Causal Inference Using Propensity Score Methods with Clustered Survey Data

Hyunshik Lee[1], Duck-He Yang[1], Ning Rui[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract
The propensity score methods have been widely used for causal inference. The majority of studies have ignored the sample design feature even when complex survey data were used. In recent years, a number of authors have shown that ignoring the sampling weight will lead to biased results. However, causal inference using the propensity score methods for clustered survey data has not been much studied. This paper tries to fill this gap by providing correct ways of incorporating the sample design feature in the calculation of the propensity score and outcome analysis to estimate the treatment effect. The proposed methods will be studied using simulated and empirical survey data.

Key Words: Logistic regression, Sampling weight, Cluster effect, Replicate variance estimator, Bootstrap variance estimator, Confidence interval.

1. Introduction

The causal inference is to statistically study the cause and effect, for which the randomized control trial (RCT) is considered the gold standard. However, in many situations, an RCT is not feasible, too expensive, or unethical. On the other hand, there are many sources of observational data that are already available or can be collected more easily and inexpensively. One of these available data sources is survey data collected by using a complex survey design. An observational study is often conducted to use such rich data sources for causal inference. However, there is a major hurdle in this situation, unlike in an RCT study, because the treatment and control groups are usually not similar or are unbalanced in confounding covariates in the observational data. The main strength of the RCT is that this issue is automatically addressed through random assignment of the study units to the treatment and control groups.

To overcome the imbalance problem in observational data, the outcome analysis has been traditionally carried out using a regression model with the outcome variable as the response variable, the treatment indicator as a dummy covariate, and all auxiliary (potentially confounding) variables as regression covariates. However, since Rosenbaum and Rubin (1983) proposed the propensity score (PS) method, this new method has gained a lot of popularity. The main advantage of the PS method is that it summarizes the effect of confounding covariates by a single number, and this single number is used to make causal inference.

The PS is defined as a probability that a study unit is (non-randomly) assigned to treatment (T) given covariate vector X as follows:

$$\pi_i = \pi(X_i) = P(T_i = 1|X_i), \quad (1)$$

where $T_i = 1$ if study unit i is assigned to treatment and $T_i = 0$ if study unit i is assigned to control. It is a balancing score, that is, for the same PS, the treatment and control groups become similar in covariates. This feature plays a key role in the PS method for causal inference. This idea is formalized by Rosenbaum and Rubin (1983) in the concept of strongly ignorable treatment assignment. It means that T is independent of the potential outcome conditional on all (confounding) covariates, namely,

$$T \perp [Y(0), Y(1)]|\mathrm{X}, \tag{2}$$

where $Y(T)$ is the potential outcome under treatment ($T = 1$) or control ($T = 0$). It also requires another assumption that the probability of receiving treatment is not zero or certainty, which is expressed as follows:

$$0 < \pi = P(T = 1|\mathrm{X}) < 1 \text{ for all X.} \tag{3}$$

Rosenbaum and Rubin (1983) show that if (2) holds, then we have

$$T \perp [Y(0), Y(1)]|\pi(\mathrm{X}), \tag{4}$$

This means that treatment assignment is independent of the potential outcomes conditional on the PS, instead of X.

Theoretically, treatment effect on study unit i is given as the difference between the potential outcomes of the unit under the treatment and control conditions as shown in the following expression:

$$Y_i(1) - Y_i(0). \tag{5}$$

In reality, we cannot use the expression in (5) because we observe either $Y_i(0)$ or $Y_i(1)$, not both. The basic idea of the PS method is to find a control unit i' that has the same or a similar $\pi_{i'}$ (i.e., $\pi_{i'} = \pi_i$ or $\pi_{i'} \approx \pi_i$), and then use $Y_{i'}(0)$ as a proxy for the unobserved $Y_i(0)$.

To implement this basic idea to estimate the treatment effect in causal inference, there are four basic methods.

- Matching – Find proxy control units by matching by the PS, where there are different matching methods such as 1:1 or 1:M (M > 1) matching and other variants.
- Stratification or subclassification – Stratify the treatment and control units together into homogeneous groups based on the PS and treat each stratum as an RCT sample.
- Inverse probability of treatment weighting (IPTW) – Use the nature of the PS as a probability to weight the full sample (both treatment and control groups), so that the weighted treatment and control groups are balanced in covariates.
- PS regression adjustment (PSRA) – Use the PS as a covariate to balance the treatment and control groups through regression.

Note that the matching method uses only the matched sample, resulting in possible waste of a lot of available data. The other three methods use the full sample. However, stratification cannot make each "homogeneous" group perfectly homogeneous (i.e., all group members have the same PS). Although this method incurs some bias, the variance is

usually smaller than that of other methods – generally the bias decreases and the variance increases as the number of strata increases. IPTW and PSRA are unbiased if all assumptions are true.

The weight is defined differently for the IPTW method, depending on whether the goal of estimation is to estimate the population average treatment effect (PATE) for the entire study population or to estimate the population average treatment effect of the treated (PATT). We also use ATE and ATT to refer to the type of weighting or the estimation type, not the parameter to be estimated.

The weight to estimate PATE is defined as

$$w_{1i} = \frac{1}{\pi_i} \text{ and } w_{0i} = \frac{1}{1-\pi_i}, \tag{6}$$

where w_{1i} is the IPTW weight for unit i if unit i is in the treatment group and w_{0i} is the IPTW weight if unit i is in the control group.

The weight for estimating PATT is defined as

$$w_{1i} = 1 \text{ and } w_{0i} = \frac{\pi_i}{1-\pi_i}. \tag{7}$$

Two IPTW basic methods to estimate the treatment effect are (1) by mean difference of the weighted treatment and control groups:

$$\hat{\tau}_{w1} = \frac{1}{n}\sum_{i=1}^{n_1} w_{1i}Y_{1i} - \frac{1}{n}\sum_{i=1}^{n_0} w_{0i}Y_{0i} \tag{8}$$

or (2) by the weighted mean difference:

$$\hat{\tau}_{w2} = \frac{\sum_{i=1}^{n_1} w_{1i}Y_{1i}}{\sum_{i=1}^{n_1} w_{1i}} - \frac{\sum_{i=1}^{n_0} w_{0i}Y_{0i}}{\sum_{i=1}^{n_0} w_{0i}}, \tag{9}$$

where the total sample size of n is divided into two sample sizes:

- Treatment group sample size, n_1;
- Control group sample size, n_0; and
- $n = n_0 + n_1$.

$\hat{\tau}_{w2}$ is based on the ratio estimators (by Hájek, 1971) and is usually more efficient than $\hat{\tau}_{w1}$, which is a Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

$\hat{\tau}_{w2}$ can be estimated using the weighted regression method with the following model and the PS weight (i.e., weighted ANOVA):

$$Y = \alpha. \tag{10}$$

Another form of regression estimator is to use the PS as a covariate:

$$Y = \alpha T + \gamma, \tag{11}$$

where P is the estimated PS. Here the PS weight is not used in regression as the weighting variable, but the PS is used as a covariate to control the effect of different PS in estimation of the treatment effect.

The PS methods have usually been used assuming that the data are from a simple random sample. Therefore, the sampling weights are all equal or conveniently equal to one. If the observational data are collected from a complex survey, ignoring the sampling weight causes a biased estimate. Researchers generally agree that the sampling weight should be used in the outcome analysis. However, there are different opinions about the use of the sampling weight in estimation of the PS. Some authors assert that it is not necessary (Zanutto, 2006), whereas others say otherwise (DuGoff, Schuler, and Stuart, 2014; Ridgeway, Kovalchik, and Griffin, 2015).

This is a relatively young area of research. Most studies have focused on point estimation with stratified simple random samples. However, cluster sampling is also often used in complex survey designs. Only recently, some authors have studied the PS method of causal inference using multistage cluster samples (Austin, Jembere, and Chiu, 2018). They also studied variance estimation using the bootstrap method. They applied the simple bootstrap, ignoring the complex design features in bootstrap sampling. However, it is well-known that the ordinary bootstrap method does not work for complex survey samples (Sitter, 1992). They also found that the simple bootstrap seriously overestimates the variance. We want to address the variance estimation issue using the proper bootstrap method and also using another popular resampling method, the jackknife with the same simulation set-up used by Austin et al. (2018).

This paper is structured as follows. In the next section, we present a more detailed discussion about causal inference using the PS method with complex survey data. Section 3 presents the set-up of the simulation study we conducted and its results. In the last section, we provide some concluding remarks along with some ideas for future study.

2. Propensity Score Methods for Observational Studies with Complex Survey Data

As mentioned in the previous section, it is generally agreed that the sampling weights should be used to estimate the treatment effect. The way in which the sampling weights should be used in the treatment estimation for causal inference using the PS depends on the particular estimation method being used.

- Matching – Typically the estimator given in (9) is used, but the weights are the sampling weights rather than the PS weights, and $n_0 = n_1$ for 1:1 matching, and the estimate is for ATT.
- Stratification or subclassification – Use (9) for each stratum with the sampling weights, and then aggregate stratum estimates using relative weights based on the number of all units in the stratum for ATE or the number of treatment units in the stratum for ATT.
- Inverse probability of treatment weighting (IPTW) – Formula (8) or (9) is used, with the weight defined by the multiple of the PS weight and the sampling weight. Either ATT or ATE can be estimated, depending on how the PS weight is calculated.
- PS regression adjustment (PSRA) – The sampling weight is used in regression estimation as the weight variable, from which an ATT estimate is obtained. ATE can

be obtained by using the multiple of the sampling weight and the appropriate PS weight instead.

As discussed earlier, researchers have differing opinions about whether the sampling weight should also be used in the estimation of the PS. From our experience, it appears that it does not greatly matter in general but that some bias may exist for unusual situations if the sampling weight is not used. Moreover, in our simulation, using the sampling weight in the PS estimate generally enhances the precision of the treatment effect estimate. Therefore, we believe that it is good practice to incorporate the sampling weight in the PS estimation as well.

In order to estimate the variance for the treatment effect estimate, simple incorporation of the sampling weight may not be sufficient. A variance estimation method appropriate for a particular sample design should be used. It seems difficult to derive a Taylor linearization variance estimator that incorporates both the PS estimation and the treatment effect estimation steps, particularly for the matching method and the stratification method. An easy way to get around this is to use resampling variance estimators such as the jackknife or the bootstrap, which do not require linearization. However, there is no guarantee that they will produce an unbiased variance estimate. One of the goals of the present study is to examine their usefulness through simulation.

Austin, Jembere, and Chiu (2018) studied a straightforward bootstrap method for the 1:1 matching. They bootstrapped the matched sample by selecting matched pairs by simple random sampling with replacement. Even disregarding the sample design, it seems obvious that bootstrap sampling should be done first from the original sample and then passed through the PS estimation and treatment estimation steps for each bootstrap sample. Applying the bootstrap to the matched sample, skipping the first step of PS estimation, is a shortcut, which, Austin et al. found, substantially overestimates the variance.

Another issue in the Austin et al. approach is ignoring the sample design features. It is well-known that the ordinary bootstrap for simple random samples does not work for complex survey samples, for which some remedies have been proposed in case of stratified simple random samples (McCarthy and Snowden, 1985; Rao and Wu, 1988). Sitter (1992) proposed a method that can handle a cluster design, which we studied.

As Austin et al. advocated, the shortcut method might be useful when the sample design information is not available in the sample data and, therefore, a design-dependent bootstrap method cannot be used.

However, if an observational study is planned and the bootstrap method is chosen, correct application of the bootstrap should be included in the plan. Therefore, we want to test the full bootstrap variance estimator along with the shortcut version.

We also studied another well-known resampling method, the jackknife. We studied both the full and shortcut versions of the jackknife variance estimator.

3. Simulation Study of the Proposed Methods

We used the same simulation set-up used by Austin, Jembere, and Chiu (2018). However, we expanded our study by including three more methods for treatment effect estimation:

stratification, IPTW, and PSRA. Furthermore, we studied variance estimation using full and shortcut versions of the bootstrap and the jackknife.

Like Austin et al., we used three approaches to estimate the PS.

- PS1: Unweighted logistic regression;
- PS2: Weighted logistic regression with the sampling weight as the regression weighting variable; and
- PS3: Unweighted logistic regression with the sampling weight as covariate.

As mentioned above, we studied four methods for treatment effect estimation, including the 1:1 matching Austin et al. used:

- Matching: 1:1 matching, denoted as MAT1:1;
- Stratification: with 10 equal-sized strata, denoted as STRAT10;
- IPTW; and
- PSRA.

To conduct the simulation study, we generated a population dataset using the same parameters and conditions employed by Austin et al. (2018). The population size was 1,000,000, which was divided into 200 equal-sized clusters, each having 5,000 study units. The 200 clusters were stratified into 10 strata, each having 20 clusters. For each unit, six covariates and one outcome variable were generated. Covariate l, denoted by $x_{l,ijk}$ for unit i in cluster i, stratum j, was generated using $N\left(u_{l,j}^{\text{stratum}} + u_{l,k}^{\text{cluster}}, 1\right)$, where $u_{l,j}^{\text{stratum}} \sim N(0, \tau_l^{\text{stratum}})$ and $u_{l,k}^{\text{cluster}} \sim N(0, \tau_l^{\text{cluster}})$. Six populations were generated by setting $\left(\tau_l^{\text{stratum}}, \tau_l^{\text{cluster}}\right)$ equal to (0.35, 0.25), (0.35, 0.15), (0.35, 0.05), (0.25, 0.35), (0.15, 0.35), and (0.05, 0.35), respectively. This set-up is designed to have covariates with different systematic variation due to stratification and clustering. In the first population (POP1), generated using $\left(\tau_l^{\text{stratum}}, \tau_l^{\text{cluster}}\right)$ = (0.35, 0.25), each covariate has 10.3 percent of its variation coming from the systematic difference between strata and 5.3 percent coming from the systematic difference between clusters. For the second and third populations (POP2 and POP3) generated with $\left(\tau_l^{\text{stratum}}, \tau_l^{\text{cluster}}\right)$ = (0.35, 0.15) and (0.35, 0.05), respectively, these figures are 10.7 percent and 2.0 percent (for POP2) and 10.9 percent and 0.2 percent (for POP3). The percentages for Populations 4, 5, and 6 (POP4, POP5, and POP6) are reversed; that is, 5.3 percent of the variation is attributed to the systematic difference between strata and 10.3 percent to that between clusters, and so on.

Treatment status, Z_i for unit i, was generated from Bernoulli (p_i), where p_i is a linear function of six covariates, $x_1, \ldots, x_6$, as follows:

$$\text{logit}(p_i) = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i} + a_4 x_{4i} + a_5 x_{5i} + a_6 x_{6i}, \tag{12}$$

with $a_0 = \log(0.0329/0.9671)$, $a_1 = \log(1.1)$, $a_2 = \log(1.25)$, $a_3 = \log(1.5)$, $a_4 = \log(1.75)$, $a_5 = \log(2)$, and $a_6 = \log(2.5)$.

We generated a single continuous outcome variable using the following model:

$$\begin{aligned} Y_i = b_0 + \delta z_i + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_6 x_{6i} \\ +0.2 z_i (b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i}) + \varepsilon \end{aligned} \tag{13}$$

where $\varepsilon \sim N(0,1)$, and $b_0 = 0$, $b_1 = 2.5$, $b_2 = -2$, $b_3 = 1.75$, $b_4 = -1.25$, $b_5 = 1.5$, $b_6 = 1.1$, and $\delta = 1$. Note that $x_1, x_2, \ldots, x_6$ are confounding covariates.

From each study population, we selected a study sample following the sample design described below.

- A simple random sample of five clusters was selected from each stratum (50 altogether).
- For stratum k, a simple random sample of n_k units was selected from each sampled cluster, where $n_k = 150, 140, 130, 120, 110, 90, 80, 70, 60, 50$, resulting in a stratum sample size of $5n_k$, for $k = 1, 2, 3, \ldots, 10$ and a total sample size of 5,000.
- Due to an unequal sampling rate (probability) across the strata, the sampling weights are unequal, ranging from 133.33 to 400.

As in Austin, Jembere, Chiu (2018), the target parameter to be estimated was the PATT.

We simulated 1,000 samples from each study population. Each sample was then used to estimate the PS by three methods (PS1, PS2, and PS3) and to estimate the PATT by four methods (MAT1:1, STRAT10, IPTW, and PSRA) with each of the estimated PS. There were 12 (treatment effect) point estimators, defined by 12 combinations of the three PS estimation methods and the four treatment effect estimation methods.

The Monte Carlo (MC) mean of a point estimator is the average of 1,000 sample estimates, which is taken as the true population mean for the point estimator, and the bias of the point estimator is the MC mean minus the population treatment effect. The percentage relative bias of the point estimator is then given by 100 times the bias divided by the population treatment effect. Table 1 presents the relative bias for 12 estimators under six study populations.

Table 1: Percentage relative bias of 12 treatment effect estimators under six populations

PS method/	*POP1*			*POP2*			*POP3*		
Estimation method	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	1.43	0.30	2.47	0.55	0.75	0.03	1.54	0.33	1.74
STRAT10	8.03	6.90	7.93	5.67	6.39	5.87	6.48	6.20	6.50
IPTW	1.65	0.60	1.62	-2.07	-0.62	-1.80	-0.39	-0.77	-0.33
PSRA	1.13	-0.17	1.03	-1.00	-0.28	-0.85	-0.01	-0.40	0.02
PS method/	*POP4*			*POP5*			*POP6*		
Estimation method	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	1.34	0.02	-0.11	1.37	1.63	1.15	3.91	2.18	2.17
STRAT10	6.26	6.35	6.20	7.38	7.96	7.27	8.50	7.65	8.49
IPTW	-0.80	0.13	-0.80	-0.03	0.93	-0.01	1.52	0.49	1.59
PSRA	-0.60	-0.34	-0.69	0.03	0.68	0.02	1.05	0.18	1.08

The relative bias for STRAT10, which is always greater than 5 percent, stands out, whereas all other estimators have a relative bias much closer to zero. As discussed earlier, the stratification method is biased to the extent that the stratum homogeneity is violated. Our

limited simulation (not presented) shows that if the number of strata is increased to 20, the bias can be reduced by half.

The MC variance is the variance of 1,000 sample estimates for a point estimator, which is taken as the true variance of the point estimator. The MC variances for the 12 point estimators included in Table 1 are presented in Table 2 for six populations.

Table 2: MC variances of 12 treatment effect estimators for six populations

PS method/	*POP1*			*POP2*			*POP3*		
Estimation	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
method									
MAT1:1	0.120	0.099	0.123	0.094	0.075	0.090	0.065	0.053	0.066
STRAT10	0.039	0.018	0.037	0.029	0.014	0.028	0.022	0.012	0.022
IPTW	0.050	0.024	0.049	0.049	0.025	0.048	0.028	0.016	0.028
PSRA	0.047	0.023	0.046	0.040	0.021	0.040	0.026	0.014	0.026
PS method/	*POP4*			*POP5*			*POP6*		
Estimation	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
method									
MAT1:1	0.077	0.059	0.073	0.075	0.061	0.067	0.065	0.060	0.072
STRAT10	0.028	0.015	0.027	0.027	0.016	0.027	0.026	0.016	0.026
IPTW	0.035	0.019	0.035	0.034	0.021	0.034	0.030	0.019	0.031
PSRA	0.031	0.017	0.031	0.029	0.016	0.028	0.027	0.016	0.028

STRAT10, which has the largest bias in point estimation, has the smallest variance. This is expected because stratification reduces the variance (at the expense of bias). PSRA has the smallest variance among (nearly) unbiased point estimators, and IPTW is very close to PSRA. MAT1:1 has a variance more than twice as large as other estimators. This is because it uses only a fraction of the control units, whereas the others use the full sample. Therefore, the variance for MAT1:1 can be reduced by increasing the number of matched control units per treatment unit.

When we compare PS estimation methods, it appears that ignoring the sampling weight does not necessarily increase the bias of the point estimators. It is interesting to see that point estimators with PS1 and PS3 behave similarly, whereas the estimators with PS2 consistently perform better in variance and mean squared error (MSE) than those with PS1 and PS3. Therefore, it appears advantageous to incorporate the sampling weight in PS estimation as the weight variable rather than as a covariate or ignoring it.

Table 3 compares the point estimators in terms of MSE for POP1. From the table, we can draw these conclusions:

- In terms of variance, STRAT10 is the best, although it is biased.
- In terms of MSE, STRAT10, IPTW, and PSRA are similar.
- MAT1:3 has much a smaller variance and MSE than MAT1:1 and becomes competitive.

The same conclusion can be drawn from the results for other populations.

Table 3: Comparison of variances and MSEs of the point estimators under POP1

Estimator	*Variance*			*MSE*		
	PS1	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	0.120	0.099	0.123	0.121	0.099	0.124
STRAT10	0.039	0.018	0.037	0.048	0.025	0.046
IPTW	0.050	0.024	0.049	0.050	0.024	0.049
PSRA	0.047	0.023	0.046	0.047	0.023	0.046
MAT1:3	0.058	0.036	0.056	0.060	0.037	0.058

To study variance estimation, we used the jackknife method with formation of replicates by dropping one cluster at a time, resulting in 50 replicates since 50 clusters were selected. The full jackknife was then applied by performing PS estimation and treatment effect estimation for each replicate. The jackknife variance estimate is an appropriately scaled sum of squared deviations between replicate estimates and the full-sample estimate. The jackknife variance estimator for a linear statistic is unbiased and is consistent for a nonlinear statistic if the statistic is a nonlinear function of sample means and totals, but the function is differential (Krewski and Rao, 1981) – the functional form can be linearized (e.g., ratio estimators, regression coefficients, correlation coefficients). Otherwise, the jackknife variance estimator is not consistent – one example of such statistics is the median. In our case, the matching and stratification methods are not linearizable, and we expected that the jackknife would not work well. Some evidence of this outcome is shown in Table 4, where the percentage relative bias of the full jackknife is presented for POP1. The table also shows the coverage rate of the 95 percent confidence interval (CI) over 1,000 simulated samples.

Table 4: Relative bias and coverage rate for 95 percent CI of the full jackknife variance estimator

Point	*Relative bias (%)*			*Coverage (%)*		
estimator	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	2,561.5	3,106.6	2,513.2	100.0	100.0	100.0
STRAT10	130.6	179.6	152.5	99.1	98.9	99.2
IPTW	51.5	5.9	52.7	97.7	94.4	97.9
PSRA	50.5	6.8	52.0	98.7	95.4	98.4

As expected, the full jackknife does not work at all for MAT1:1 and STRAT10. However, it works much better for IPTW and PSRA under PS1 and PS3 but still considerably overestimates, whereas it works very well under PS2.

What would happen if we used the shortcut jackknife? It bypasses the PS estimation step in each replicate, but the full-sample PS is reused in each replicate sample to estimate the replicate treatment effect. Aggregation of the replicate estimates is the same as the full jackknife. The result for the shortcut jackknife is shown in Table 5 for POP1.

Table 5: Relative bias and coverage rate for 95 percent CI of the shortcut jackknife variance estimator

Point	*Relative bias (%)*			*Coverage (%)*		
estimator	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	-10.0	9.0	-13.6	91.2	91.5	94.2
STRAT10	69.5	30.0	124.2	97.6	90.1	98.4
IPTW	54.8	-53.8	55.7	97.6	77.6	97.7
PSRA	52.0	-54.9	53.5	98.1	77.0	97.5

The shortcut worked quite well for MAT1:1. For STRAT10, the bias was substantially reduced from that of the full jackknife. However, it is still not negligible, especially under PS1 and PS2, whereas it is somewhat acceptable under PS2. On the other hand, it performed poorly for IPTW and PSRA. It underestimates the variance under PS2 and gives poor coverage of the 95 percent CI, for which the full jackknife worked very well. Interestingly, there is not much difference between the full and shortcut jackknife methods for IPTW and PSRA under PS1 and PS3.

It is known that the bootstrap variance estimator is better than the jackknife at handling severe nonlinearity of the functional form of the estimate (Kovar, Rao, and Wu, 1988). This is demonstrated in our simulation as well. The full bootstrap selected bootstrap samples from the original sample using the method appropriate for the stratified cluster design (Sitter, 1992), and then the PS and treatment estimation steps were carried out for each bootstrap sample. We selected 200 bootstrap samples, which produced 200 bootstrap treatment effect estimates. To obtain the bootstrap variance estimate, these estimates were aggregated using a simple formula. To construct the confidence interval, we can use either the usual normal theory CI, which will work well if the bootstrap estimates are symmetrical, or the quantile CI calculated from the empirical distribution of the bootstrap estimates, which performs better when the distribution is asymmetric. Table 6 presents the results for variance estimation and quantile 95 percent CI for the full bootstrap for POP1.

Table 6: Relative bias and coverage rate for quantile 95 percent CI of the full bootstrap variance estimator

Point	*Relative bias (%)*			*Coverage (%)*		
estimator	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	14.9	23.0	11.8	100.0	100.0	99.9
STRAT10	-2.2	33.5	2.0	91.6	94.3	91.3
IPTW	-18.8	-13.6	-18.0	91.5	92.1	92.2
PSRA	-21.5	-16.7	-20.3	91.7	93.8	92.2

The relative bias is not so bad, although it is negatively biased for IPTW and PSRA. The quantile CI coverage is generally lower than the nominal value; however, for MAT1:1, it incorrectly gives the 100 percent coverage. This is quite surprising because the relative bias was quite contained. This overcoverage issue disappears when the normal theory CI is used, as seen in Table 7.

Table 7: Comparison of coverage rates of the quantile and normal 95 percent CIs with the full bootstrap variance estimator

Point	*Quantile CI*			*Normal CI*		
estimator	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	100.0	100.0	99.9	95.7	96.6	96.4
STRAT10	91.6	94.3	91.3	91.7	95.3	92.0
IPTW	91.5	92.1	92.2	92.1	92.3	91.9
PSRA	91.7	93.8	92.2	91.8	93.0	91.8

We also tried the shortcut bootstrap variance estimator, which bypasses bootstrapping the original sample and PS estimation step. Therefore, once the full-sample PS estimates are obtained, they are reused in bootstrapping the treatment effect estimation step. Unlike the jackknife, the shortcut bootstrap did not work at all for all estimators, as shown in Table 8. The simulation was run only for PS2 with POP1.

Table 8: Relative bias and coverage rate for 95 percent CI of the shortcut bootstrap variance estimator

Point	*Relative bias (%)*			*Coverage (%)*		
estimator	*PS1*	*PS2*	*PS3*	*PS1*	*PS2*	*PS3*
MAT1:1	NA	167.0	NA	NA	99.7	NA
STRAT10	NA	781.5	NA	NA	100.0	NA
IPTW	NA	684.7	NA	NA	100.0	NA
PSRA	NA	657.3	NA	NA	100.0	NA

4. Summary and Concluding Remarks

Below we summarize our study results for the four treatment effect point estimators.

- MAT1:1, IPTW, and PSRA are (nearly) unbiased.
- STRAT10 is biased but has the smallest variance.
- The bias of STRAT10 can be reduced by increasing the number of strata.
- In terms of MSE, STRAT10, IPTW, and PSRA are similar, but IPTW and PSRA are preferable because there is no bias.
- MAT1:1 uses a much smaller sample and suffers a larger variance, but the variance can be reduced by increasing the number of matching control units.
- It appears that the sampling weight does not make much difference in estimation of the PS in our simulation data. However, the sampling weight-incorporated PS2 gives better treatment effect estimates with a smaller variance.

The study results for variance estimation are summarized below.

- The full jackknife works well for IPTW and PSRA under PS2 in terms of bias and coverage rate.
- The shortcut jackknife works reasonably well for MAT1:1 and STRAT10 (under PS2) but not for IPTW and PSRA, which is opposite to the full jackknife.
- The full bootstrap variance estimator works reasonably well for all point estimators in terms of bias although it underestimates for IPTW and PSRA.

- The coverage rate for the quantile CI based on the empirical distribution of bootstrap estimates is somewhat lower than the nominal value; however, it is not too bad except for MAT1:1, for which the coverage rate is too high (100%).
- If the normal CI is used with the bootstrap variance estimate, the coverage is pretty good for all point estimators.
- The shortcut bootstrap does not work at all for any point estimator.

Based on the results of our study, we provide the following preliminary directions.

- We recommend using the sampling weight in estimation of both the PS and the treatment effect.
- If feasible, use more matched control units in the matching method.
- It is advisable to have a larger number of strata (>10) to the extent that this is affordable.
- IPTW and PSRA use the full extent of all available data and perform better in general in terms of bias and better or equal in terms of MSE, so they should be considered for general use.
- Use the full jackknife for IPTW and PSRA, whereas the shortcut jackknife may be used for the matching and stratification methods, especially with PS2.
- If feasible, the full bootstrap appropriate for the sample design may be used for any point estimators.
- A caution is needed when constructing the quantile CI with the full bootstrap for the matching method; it may be advisable to form both quantile and normal CIs and compare them, as the normal method appears to work well for all point estimation methods provided that the distribution of bootstrap estimates are not asymmetrical.
- The shortcut bootstrap should not be used unless there is strong evidence that it works for a particular situation.

We consider the following items for future study:

- Taylor linearization method for variance estimation;
- Unequal cluster size design with and without probability proportional to size sampling;
- Binary outcome variables;
- Estimation of PATE; and
- Application to real data.

References

Austin, P. C., N. Jembere, and M. Chiu. 2018. Propensity score matching and complex surveys. *Statistical Methods in Medical Research,* 27(4) 1240-1257.

DuGoff, E. H., M. Schuler, and E.A. Stuart. 2014. Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research,* 49, 284-303.

Hájek, J. 1971. Comment on a paper by D. Basu. In: Godambe, V.P., and Sprott, D.A. (eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston, p. 236.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu. 1988. Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16 (Supplement), 25-45.

Krewski, D., and J. N. K. Rao. 1981. Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

McCarthy, P. J., and C. B. Snowden. 1985. The bootstrap and finite population sampling. In *Vital and Health Statistics* (Ser. 2, No. 95), Public Health Service Publication. Washington, DC: U.S. Government Printing Office.

Rao, J.N.K., and C. F. J. Wu. 1988. Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Ridgeway, G., S. A. Kovalchik, and B.A. Griffin. 2015. Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3, 237-249.

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

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Zanutto, E. L. 2006. A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, 4, 67-91.