Compare bootstrap methods for propensity score matching

Yilei Yang yy3421

Yumeng Qi yq2378

Kindle Zhang qz2527

1. Background

1.1 Propensity score matching

Propensity-score matching (PSM) is a statistical matching technique introduced by Paul R. Rosenbaum and Donald Rubin in 1983[i]. When comparing outcomes between treated and untreated groups, PSM attempts reduce the bias due to confounding variables. Unlike randomized trials, where treatment groups are typically balanced on average for each covariate, in observational studies, the assignment of treatments to subjects is typically not random. Thus, by mimicking randomization, PSM creates a sample of units that received the treatment and makes them comparable on all covariates to the untreated samples.

1.2 Bootstrapping under propensity score matching

The precision in variance estimation is vital for constructing confidence intervals, ensuring correct type I error rates in statistical significance tests. However, a crucial challenge lies in accurately estimating the standard error of these effects. Traditional methods for variance estimation assume independence between observations. However, in PSM, treated and control units are matched based on their propensity scores, creating dependencies between matched pairs or sets, which violates the independence assumption.

Bootstrapping is a widely employed resampling technique that allows estimation of the sampling distribution of almost any statistics using random sampling methods.[ii] The basic idea of bootstrapping is that inference about a population from sample data can be modeled by resampling the sample data and performing inference about a sample from resampled data[iii].

In this study, we introduced and evaluated two distinct bootstrap methods tailored for PSM. We aimed to clarify two issues: whether the bootstrap is a suitable approach for treatment effect variation estimation under PSM and, if so, which bootstrap method should be employed.

2. Monte-Carlo simulations

2.1 Data generating processes

The study generated a dataset by incorporating the principles of PSM. We generated 10 baseline covariates (X_1～X_{10}). These covariates were simulated from independent standard normal distributions. Among these 10 covariates, seven influenced treatment selection (X_1～X_7), while another set of seven affected the outcome (X_4～X_{10}). Moreover, covariates were assigned varying degrees of impact, categorized as weak, moderate, strong, or very strong, with respect to their influence on treatment selection or outcome.

For each participant, the likelihood of being selected for treatment was determined using the following logistic model:

$$log(\frac{pi}{1-pi}) = \beta_{treat} + \beta_{low}X_1 + \beta_{med}X_2 + \beta_{high}X_3 + \beta_{low}X_4 + \beta_{med}X_5 + \beta_{high}X_6 + \beta_{very}X_7$$

The intercept of the treatment selection model β_{treat} was chosen to ensure that the proportion of treated subjects in the simulated sample remained fixed at the desired percentage. This study applied simulation at prevalence of 5%, 10%, 20%, 25%, 30%, 40% and 50%. The regression coefficients β_{low}, β_{med}, β_{high} and β_{very} were assigned values of $log(1.25)$, $log(1.5)$, $log(1.75)$ and $log(2)$ respectively. Subsequently, for each participant, the treatment status was generated from a Bernoulli distribution with a subject-specific parameter p_i: $Z_i \sim$ Bernoulli (p_i). We simulated two types of outcomes for each subject: a continuous outcome, a binary outcome. The continuous outcome was generated using the formula $Y_i = Z_i + \beta_{low}X_4 + \beta_{med}X_5 + \beta_{high}X_6 + \beta_{very}X_7 + \beta_{low}X_8 + \beta_{med}X_9 + \beta_{high}X_{10} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$. Set $\sigma = 3$ and Z_i means treatment. Thus, the treatment resulted in an increase of one unit in the mean outcome. For the binary outcome, we randomly generated a dichotomous outcome using a logistic model, $Y_i \sim$ Bernoulli $(p_i$, outcome $)$. Where the logit $(p_i,$ outcome $)$ can be calculated as

$$\beta_{outcome} + \beta_{effect}Z_i + \beta_{low}X_4 + \beta_{med}X_5 + \beta_{high}X_6 + \beta_{very}X_7 + \beta_{low}X_8 + \beta_{med}X_9 + \beta_{high}X_{10}$$

The binary outcome was simulated such that the treatment caused a 0.02 reduction in the probability of the event occurring. Additionally, the marginal probability of the event occurrence for all subjects without treatment was set at 0.10.($\beta_{outcome}$ is different from the β_{treat} which is the prevalence, it's an incidence rate which I control nearly to 1%)[iv]

In our Monte Carlo simulations, we introduced an additional variable by allowing control over the sample size in each simulation. While varying the percentage of treated subjects at different prevalence, we conducted simulations for a total of 1000 in both continuous and binary outcomes. The sample size for each dataset was also adjustable, providing flexibility in exploring the impact of different sample sizes on the simulation outcomes.

Refer to the Appendix for the R code used to generate these simulated datasets.

2.2 Analyses in simulated datasets

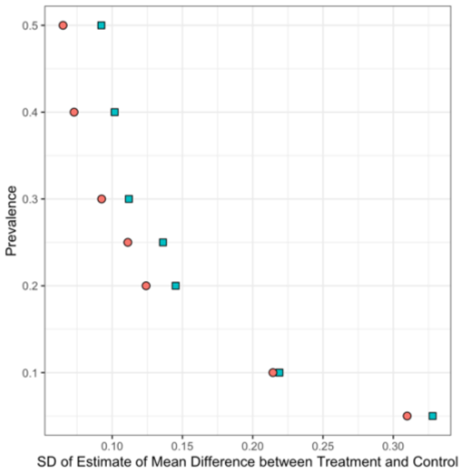
There are many algorithms for matching on the propensity score, the study used greedy nearest neighbor (NNM) on the logit of the propensity score for simplicity. The logistic regression model regressed treatment status on the seven baseline covariates that affected the outcome. Matching without replacement was used, so that each untreated subject was included in at most one matched set.

The study applied two ways of bootstrapping used under PSM. For simple bootstrap, the bootstrap is applied to the matched sample that was obtained using PSM. Notice that one must bootstrap matched pairs, rather than individual subjects. Let the matched sample consist of N matched pairs: $M_1, M_2, ..., M_N$. A bootstrap sample was then drawn from the set of N matched pairs, ending up with a resampled dataset with N pairs, and the effect of treatment is estimated. If simulated multiple times, 1000 in this study, the standard deviation of the treatment effects across the 1000 bootstrap samples is used as an estimate of the standard error of the estimated treatment effect in the original propensity-score-matched sample.[v]

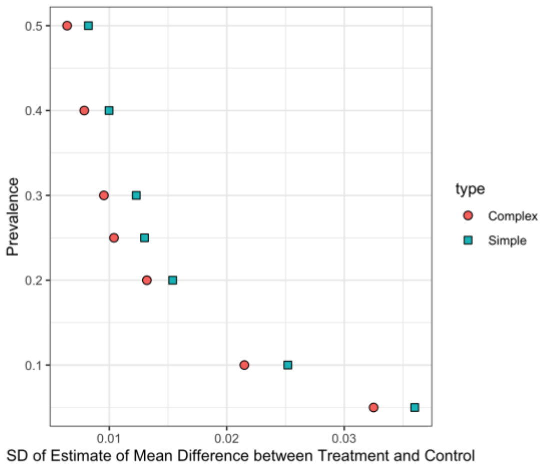
For complex bootstrap, samples are drawn from the original treated group. Based on the bootstrap sample, from treated group only, and the original untreated group, the PSM is applied. Then, the treatment is estimated in the matched samples. The standard deviation of the estimated treatment effects across 1000 propensity score matched samples is then calculated. This serves as an estimate of the sampling distribution of the treatment effect in the original sample.

3. Results and integrations

Using the two bootstrapping methods, variances of treatment effect estimation are collected under different prevalence, both continuous and binary outcomes. These data are further visualized into the graphs below.



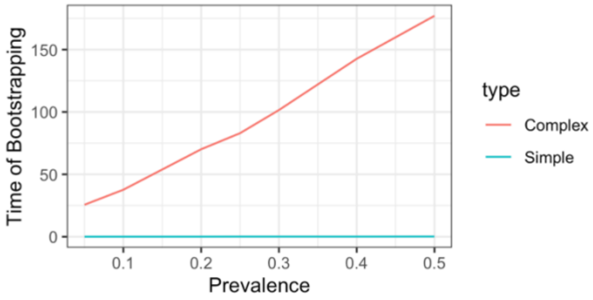
Graph 1: Continuous Outcome



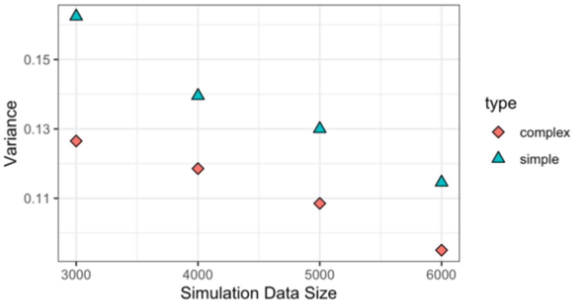
Graph 2: Binary Outcome

Using the above graphs, one significant observation is a pattern of lower average of variance for the complex bootstrap compared to the simple bootstrap for both continuous and binary outcome. Based on this result of low variance of the complex bootstrap, it could be interpreted that the simulation study using the complex bootstrap generates a better stability compared to the simulation using the simple bootstrap.

Furthermore, a decreasing trend could be observed from graph 1 and 2. This trend could indicate a negative relationship between prevalence level and variance of estimate of treatment effect for both complex and simple bootstrap. This similar relationship could also be found in the results of variance collected from simulation using different data sizes. The simulation data sizes used in the data collection process include 3000, 4000, 5000, and 6000. Based on these 4 simulations, a negative relationship between simulation data size and variance of treatment effect could be discovered (Graph 3).



Graph 3: Running time of simple and complex bootstrap



Graph 4: Contiuous Outcome Simulation usin Different Simulation Data Size

This common negative relationship found in simulations using different prevalence levels and different simulation data sizes could lead to a joint interpretation. The increasing prevalence level and data size both lead to an increase in sample data size in the simulation study, increasing the stability of the simulation as indicated by the result of a decrease in variance of treatment effect estimate.

In addition, to quantify the performance of simple and complex bootstrap, data on the running time of each simulation is collected for comparison. Based on the result (Graph 4), a stable running time below 1 second could be discovered for simple bootstrap as the prevalence level increases from 0.05 to 0.5. In comparison, simulations using complex bootstrap have a substantially higher running time compared to those using simple bootstrap at every prevalence level. This is because complex method requires additional implementations of the PSM algorithm, whereas the simple bootstrap method only involves drawing bootstrap samples from the original propensity-score-matched sample. In general, matching algorithms are more computationally intensive than are random sampling algorithms.

Furthermore, there is a noticeable increase in simulation time for complex bootstrap given the increase in prevalence level. Using this result, a smaller computational cost of simulation using simple bootstrap could be deduced compared to those using complex bootstrap.

4. Discussions

To compare between simple and complex bootstrap, the two parameters used in the current simulation study are variance of treatment effect estimate and simulation running time. Therefore, limited quantitative measures are used in this study to determine between the two bootstrap methods. Furthermore, while a difference in variance could be observed between simple and complex bootstrap, it is indefinite to conclude that this observed difference is significant to influence the stability of the current simulation. To further answer this question, an empirical value of variance or a naïve parametric estimator might be ideal as a baseline for evaluation.

To draw a conclusion, in this simulation study, under the PSM framework, simple and complex bootstrap were performed in address of the variance estimation issue. At different level of prevalence, 4 levels of impact to outcomes and 2 types of outcomes, dataset was generated, and bootstrapping was simulated.

Through simulation, a smaller variance is obtained for complex bootstrap. Such result is consistent through different prevalence and different type of outcome. Based on this result, it could be concluded that complex bootstrap method would be preferred for higher stability along with the use of propensity score matching. However, if lower computational cost is more desired for the study, simple bootstrap could be selected.

5. Contribution

Yilei Yang completed sections for data visualization, result discussion, and conclusion section in coding, presentation, and report.

Yumeng Qi completed coding two bootstrapping methods, report sections for background, analyses in simulated datasets and presentation.

Kindle Zhang was in charge of handling modeling and coding tasks within the team, and also taking care of the preparations for the group presentation.

All members conducted literature review of background information about propensity score matching and two bootstrap methods, actively engaged in several group meetings, discussed various analysis plans, and helped edit the final paper.

References

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

[ii] Varian, H.(2005). "Bootstrap Tutorial". *Mathematica Journal*, 9, 768–775.

[iii] Good, P. I. (2006). *Resampling methods*. Birkhũser Boston.

[iv] Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*, 26(4), 734–753.

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