Journal of Statistical Research *ISSN 0256 - 422 X*
2018, Vol. 52, No. 2, pp. 187-200

BOOTSTRAP BIAS CORRECTION FOR AVERAGE TREATMENT EFFECTS WITH INVERSE PROPENSITY WEIGHTS

GUBHINDER KUNDHI*
Department of Economics, Memorial University
230 Elizabeth Avenue, St. John's, NL, A1C 5S7, Canada
Email: gkundhi@mun.ca

MARCEL VOIA
Department of Economics, Carleton University
Loeb Building, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada and
Laboratoire d'Économie d'Orléans, Faculté de Droit d'Économie et de Gestion
Rue de Blois - BP 26739, 45067 ORLÉANS Cedex 2
Email: Marcel.Voia@carleton.ca

SUMMARY

The estimated average treatment effect in observational studies is biased if the assumptions of ignorability and overlap are not satisfied. To deal with this potential problem when propensity score weights are used in the estimation of the treatment effects, in this paper we propose a bootstrap bias correction estimator for the average treatment effect (ATE) obtained with the inverse propensity score (BBC-IPS) estimator. We show in simulations that the BBC-IPC performs well when we have misspecifications of the propensity score (PS) due to: omitted variables (ignorability property may not be satisfied), overlap (imbalances in distribution between treatment and control groups) and confounding effects between observables and unobservables (endogeneity). Further refinements in bias reductions of the ATE estimates in smaller samples are attained by iterating the BBC-IPS estimator.

Keywords and phrases: Average Treatment Effects, Propensity Score, Bias Correction, Bootstrap

JEL codes: C1, C150, C9

1 Introduction

Observational studies and the associated methods of treatment effects are used more often to measure differences between groups of individuals. In these studies, when confounding between what is called the treatment effect and the observables from the data is present, the methods employed are biased.

In the treatment literature a few popular methods are used to deal with confounding: matching, covariate/regression adjustment, and stratification, see Hennekens and Buring (1987). These methods while popular can be subject of misspecification. In particular the matching method may fail

* Corresponding author

when relevant matching covariates are not available and in this case the presence of unmeasured confounding (unobserved heterogeneity) can bias the results. The regression adjustment based methods on the other hand (linear, logistic for example) depend on correct specification of the model relating the covariates to the outcome and in case the distributions of the covariates in the treatment and control groups are different corrections are hard to implement. The stratification method can also fail if the strata in the two groups does not contain information about subjects from either of the two groups. These cases of non-informative strata can happen in models with a large number of strata for example, models with a large numbers of covariates.

One important statistical tool that can be used as a basis for matching, stratification, regression adjustment and data reduction is the propensity score (PS). The advantage of using the PS method for matching and stratification rely on the fact that it can control for covariates and in this case, both matching and stratification can be done on a single scalar variable. The use of PS in observational studies therefore became popular in the estimation of the average treatment effects (ATE) and was accepted as a tool to adjust for confounding (reduce the bias of the estimated ATE) when it was present, see Rosenbaum (1995), Joffe and Rosenbaum (1999) and Lunceford and Davidian (2004). The literature that looks at the misspecification of the propensity score is thin, while the applications are skyrocketing. There are two strains in the literature, one focusing on the testing side and the other one on assessing the issues of misspecification. In particular the paper of Shaikh et al. (2009) proposed a PS score specification test build on some restrictions between the estimated densities of the treatment and comparison groups. Lee (2007) proposed a regression based method to detect a misspecified propensity score in the case when the propensity score model is under-specified as well as in the case when a relevant covariate is inadvertently excluded. While Lee (2013) proposed a balancing test for the PS score and shows that a nonparametric version of the balancing test is working better, he also shows that balancing tests are of little utility if the conditional independence assumption underlying matching estimators is not fulfilled. On assessing the misspecification issue, Millimet and Tchernis (2009) suggest that overspecifiyng the propensity score model can be beneficial. Using a similar argument Rubin (2009) claims that not controlling for an observed covariate is bad practical advice and argues that even if one were to condition on more covariates the result would be inefficient, but not biased. In another note Clarke et al. (2015) and Clarke et al. (2016) show that while the standard practice when estimating a treatment effect is to include all available pre-treatment variables, this approach may not always be optimal when the goal is bias reduction. In particular they show that conditioning on an observed covariate can increase bias when there is a confounding effect between two covariates. The work of Imai and Ratkovic (2014) introduce a covariate balancing propensity score methodology based on the GMM or Empirical Likelihood to improve the properties of the propensity score. In the case of matching estimators, Abadie and Imbens (2011) proposed a nonparametric bias correction method that makes the matching estimator consistent and asymptotically normal but less efficient than the regression adjustment and weighting estimators. Zhao et al. (2009) use a PS approach named genomic propensity score (GPS) to correct for bias due to population stratification using genetic and non-genetic factors. For inverse probability weighting using PS weights, Peng and Feng (2011) employed a bootstrap method that takes into account the dependent structure of the propensity score stratified data to construct confidence

intervals for the estimates of ATE.

As seen from the discussion above, misspecification of the PS score due to omitted relevant variables, balancing or confounding effects (endogeneity) has important implications on the bias of the PS score, which translates to bias estimation of the treatment effects. This paper addresses the problem of misspecification of the PS due to: omitted variables, endogeneity and lack of overlap by doing a bootstrap bias correction to the inverse probability weighting estimator (BBC-IPS), which has misspecified the PS score weights. The misspecifications of the PS score weights will induce bias in the estimation of the ATE.

Kim and Yixiao (2016) have shown that the bootstrap bias correction for maximum likelihood (ML) estimators is an effective tool in reducing the bias of the fixed effects estimator and in improving the coverage accuracy of the associated confidence interval in nonlinear panels. It is well known that fixed effects estimators for nonlinear panels with short time periods (the case considered by Kim and Yixiao (2016)) suffers from inconsistency because of the incidental parameters problem. Even if the time series dimension grows at a slower rate than the cross-section dimension the ML estimator is asymptotically biased and therefore the associated confidence intervals have a large coverage error.

The BBC-IPS estimator in our paper is based on a nonlinear function of the PS weights, which are estimated by maximum likelihood, which is a model similar to the one analyzed by Kim and Yixiao (2016) and is also estimated via maximum likelihood. We use their findings about the bootstrap bias correction for the maximum likelihood (ML) estimators obtained from nonlinear models to motivate that a bias-correction procedure can reduce the bias for propensity score estimators. The misspecification cases that we propose to investigate induce bias in the estimated propensity score, therefore having a method that reduces the bias of the estimated propensity score would be very useful in applications that use propensity scores. Although bootstrap bias corrections are well established their application to propensity scores methods is rare in the literature.

We show in our simulations that a bias-correction approach via a bootstrap procedure for the inverse propensity score (BBC-IPS) performs well. It helps in correcting the finite sample bias of the ATE for all these types of misspecifications of the propensity score (PS) and different degrees of endogeneity (correlation between one of the covariates and the unobeservable) and choices of distributions for the unobservables. The BBC-IPS estimator is iterated in smaller samples for further refinements in the bias correction of the ATE estimates and this procedure works well in reducing bias for all types of mispecifications considered in our paper.

The rest of the paper is organized as follows: Section 2 discusses the methodology, Section 3 presents the simulation exercise results and Section 4 concludes.

2 Bootstrap Bias Correction for the Inverse Propensity Score Weighting Estimator

The PS based method is widely and increasingly used in the estimation of ATE for its attractive properties. Two important assumptions are required for the validity of this method. In particular the assumptions of ignorability and overlap are the pillars of this method.

Consider a sample $\{y_i, z_i, \mathbf{x}_i'\}$ with $i = 1, \ldots, N$ individuals, where y is the outcome variable of interest, z is an indicator for the treatment and $\mathbf{x}'$ is a set of covariates. Now suppose that observation i with characteristics $\mathbf{x}_i$ receives a treatment then an outcome y_{1i} is observed. If the ith individual with characteristics $\mathbf{x}_i$ did not receive a treatment then we observe the outcome y_{0i}. Let y_i denote the observed outcome and z_i denote the indicator of whether individual i received treatment or not, then the observed outcome can be written as:

$$y_i = z_i y_{1i} + (1 - z_i) y_{0i}.$$

If the sample is drawn from the joint distribution of $(y, z, \mathbf{x}')' \in \mathcal{Y} \times [0, 1] \times \mathcal{X}$, the following assumptions are considered for the identification of treatment effects:

1. Ignorability: (y_1, y_0) and z are independent conditional of $\mathbf{x}$.
2. Overlap: For all $\mathbf{x} \in \mathcal{X}$, $0 < \Pr\{z = 1|\mathbf{x}\} = p(\mathbf{x}) < 1$.

Assumption 1 which was initially defined by Rosenbaum and Rubin (1983) has different names: Heckman and Robb (1984) referred to it as "selection on observables", Lechner (2001) named it the "conditional independence assumption", we can also find it in statistical literature as "ignorability of treatment", "unconfoundedness", while in the missing data literature is referred to as "missing at random". The assumptions state that if $\mathbf{x}$ contains enough information that determines treatment then the joint distribution (y_1, y_0) can be independent of z. In other words, when we condition on $\mathbf{x}$, even if the joint distribution (y_1, y_0) and z can be correlated, once the $\mathbf{x}$'s are conditioned on (y_1, y_0) and z's become independent.[1] In other words there is no unobserved factor that influences both outcomes (y_1, y_0) and treatment z simultaneously.[2] Assumption 2 guarantees that one observes individuals with the same characteristics $\mathbf{x}$ in both the control ($z = 0$) and treatment ($z = 1$) groups. Here $p(\mathbf{x})$ is known as the propensity score.

Finally, Rosenbaum and Rubin (1983) refer to the combination of the two assumptions as "strongly ignorable treatment assignment", conditions that once violated will induce bias in the estimation of the treatment effects. Cases where some of these covariates are not observed will require additional strong assumptions (based on instrumental variables) for possible identification. In the absence of these assumptions, Manski (1990) shows that only bounds can be identified. In this paper we try to overcome this problem by introducing a correction to the ATE estimator obtained via inverse probability weights.

The literature focuses on the following two parameters of interest for the evaluation of the treatment effects: the Average Treatment Effect (ATE) and the Average Treatment on the Treated Effect (ATT). The ATE describes the expected effect of treatment for an arbitrary observation i chosen at random from the population, while the ATT is the mean effect for those that actually participate in the treatment. The focus of this paper is on finding of the ATE measure but the identification of the ATT measure can be done in an analogous way.

[1] In this case $\mathbf{x}$ can be comprised of pre-treatment variables which values do not change during the time treatment takes effect.

[2] This assumption is not testable.

Under the satisfaction of above two conditions, the ATE is defined as follows:

$$ATE = \mathbb{E}[y_{1i} - y_{0i}].$$

By employing the inverse propensity score weighting, we can identify ATE (θ) as follows:

$$\theta = \mathbb{E}\left[\frac{yz}{p(x)} - \frac{y(1-z)}{1-p(x)}\right].$$

Estimation

If $\widehat{p}(\mathbf{x}_i)$ represents a consistent estimator of $p(\mathbf{x})$ obtained by maximizing the log-likelihood (ll):

$$ll = \frac{1}{N}\sum_{i=1}^{N}\big\{z_i \ln p(x_i) + (1-z_i)\ln(1-p(x_i))\big\},$$

where $p(x_i) = \exp(x_i\beta)/\big(1+\exp(x_i\beta)\big)$, then using the entire random sample of size N one has by the analogy principle

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i z_i}{\widehat{p}(x_i)} - \frac{y_i(1-z_i)}{1-\widehat{p}(x_i)}\right\}.$$

A logistic model is used to obtain the Maximum likelihood estimates of the propensity scores $\widehat{p}(X_i)$ and thereafter used to estimate the ATE, $\widehat{\theta}$.

If subject i is randomly assigned to the treatment group and the control group, then both these assumptions, ignorability and overlap hold by construction and ATE is unbiased. If any of the two assumptions listed above do not hold then the ATE is biased.

Bias Correction

Next, we focus on the bias correction of the estimated ATE. The bias of the ATE estimator $\widehat{\theta}$ is given by,

$$Bias = E[\widehat{\theta}] - \theta_0.$$

The bootstrapped bias is

$$Bias^* = E^*[\widehat{\theta}^*] - \hat{\theta},$$

where $E^*[\widehat{\theta}^*] = \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_i^*$ is the average of the bootstrap estimates of the ATE, $\widehat{\theta}_i^*$ obtained from B bootstrap re-samples drawn in each of replication of a Monte Carlo simulation experiment.

The bias adjusted ATE estimator, $\tilde{\theta}_C$ is therefore given by,

$$\tilde{\theta}_C = \widehat{\theta} - Bias^* = \widehat{\theta} - (E^*[\widehat{\theta}^*] - \widehat{\theta}) = 2\widehat{\theta} - E^*[\widehat{\theta}^*].$$

Note that the bootstrapped bias correction specified above can be iterated for further refinements and to improve its accuracy as discussed in Hall (1992) who provides a formula for this procedure. The derivation of this is straightforward. Let $E^*[\widehat{\theta}^*]$ in the above equation denote the average of the

ATE estimates obtained using an "outer" bootstrap. To iterate the bootstrap bias correction the "bias of the bias", i.e., $E[Bias^* - Bias]$ can be estimated using an "inner bootstrap" such that:

$$\widetilde{Bias} = E^*[Bias^{**} - Bias^*] = E^*[(E^{**}[\widehat{\theta}^{**}] - E^*[\widehat{\theta}^*])] - (E^*[\widehat{\theta}^*] - \hat{\theta}]).$$

This can be used to further reduce the bias from the bias such that the iterated bias correction is given by:

$$\overline{\theta_C} = \widehat{\theta} - (Bias^* - \widetilde{Bias}) = E^*(E^{**}[\widehat{\theta}^{**}]) - 3E^*[\widehat{\theta}^*] + 3\hat{\theta}. \tag{2.1}$$

To implement the "inner" bootstrap another B bootstrap resamples are drawn within each "outer" bootstrap resample and estimates of the ATE, $\widehat{\theta}_i^{**}$ are obtained for each of these resamples and averaged over B to obtain $E^{**}[\widehat{\theta}^{**}]$. $E^*(E^{**}[\widehat{\theta}^{**}])$ is therefore obtained by averaging B estimates for the ATE from the "outer" bootstrap resamples. Further refinements using this iterative bias correction procedure can be helpful in reducing bias especially in smaller samples where bias is more pronounced.

Note that the application of a bootstrap bias correction might be computationally complex in nonlinear propensity score models specified in higher dimension. In multiple treatment examples complications may arise as one may need to estimate a bivariate or multivariate propensity score. The finite sample bias in such models can be quite large such that bias reductions may involve further iterations which can be computationally intensive since with every additional iteration a larger number of bootstrap resamples are drawn within each replication (M) of a simulation.

3 Simulations

Monte Carlo simulations are conducted to analyze the bias in the estimates of the Average Treatment effect (ATE) for three special cases of misspecifications commonly encountered in applied work. Firstly, we consider the case of a missing covariate (Ignorability assumption fails) in the treatment response model. In the second case one of the covariates is endogenous and we examine the bias at various levels of endogenity of this variable in our simulations. Lastly, we look at cases where there is an imbalance (Overlap assumption fails) in the distributions of the control and treatment groups.

Let $N = N_t + N_c$, where N_t and N_c are the treatment and control samples. A bootstrap bias correction is applied to the ATE estimates under these three cases of misspecifications. Non-parametric bootstrap re-samples are drawn with replacement in each replication of the experiment to estimate the bias of the ATE. The finite sample performance of the bias correction is compared across various sample sizes; (N_t, N_c) = (30, 50) (60, 80) (100, 150), (200, 300), (500, 750), (1000, 1500) and true values of the treatment parameter θ_0 = 0.5, 1, 2 respectively. The number of simulations (M) and bootstrap re-samples (B) are set to $M = 1000$ and $B = 999$. For the iterated bias corrections of the ATE estimates, the number of bootstrap resamples B is set equal to 99.

Let y_i and z_i be vectors of responses and assignments to a treatment. The responses y_i are linearly related to z_i, the treatment variable and three other covariates, x_1, x_2 and x_3. In the absence of any misspecification therefore this relationship is given by the following equation:

$$y_i = \theta z_i + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i. \tag{3.1}$$

Table 1: Bootstrapped Bias and Bias-corrected ATE for the case of a missing covariate x_3

θ_o		(N_t, N_c):	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
0.5	$\mathcal{N}$	$Bias^*$	0.9074	0.8057	0.5085	0.5190	0.4581	0.3679
			(2.4154)	(1.9294)	(1.3459)	(0.6179)	(0.3754)	(0.2434)
		θ_C	0.5175	0.4805	0.5360	0.5135	0.5090	0.5099
			(2.5248)	(2.0546)	(1.4156)	(0.7087)	(0.4035)	(0.3483)
	t	$Bias^*$	0.9675	0.8213	0.5886	0.5163	0.4687	0.3794
			(2.7301)	(1.7439)	(1.2806)	(0.7354)	(0.3876)	(0.2789)
		θ_C	0.5359	0.4937	0.5486	0.5660	0.5247	0.5576
			(2.8100)	(1.8242)	(1.3231)	(0.8024)	(0.4224)	(0.3484)
1	$\mathcal{N}$	$Bias^*$	0.9578	0.6792	0.5877	0.4831	0.4454	0.3491
			(2.3491)	(1.6792)	(1.2484)	(0.6802)	(0.3323)	(0.2363)
		θ_C	0.9672	0.9577	0.9008	0.9130	0.9899	0.9829
			(2.4214)	(1.7325)	(1.3278)	(0.7419)	(0.4192)	(0.3332)
	t	$Bias^*$	1.0777	0.7977	0.5448	0.4914	0.4566	0.3601
			(2.3884)	(1.6757)	(1.4454)	(0.5655)	(0.3264)	(0.1948)
		θ_C	0.9627	0.9761	0.9793	0.9860	1.0204	1.0245
			(2.4822)	(1.7611)	(1.5391)	(0.6573)	(0.4663)	(0.2836)
2	$\mathcal{N}$	$Bias^*$	0.9032	0.7211	0.5467	0.4806	0.4201	0.3679
			(1.4474)	(1.3693)	(0.8371)	(0.5188)	(0.2175)	(0.1470)
		θ_C	1.8858	1.8722	1.8860	1.8530	1.8811	1.8773
			(1.5146)	(1.4200)	(0.9053)	(0.5336)	(0.3141)	(0.2194)
	t	$Bias^*$	0.9856	0.7611	0.5320	0.4688	0.4306	0.3801
			(1.4297)	(1.1356)	(0.9788)	(0.5211)	(0.2177)	(0.1747)
		θ_C	1.8830	1.9457	1.9434	1.8743	1.8827	1.9484
			(1.5092)	(1.2769)	(1.0223)	(0.6592)	(0.3118)	(0.2237)

Note: (1) For the model in equation 3.1: $y_i = \theta z_i + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$, x_3 is missing and the treatment effect takes the values $\theta = \{0.5, 1, 2\}$, the covariates, x_i are continuous and drawn from a normal ($\mathcal{N}$) and student's t (t) distributions with six degrees of freedom. (2) $Bias^*$ is the bias associated to the estimation of θ using IPS without correction. θ_C is the bias corrected ATE estimator. (3)The standard errors of the ATE estimates before and after bias correction are in parentheses.

The covariates, x_i are continuous and drawn from a normal ($\mathcal{N}$) and student's t (t) distributions with six degrees of freedom. The mean is set to $E[x_i] = 0$ with varying variances $Var[x_i] = 0.4$, 0.5 and 1.5 respectively. The error u_i is standard normal and all the coefficients of the covariates are set to 1 for simplicity.

As shown in the results in the Tables 1–6, the bias of the ATE estimates is substantial when the model is misspecified especially in smaller samples. For samples sizes greater than 100 (treatment and control groups), the BBC-IPS estimator works well in correcting the bias of the ATE estimates for all types of misspecifications. For samples sizes smaller than 100, the bias of the ATE is more pronounced as expected. We find that the BBC-IPS estimator does not reduce bias as effectively

Table 2: Bootstrapped Bias and Bias-corrected ATE for the case of an endogenous covariate x_1

θ_o	ρ		(N_t, N_c) :	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
0.5	0.25	$\mathcal{N}$	$Bias^*$	-0.6078	-0.4841	-0.4641	-0.4522	-0.3956	-0.3353
				(0.7833)	(0.5727)	(0.5502)	(0.3045)	(0.1730)	(0.1518)
			θ_C	0.5339	0.5282	0.5617	0.5707	0.5308	0.5033
				(0.8289)	(0.6714)	(0.5975)	(0.4149)	(0.2338)	(0.2036)
		t	$Bias^*$	-0.7077	-0.5652	-0.5302	-0.5333	-0.5025	-0.4153
				(1.0585)	(0.7472)	(0.7318)	(0.3667)	(0.2080)	(0.1500)
			θ_C	0.5189	0.5341	0.5459	0.5235	0.5431	0.5314
				(1.0848)	(0.8116)	(0.8076)	(0.4586)	(0.2937)	(0.2282)
	0.5	$\mathcal{N}$	$Bias^*$	-0.6516	-0.5693	-0.5248	-0.5163	-0.5042	-0.4173
				(0.6532)	(0.5690)	(0.5002)	(0.2357)	(0.1446)	(0.1043)
			θ_C	0.5349	0.4822	0.5530	0.5126	0.5121	0.5076
				(0.7151)	(0.6046)	(0.5929)	(0.3138)	(0.2207)	(0.2094)
		t	$Bias^*$	-0.7418	-0.6497	-0.6319	-0.6141	-0.5802	-0.4100
				(0.9331)	(0.5590)	(0.4205)	(0.3344)	(0.1941)	(0.1337)
			θ_C	0.5155	0.5411	0.5772	0.5226	0.5338	0.5256
				(1.0396)	(0.6792)	(0.5263)	(0.4837)	(0.2124)	(0.2330)
	0.9	$\mathcal{N}$	$Bias^*$	-0.7044	-0.6540	-0.5962	-0.5688	-0.5152	-0.4801
				(0.7047)	(0.4574)	(0.3069)	(0.2044)	(0.1173)	(0.0815)
			θ_C	0.5411	0.5271	0.5483	0.5186	0.5075	0.5171
				(0.7870)	(0.5125)	(0.4408)	(0.3131)	(0.2009)	(0.1364)
		t	$Bias^*$	-0.8176	-0.7453	-0.6875	-0.6893	-0.6003	-0.4854
				(1.0484)	(0.6070)	(0.3598)	(0.3214)	(0.1618)	(0.1278)
			θ_C	0.5470	0.5266	0.5604	0.5293	0.5274	0.5208
				(1.1027)	(0.7450)	(0.4558)	(0.4067)	(0.2020)	(0.2061)

in these samples for all types of misspecifications in our experiment. The BBC-IPS estimator is therefore iterated as per equation 2.1 and this procedure performs well in reducing bias in smaller samples such that the bias corrected ATE estimates are close to the true parameter values θ_o. The results for these corrections are reported in Tables 1–6.

The bootstrapped bias as discussed in Section 2 is computed for three cases of misspecifications. The first case relates to a missing covariate where the variable x_3 is dropped in the simulations and the effect on the bias of the ATE is observed over various samples sizes and values of θ_o. The results are reported in Table 1.

It is evident that the bias of the ATE is quite large and decreases as the sample size gets larger for covariates that have either thin (normally distributed) or thicker tails (student's t distributed), however covariates with thicker tails induce a slightly higher bias. A bootstrapped bias correction works quite well in reducing the bias over various values of θ and sample sizes such that the bias corrected ATE estimates are quite close to the true θ_o. It is also interesting to note that covariates

Table 3: Bootstrapped Bias and Bias-corrected ATE for the case of an endogenous covariate x_1

θ_o	ρ		(N_t, N_c) :	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
1	0.25	$\mathcal{N}$	$Bias^*$	-0.6787	-0.5952	-0.5764	-0.5243	-0.5047	-0.4022
				(0.7529)	(0.5289)	(0.4691)	(0.3178)	(0.1981)	(0.1253)
			θ_C	1.0126	1.0305	1.0079	1.0102	1.0048	0.9788
				(0.8857)	(0.5813)	(0.5443)	(0.4210)	(0.2160)	(0.2231)
		t	$Bias^*$	-0.6975	-0.6210	-0.6037	-0.5850	-0.5583	-0.4648
				(0.9934)	(0.6226)	(0.4046)	(0.3354)	(0.3196)	(0.1728)
			θ_C	1.0893	1.0494	1.0396	1.0976	1.0953	1.0765
				(1.0359)	(0.7480)	(0.5126)	(0.4976)	(0.4018)	(0.2511)
	0.5	$\mathcal{N}$	$Bias^*$	-0.7881	-0.7275	-0.6980	-0.6185	-0.5986	-0.5044
				(0.7750)	(0.7347)	(0.3621)	(0.2512)	(0.1748)	(0.1128)
			θ_C	1.0297	1.0378	1.0014	1.0053	0.9878	0.9822
				(0.8865)	(0.8378)	(0.3948)	(0.2769)	(0.2554)	(0.2157)
		t	$Bias^*$	-0.8130	-0.7685	-0.7013	-0.6988	-0.6705	-0.5719
				(1.0183)	(0.5977)	(0.3785)	(0.3331)	(0.2171)	(0.1276)
			θ_C	1.0815	1.0319	1.0768	1.0012	1.0934	1.0808
				(1.0842)	(0.7642)	(0.4428)	(0.3982)	(0.3075)	(0.2131)
	0.9	$\mathcal{N}$	$Bias^*$	-0.8520	-0.8142	-0.7753	-0.6921	-0.6190	-0.5780
				(0.7281)	(0.5615)	(0.3537)	(0.2008)	(0.1411)	(0.0897)
			θ_C	1.0541	1.0430	1.0143	1.0137	0.9886	0.9828
				(0.8500)	(0.6194)	(0.4860)	(0.3242)	(0.1840)	(0.1186)
		t	$Bias^*$	-0.8472	-0.8224	-0.7832	-0.7716	-0.7356	-0.6764
				(0.7802)	(0.5760)	(0.3797)	(0.2481)	(0.2322)	(0.1034)
			θ_C	1.0454	1.0546	1.0950	1.0677	1.0952	1.0748
				(0.8675)	(0.7900)	(0.4016)	(0.3213)	(0.3095)	(0.2130)

with thicker tails provide a greater reduction of the bias. Also, the bias of the ATE estimator is quite stable for different values of the treatment effect.

In the second case we allow for one endogenous covariate x_1 such that it is correlated with the error u_i. We examine the bias of the ATE estimates over various strengths of endogeneity by setting the correlation coefficient (ρ) between the covariate x_1 and the error u_i in equation 3.1 to be ρ = 0.25, 0.5 and 0.9. The results of our Monte Carlo simulations are reported in Tables 2, 3, and 4.

It is noticeable that for all sample sizes and values of θ the bias of the estimates gets larger in magnitude with stronger endogeneity for instance, $\rho = 0.9$. The bias is relatively smaller with weaker endogeneity of x_1 as expected and also in larger samples. Contrary to the case of a missing covariate, the bias associated to a correlation between an observable and the unobservable becomes negative and is not stable over different values of the treatment effect. Also, fatter observables induce more bias. The bootstrap bias correction works quite well in reducing bias substantially over all sample sizes and values of ρ and θ. It is noticeable in Tables 2, 3, and 4 that the bias corrected

Table 4: Bootstrapped Bias and Bias-corrected ATE for the case of an endogenous covariate x_1, continued:

θ_o	ρ		(N_t, N_c) :	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
2	0.25	$\mathcal{N}$	$Bias^*$	-0.8112	-0.6920	-0.5692	-0.5693	-0.5424	-0.4630
				(0.7697)	(0.6984)	(0.4141)	(0.2792)	(0.1623)	(0.1214)
			θ_C	1.8875	1.8693	1.8799	1.8848	1.8626	1.8756
				(0.8020)	(0.7170)	(0.5432)	(0.3428)	(0.2599)	(0.1873)
		t	$Bias^*$	-0.7504	-0.7272	-0.6818	-0.6319	-0.5607	-0.4161
				(0.8848)	(0.6283)	(0.4534)	(0.2936)	(0.2005)	(0.1369)
			θ_C	1.9669	1.8952	1.9616	1.9478	1.9376	1.9408
				(0.9113)	(0.8770)	(0.5216)	(0.3620)	(0.2987)	(0.2008)
	0.5	$\mathcal{N}$	$Bias^*$	-0.8955	-0.8371	-0.7762	-0.7013	-0.5917	-0.4987
				(0.7150)	(0.5861)	(0.3840)	(0.2449)	(0.1458)	(0.1125)
			θ_C	1.8791	1.8820	1.8546	1.8754	1.8917	1.8855
				(0.8472)	(0.6901)	(0.4744)	(0.3547)	(0.2344)	(0.1655)
		t	$Bias^*$	-0.9108	-0.8446	-0.7800	-0.7525	-0.7076	-0.5621
				(1.0562)	(0.5508)	(0.3918)	(0.3066)	(0.1800)	(0.1324)
			θ_C	1.8854	1.8775	1.8834	1.8604	1.8706	1.8806
				(1.1413)	(0.7042)	(0.4249)	(0.4003)	(0.2559)	(0.1855)
	0.9	$\mathcal{N}$	$Bias^*$	-0.9524	-0.9067	-0.8212	-0.7980	-0.7143	-0.6174
				(0.7779)	(0.6912)	(0.3345)	(0.2359)	(0.1367)	(0.1126)
			θ_C	1.8854	1.8885	1.8777	1.8830	1.8851	1.8956
				(0.8949)	(0.7760)	(0.4017)	(0.3011)	(0.2292)	(0.1688)
		t	$Bias^*$	-0.9680	-0.9169	-0.8860	-0.8724	-0.7973	-0.6886
				(1.0001)	(0.8232)	(0.4330)	(0.3207)	(0.1858)	(0.1942)
			θ_C	1.8776	1.8848	1.8707	1.8821	1.8772	1.9340
				(1.0921)	(0.8868)	(0.5265)	(0.4319)	(0.2648)	(0.2231)

Note: Same notation as in Table 1 is used. The treatment effect takes the values $\theta = \{0.5, 1, 2\}$, the covariate x_1 is correlated with the unobserved error u with correlations ρ = 0.25, 0.5 and 0.9. The standard errors of the ATE estimates before and after bias correction are in parentheses.

ATE estimates are quite close to the true values of θ. As in the previous discussed case the bias corrected ATE estimated via IPS works better when the covariates have fatter tails.

In the third case we consider an imbalance in the distributions between the treatment and control groups for one of the covariates, x_1. We generate this imbalance by varying the standard deviations for the treatment group $S_t = 0.8$, 1.2, while keeping the standard deviation of the control group constant at $S_c = 0.5$. Note that the $E[x_1] = 0$ is the same for both groups. The results for this experiment are reported in Tables 5 and , 6.

As expected the ATE estimates are biased as a result of this imbalance and in particular the bias gets larger when the standard deviation for the treatment group increases from 0.8 to 1.2. The bias decreases with a larger sample size. As in the case where there is a confounding/endogenity

Table 5: Bootstrapped Bias and Bias-corrected ATE: Imbalanced distributions between the control and treatment groups

θ_o	(S_t, S_c)		(N_t, N_c)	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
0.5	(0.8, 0.5)	$\mathcal{N}$	$Bias^*$	-0.8008	-0.6560	-0.4795	-0.4158	-0.3882	-0.2917
				(1.8870)	(1.6492)	(0.7118)	(0.4690)	(0.2913)	(0.2018)
			θ_C	0.5125	0.5223	0.5119	0.5029	0.5026	0.5091
				(1.9728)	(1.7423)	(0.8349)	(0.5726)	(0.3277)	(0.2705)
		t	$Bias^*$	-0.8510	-0.7399	-0.5237	-0.4366	-0.4032	-0.3084
				(1.6403)	(1.2273)	(0.7549)	(0.4824)	(0.3291)	(0.2182)
			θ_C	0.4896	0.5264	0.5081	0.5041	0.5097	0.5164
				(1.7351)	(1.3629)	(0.8379)	(0.5335)	(0.4248)	(0.2918)
	(1.2, 0.5)	$\mathcal{N}$	$Bias^*$	-0.8073	-0.7091	-0.6064	-0.5161	-0.4543	-0.3366
				(1.8910)	(1.4015)	(0.7415)	(0.4760)	(0.2778)	(0.2282)
			θ_C	0.5170	0.4702	0.5362	0.5077	0.5075	0.5064
				(1.9280)	(1.4777)	(0.8506)	(0.5252)	(0.3165)	(0.2800)
		t	$Bias^*$	-0.8702	-0.7622	-0.6618	-0.5276	-0.4932	-0.3681
				(1.7706)	(1.4111)	(0.7622)	(0.4926)	(0.2962)	(0.1988)
			θ_C	0.5341	0.5180	0.5087	0.5251	0.5434	0.5395
				((1.8257)	(1.5200)	(0.8401)	(0.5667)	(0.3525)	(0.2652)

Note: Same notation as in the previous two tables is used. Additionally, the imbalance in the distributions between the control and treatment groups is obtained by varying the standard deviations for the treatment group $S_t = 0.8, 1.2$, while keeping the standard deviation of the control group constant at $S_c = 0.5$. The standard errors of the ATE estimates before and after bias correction are in parentheses.

effect between observables and unobservables, the bias of the ATE estimates obtained using the IPS estimator is negative and larger in magnitude than case 2 (endogeneity case) but a bit smaller in absolute value when it is compared with the missing covariate case (case 1). The bootstrap bias correction performs quite well over all values of θ, S_c and S_t and sample sizes. Again, fatter covariates help more on the bias reduction.

The standard errors of the ATE estimates before and after bias correction are reported in parentheses in the tables for all types of misspecification. It is interesting to note that the standard errors for the bias-corrected ATE estimates are quite close in magnitude to the standard errors of the ATE estimates prior to bias correction. Bias reductions using the bootstrap correction therefore lead to efficiency gains in terms of the Mean Squared Error (MSE) across all sample sizes and parameter values for all cases of misspecification considered in this experiment.

4 Conclusion

This paper proposes a bootstrap bias correction for the ATE obtained using the estimator for the inverse propensity score (BBC-IPS). We show in our simulations that the BBC-IPC performs well in correcting the finite sample bias leading to efficiency gains in terms of the Mean Squared Error

Table 6: Bootstrapped Bias and Bias-corrected ATE: Imbalanced distributions between the control and treatment groups

θ_o	(S_t, S_c)		(N_t, N_c)	(30, 50)	(60, 80)	(100, 150)	(200, 300)	(500, 750)	(1000, 1500)
1	(0.8, 0.5)	$\mathcal{N}$	$Bias^*$	-0.7917	-0.6996	-0.4838	-0.4312	-0.3956	-0.2867
				(1.7430)	(1.3146)	(0.8368)	(0.4156)	(0.2434)	(0.1781)
			θ_C	1.0682	1.0782	1.1152	1.0049	1.0849	1.1089
				(1.8151)	(1.3918)	(0.9072)	(0.5042)	(0.3279)	(0.2673)
		t	$Bias^*$	-0.8789	-0.7323	-0.5279	-0.4884	-0.3549	-0.2981
				(1.6036)	(1.2331)	(0.8631)	(0.4795)	(0.2659)	(0.1466)
			θ_C	1.1139	1.0643	1.0937	1.1969	1.0769	1.0597
				(1.7153)	(1.3348)	(0.9475)	(0.5386)	(0.3135)	(0.2365)
	(1.2, 0.5)	$\mathcal{N}$	$Bias^*$	-0.8212	-0.7332	-0.6338	-0.5210	-0.4793	-0.3560
				(1.8450)	(1.2571)	(0.7823)	(0.4373)	(0.2820)	(0.1967)
			θ_C	1.0822	1.0666	1.0956	1.0914	1.0678	1.0580
				(1.9244)	(1.4036)	(0.8468)	(0.5352)	(0.3197)	(0.2674)
		t	$Bias^*$	-0.8882	-0.7507	-0.6588	-0.5533	-0.5274	-0.4111
				(1.7495)	(1.1401)	(0.7643)	(0.4782)	(0.2236)	(0.1527)
			θ_C	1.0671	1.0806	1.0683	1.0457	1.1646	1.0429
				(1.8287)	(1.2551)	(0.8257)	(0.5275)	(0.3154)	(0.2425)
2	(0.8, 0.5)	$\mathcal{N}$	$Bias^*$	-0.8791	-0.6652	-0.5190	-0.4289	-0.4078	-0.3032
				(1.8910)	(0.8653)	(0.7802)	(0.4340)	(0.1979)	(0.1316)
			θ_C	1.9091	2.0377	2.0211	1.9713	1.9763	1.9670
				(1.9464)	(0.9596)	(0.8532)	(0.4838)	(0.2231)	(0.2022)
		t	$Bias^*$	-0.8336	-0.6052	-0.4971	-0.4727	-0.4047	-0.3249
				(1.5949)	(0.7255)	(0.6590)	(0.4255)	(0.2376)	(0.1861)
			θ_C	1.9636	1.9492	2.0190	2.0410	2.1044	2.0268
				(1.6895)	(0.8650)	(0.7072)	(0.4984)	(0.3170)	(0.2734)
	(1.2, 0.5)	$\mathcal{N}$	$Bias^*$	-0.9075	-0.7293	-0.6408	-0.5358	-0.5037	-0.4076
				(1.8733)	(1.4650)	(0.6705)	(0.4223)	(0.1863)	(0.1213)
			θ_C	1.9015	2.0693	2.1375	2.0845	2.0615	2.1269
				(1.9334)	(1.4874)	(0.7660)	(0.4515)	(0.2665)	(0.1807)
		t	$Bias^*$	-0.8883	-0.7465	-0.6097	-0.5535	-0.5161	-0.4177
				(1.6495)	(1.1332)	(0.5940)	(0.4159)	(0.1945)	(0.1649)
			θ_C	1.9683	2.0595	2.1267	2.0306	2.0323	2.1054
				(1.7890)	(1.2109)	(0.6513)	(0.5059)	(0.2489)	(0.2530)

(MSE) when we have misspecifications of the propensity score (PS) due to: omitted variables (ignorability property may not be satisfied), overlap (imbalances in distribution between treatment and control groups), endogeneity (confounding effect between observables and unobservables). Depending on the type of misspeciffication the ATE estimate obtained via IPS estimator is overestimated

(omitted variables) or underestimated (imbalances or endogeneity). The bias correction IPS estimator (BBC-IPS) reduces the bias in all the above cases effectively in larger samples. Also, the bias is further reduced if the observed covariates have fatter tails. In smaller samples, the bias of the ATE is more pronounced as expected and further refinements to bias reductions are attained by iterating the BBC-IPS estimator. This procedure works well in correcting the bias of the ATE in these samples.

References

Abadie, A. and Imbens, G. (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1–11.

Clarke, K., Kenkel, B., and Rueda, M. (2015), "Misspecification and the Propensity Score: The Possibility of Overadjustment," *University of Rochester Working Paper*.

— (2016), "Omitted Variables, Countervailing Effects, and The Possibility of Overadjustment," *Political Science Research and Methods*, 6, 343–354.

Hall, P. (1992), "The Bootstrap and the Edgeworth Expansion," *New York: Springer-Verlag*.

Heckman, J. and Robb, R. (1984), "Alternative Methods for Evaluating the Impact of Interventions," *Heckman and Singer (eds.), Longitudinal Analysis of Labor Market Data*.

Hennekens, C. and Buring, J. (1987), "Regression shrinkage and selection via the lasso," *Epidemiology in Medicine*.

Imai, K. and Ratkovic, M. (2014), "Covariate balancing propensity score," *Journal of the Royal Statistical Society. Series B (Methodological)*, 76, 243–263.

Joffe, M. and Rosenbaum, P. (1999), "Invited commentary: propensity scores," *American Journal of Epidemiology*, 150, 327–333.

Kim, M. and Yixiao, S. (2016), "Bootstrap and k-step bootstrap bias corrections for the fixed effects estimator in nonlinear panel data models," *Econometric Theory*, 32, 1523–1568.

Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," *Econometric Evaluations of Active Labor Market Policies in Europe*, 43–58.

Lee, W. (2007), "On Assessing the Specification of Propensity Score Models," *Working Paper*.

— (2013), "Propensity score matching and variations on the balancing test," *Empirical Economics*, 44, 47–80.

Lunceford, J. and Davidian, M. (2004), "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, 23, 2937–2960.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319–323.

Millimet, D. and Tchernis, R. (2009), "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies," *Journal of Business and Economic Statistics*, 27, 33–46.

Peng, X. and Feng, Z. (2011), "Bootstrap Confidence Intervals for the Estimation of Average Treatment Effect on Propensity Score," *Journal of Mathematics Research*, 3, 52–58.

Rosenbaum, P. (1995), "In: Observational Studies," *Epidemiology in Medicine*, 1–12.

Rosenbaum, P. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. (2009), "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" *Statistics in Medicine*, 28, 1420–1423.

Shaikh, A., Simonsen, M., Vytlacil, E., and Yildiz, N. (2009), "A specification test for the propensity score using its distribution conditional on participation," *Journal of Econometrics*, 151, 33–46.

Zhao, H., Rebbeck, T., and Mitra, N. (2009), "A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors," *Genetic Epidemiology*, 33, 679–690.

Received: July 2, 2018

Accepted: January 30, 2019