $\beta_1, \beta_2, \beta_3$, which have direct effects on the level of confounding;and 2) the differential distributions of X in the two treatment groups, which affect the level of confounding indirectly.

For each parameter structure,we conduct 1000 simulations in an iterative model. We use 2000 Bootstrap to construct 95% and 90% confidence intervals within each iteration. The empirical coverage probabilities under different parameter configurations are reported to assess of the performance of the proposed procedure.In order to understand how the coverage property changes in different sample size situations, we consider sample sizes as $n_c = n_t = 500$, for each of the parameter settings in Table 1 and 2. The simulation results are reported in Table 3-1 and 3-2, the results of simulation 1, and Table 4-1 and 4-2, the results of simulation 2.$\bar{L}$, $\bar{U}$ denote the mean lower and upper confidence limit in 1000 simulation, and $E(\hat{\theta})$ is the mean estimate of θ.B-p means the Bootstrap-p method and B-S means Bootstrap stratified method.

Propensity score methodology has been applied to many clinical and epidemiological studies successfully since Rubin and Rosenbaum's early creative work.It has become a widely used tool to reduce the potential bias in treatment effect estimation during observational data analysis.We have proposed a Bootstrap method based on inference procedure for the treatment effect within the framework of propensity score and propensity score subclassification in this paper.

Our study shows that the proposed method provides valid causal inferences in large observational studies.It has several advantages in summary: First, it does not require a variance estimation. Our experience suggests that it is difficult to directly estimate the analytical derivation of the variance in the treatment effect estimate generally, if not entirely impossible.Secondly, it does not rely on any restrictively distributional assumption on the covariates.This method is particularly important in practice because there are rarely explanatory variables, all of which are normally distributed. Thirdly, the Bootstrap intervals consider the variation that arises from the estimation of propensity scores, and they accommodate the dependency among the responses both within and between subclasses due to the ordering structure introduced by the subclassification. Finally, it is relatively easy to implement the new Bootstrap procedure in most computing platforms.

Our simulation suggests that the empirical coverage of the procedure are reasonable. While the empirical coverage of the probabilities are below the nominal level (95% and 90%), they are closer to the nominal level when the sample sizes are greater than 1000 per group. The simulation results also show that the BCa confidence interval is slightly better than percentile method on the coverage probability, as well as the accuracy of estimated intervals. Propensity score subclassification is better than percentile method in the two aspects. The size of the treatment effect is an additive component in a linear model; δ only stands for a shift in the central locations between the two treatment groups when the responses are generated from this model.The simulation also shows that when all of the covariates are used in the logistic regression model to estimate the unknown propensity scores, the proposed method adjusts for the effects with the systematically different covariates quite effectively.

Although the first simulation results are promising, more extensive simulation studies are apparently needed to establish the operating characteristics of the proposed method for various practical data situations.In this case, the current simulation has several limitations: First, it only considers balanced designs while few observational data have balanced group sizes. For example, the sample $n_c = n_t = 500$ is used to reduce the imbalance.Second, the values range of the parameters used in the current simulation is still limited. For example, only values of β in linear relationship $Y = Z\delta + X^t\beta + \varepsilon$ are used to produce random responses. Since β has a direct effect on the level of confounding between the observed covariates X and the treatment assignment X, it is interesting to examine the performance of the proposed method under many different values of β.

A smaller value of β decreases the level of confounding when we hold other parameters constant(in the most extreme case of $\beta = 0$, we have $Y = Z\delta + X^t\beta + \varepsilon$, indicating no confounding effects from X. In addition, parameters p_{2c}, p_{2t}, p_{3c} and p_{3t} control the different distributions between the treatment groups for a set of pre-selected β values. The size of d, and the difference between p_{2c} and p_{2t}(or that between p_{3c} and p_{3t}) reflect the separation of covariate distribution between the two treatment groups.In the current simulation, we only consider one d value, and a limited number of binomial probabilities. We feel that further investigation is certainly necessary to have a more comprehensive understanding of the new method's operating characteristics according to these observations.

Our current work pays more attention on a re-sampling based on the approach for constructing a simple confidence interval of an unknown treatment effect.Several related issues have yet to be explored.Treatment effects estimated by other measures, such as Stratified Matching Method and Stratifying Regression Method, also necessary to be discussed.

References

Bradley Efron, Robert J. Tibshirani. (1993). *An introduction to the Bootstrap*. New York : Chapman & Hall .

Efron, B. (1985). Bootstrap confidence interval for a class of parametric problems. *Biometirka*, Vol. 72, No. 1. 45-58.

Hahn,J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66 , 315-332.

Hall, Martim,M. (1988). On the Bootstrap and two sample problemTreatment Effect. *NBER technical working paper*, No. 283. 179-192.

Paul R. Rosenbaum, Donald B. Rubin. (1983). The central role of the propensity score in observational studies for causal effects. *Biometirka*, Vol. 70, No. 1. 41-55.

WanZhu Tu, Xiao-Hua Zhou. (2003). A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification. *Health Services and Outcomes Research Methodology*, Vol. 3, No. 2. 135-147.

Table 1. The parameters used in simulation 1

	δ	d	σ_1	σ_ε
1	0.0	0.5	1.0	1.0
2	0.5	0.5	1.0	1.0
3	1.0	0.5	1.0	1.0
4	2.0	0.5	1.0	1.0

Table 2. The parameters used in simulation 2

	δ	d	σ_1	σ_ε	p_{2c}	p_{3c}	p_{2t}	p_{3t}
5	0.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4
6	0.5	0.5	1.0	1.0	0.3	0.7	0.8	0.4
7	1.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4
8	2.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4

Table 3-1. The results of simulation 1

	95%		90%		$E(\hat{\theta})$
$Y = Z\delta + X^t\beta + \varepsilon$	BCa	B-p	BCa	B-p	
1 $\delta = 0$	0.956	0.959	0.918	0.932	0.0170
2 $\delta = 0.5$	0.955	0.956	0.922	0.930	0.5156
3 $\delta = 1$	0.962	0.967	0.934	0.943	1.0143
4 $\delta = 2$	0.964	0.966	0.942	0.942	2.0116

Table 3-2. The results of simulation 1

	95%$\bar{L}$		95%$\bar{U}$		90%$\bar{L}$		90%$\bar{U}$	
Y	BCa	B-p	BCa	B-p	BCa	B-p	BCa	B-p
1	-0.32598	-0.39608	0.27252	0.31729	-0.29360	-0.36158	0.29122	0.33494
2	0.14824	0.11755	0.82780	0.81168	0.21355	0.10527	0.78369	0.76705
3	0.68547	0.61868	1.27195	1.31832	0.71072	0.64612	1.28670	1.33300
4	1.65617	1.58623	2.30806	2.36419	1.67832	2.32286	2.32286	2.37897

Table 4-1. The results of simulation 2

	95%		90%		$E(\hat{\theta})$	
Y	B-p	B-S	B-p	B-S	B-p	B-S
5 $\delta = 0$	0.972	0.949	0.953	0.892	-0.00172	-0.00148
6 $\delta = 0.5$	0.972	0.949	0.956	0.892	0.49842	0.49852
7 $\delta = 1$	0.974	0.947	0.958	0.895	0.99856	0.99852
8 $\delta = 2$	0.976	0.943	0.965	0.894	1.99884	1.99852

Table 4-2. The results of simulation 2

	95%$\bar{L}$		95%$\bar{U}$		90%$\bar{L}$		90%$\bar{U}$	
Y	B-p	B-S	B-p	B-S	B-p	B-S	B-p	B-S
5	-0.24714	-0.16471	0.24533	0.16471	-0.20785	-0.14095	0.20530	0.13782
6	0.23904	0.33266	0.76260	0.66471	0.28032	0.35905	0.71938	0.63782
7	0.72102	0.83266	1.28436	1.16471	0.76487	0.85902	1.28436	1.13782
8	1.67559	1.83266	2.33764	2.16471	1.72594	1.85305	2.28089	2.13782