Author's Name: Francis Apaloo

Title: Comparing Performance of Propensity Scores Techniques and Ordinary least Square methods in Estimating Treatment Effects: A Monte Carlo Simulation Study

Date of Publication: February 17, 2013

Introduction and Literature Review

A key issue in quasi-experimental studies and also with many evaluations which required a treatment effects (i.e. a control or experimental group) design is selection bias (Shadish el at 2002). Selection bias refers to the selection of individuals, groups or data for analysis such that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed (Shadish el 2002). There are many ways in which selection bias threatens the validity of study conclusions. One is internal validity, which refers to the causal link between independent variables (which, for example, describe the participants or features of the service they receive) and dependent variables (particularly the outcome of the program). Here we are concerned with whether the program or intervention is the cause responsible for the observed effects rather than extraneous factor.

The other is external validity. Here we are also concerned whether the findings of an evaluation or a study can be generalized to another study of a similar type. One way to reduce selection bias is to conduct a randomized experiment in which units are assigned randomly to experimental and control groups. This leads to high internal validity. When the units were very much the same at the beginning and had much the same experiences during the period of the program except for the receipt of the program itself. Any differences that are observed between the two groups at the end are fairly certainly due to the program.

Random selection would to some extent control for selection bias, because units would be assigned to treatment and control groups prior to program implementation on random basis and this "process would tend to ensure fair distribution of observed and any other unobserved biasing factors in the assignment process" (Shenyang G. & Fraser M., 2010).

However, in observational studies it is impossible to make this assumption. All that we know is that there are several variables believed to be correlated to the outcome variable. The only alternative in "retrospective studies, is to identify these influences and attempt to control for them" (Rubin & Rosenbaum, 1984). This is challenging because while in treatment and control groups, there are many background variables and proxy variables that may bias outcome in an observational study.

While Rubin and Rosenbaum recommend propensity score matching in order to fairly balance the treated group and control groups at baseline, another biasing issue that arises with the use of propensity score matching (PSM) in addition to the issue of selection bias in observational studies is that of specification error. A specification error refers to the exclusion of any influential variable among the covariates in a regression model such as PSM. For example, consider the recently proposed theories of "Multiple Intelligences." If multiple intelligences are an important determinant in school performance among kindergarten aged students and this factor is not included among the covariate variables included in the PSM study then this difference could explain any observed differences in student performance at outcome.

In summary, these issues of selection bias and specification error are the major obstacles to be faced when using PSM to adjust for pre-existing group comparisons in observational studies. We seek an approach that removes all the problems associated with observational data, but this is impossible. Nevertheless, despite the problems, PSM and other matching techniques is

a step up from simply explaining the limitations of observational studies and then reporting group performance results at outcome as if the study were experimental as is often done in traditional observational studies.

The purpose this study is to compare the performance of propensity score techniques and OLS method using a Monte Carlo simulation. A Monte Carlo study is a simulation exercise "designed to shed light on properties of competing estimators for a given estimating problem" (Shenyang G. & Fraser M., 2010). The current study chose to use a Monte Carlo study to compare the performance of PSM techniques and OLS because such a simulation approach allows us to examine comparative results in a way that is more intuitive and less technical.

Monte Carlo studies are very popular in methodological studies. Because they are most often used to "show the importance of checking the tenability related of assumptions related to corrective methods and also to compare models under different scenarios of data generation" (Shenyang G. & Fraser M.,2010). It is the latter use of Monte Carlo simulation which is useful to this study.

To that end, many methodology scholars have found Monte Carlo simulation exercise very useful in many areas of research. For instance, in 2010, Shenyang and Fraser conducted Monte Carlo studies that compared the performance of the following models (i.e OLS regression, Heckit treatment effect and matching estimators). They stimulated two data generation settings and compared the performance across the three models in each setting. They concluded that the OLS regression did poorly in each setting of data generation.

Similarly, Zhao (2004) also conducted a Monte Carlo study to compare propensity score matching with covariate matching estimators under different conditions. Zhao found that "selection bias due only to observables was a strong assumption, however, with a proper data set and if the selection-only-on-observables assumption was justifiable, matching estimators were useful for estimating treatment effects". Furthermore, Zhao found no clear winner among different matching estimators and that the propensity score matching estimators rely on the balancing property.

This study expands on Shenyang and Fraser work by using a similar data generation strategy for each setting to compare the performance of OLS regression and propensity score matching techniques (i.e. nearest neighbor, optimal matching and subclassification).The rest of this study is organized as follows: Models, design, findings and conclusion

Models

Ordinary Least squares Regression

Ordinary least-squares (OLS) regression is a generalized linear modelling technique that may be used to model a single response variable which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been appropriately coded (Hutcheson, G. D. 2011)

Ordinary Least squares Regression is represented as follows**:** $Y = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3$

Where Y is the outcome variable

X1,x2 and x3 are explanatory variables

$\beta 1$, $\beta 2$ and $\beta 3$ are regression coefficients

Propensity Score Techniques

To fully comprehend the complexities of propensity score models, a key variable called propensity score is fully delineated.

Propensity score. Rosenbaum and Rubin (1983) defined propensity score as the conditional probability of assignments to a particular treatment given a vector of observed covariates. Propensity scores balance observed differences between treated and control participants in a sample. Rosenbaum (2002b, p298) showed that a treated and control participants with the same value of propensity score have the same distribution of the observed covariate X. Also treatment assignment and observed covariates are conditionally independent given the propensity score. This property links the propensity score to the assumption regarding ignorable treatment assignment. In other words, conditional on propensity score, the covariates may be independent of assignment to treatment. Therefore, for observations with the same propensity score, the distribution of covariates should be the same across the treated and the control groups. In addition, the propensity score of each participant has the same probability of assignment to treatment as in randomized experiment.

Furthermore, according to Rubin, if the ignorable assignment assumption holds and the estimated propensity score $e(x_i)$ is a balancing score, then the expected difference in observed responses to treatments conditions at $e(x_i)$ is equal to ATE at $e(x_i)$. This links the propensity score to the counterfactual framework and shows how the problem of not observing outcome for the treated participants under the control group can be resolved. It follows that the mean differences of the outcome variable between treated and control participants for all units with the same value of propensity score is an unbiased estimate of ATE at that propensity score. This idea forms the foundation for all propensity score matching techniques.

The technique for estimating propensity score is not the focus of this study but the different matching methods is component of this study. One of the requirements of this study is to a create a matched sample using three propensity score models so that a post matching analysis can be conducted on the OLS regression based on the matched sample. The following techniques are used to create the matched sample:

Nearest Neighbor Matching

Based on the propensity score, the nearest neighbor procedure then matches each treated case to a control case (i.e. a 1-to-1 match) using nearest neighbor within a caliper (a caliper in this case is the standard deviation of the estimated propensity scores).

Optimal Matching

The optimal matching algorithm is not a type of greedy matching algorithm which divides a large decision making problem into a series of smaller, simpler decisions each of which is handled optimally. It rather involves the process of developing matched sets with a size such that the total sample distance of propensity scores is minimized. In other words, the optimal matching algorithm creates S sets and identifies which controls are matched to which treated participants in such a way that matching optimizes the total distance for the given data set **(**Guo, S., & Fraser**,** M. 2010**).**

Subclassification

According to Rosenbaum and Rubin (1984) , prior work on sub classification has shown that creating subclasses based on propensity score balances all observed covariates. There are many ways of forming subclasses based on propensity score. Subclasses can be formed based on the

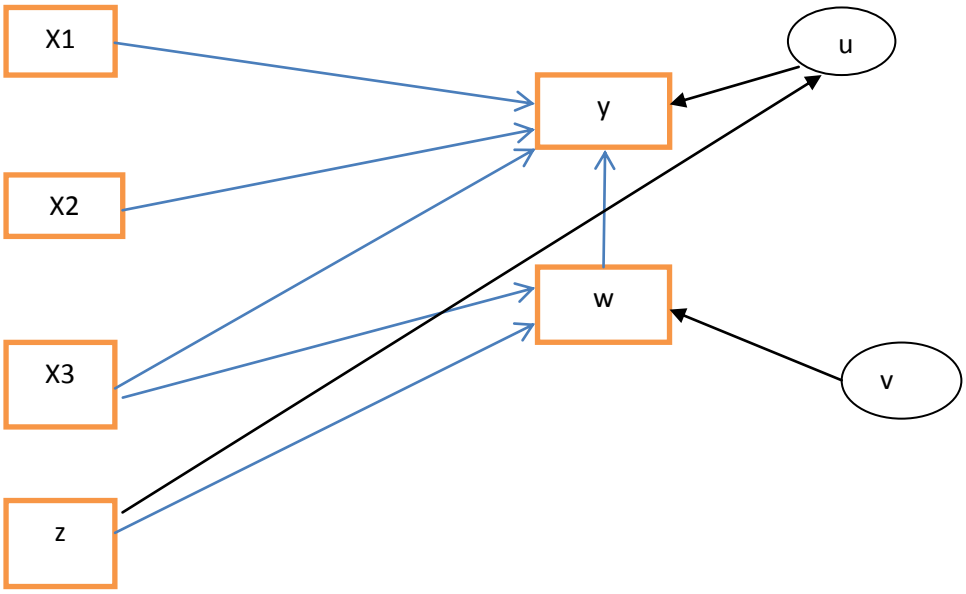
number of observations with in a subclass or subclasses can be formed based on the inverse variance of subclass-specific treatments (Rosenbaum and Rubin 1984). Based on the latter approach , we can form propensity score subclass this way and then estimate an overall treatments effect using the weighted mean of the subclass specific estimates.

Design of the Monte Carlo Study

In this subsection, I show specification for the two settings of selection bias simulated by the Monte Carlo Study. I also demonstrate the model specification for each of the four techniques under these settings and the evaluation I used.

There are two scenarios whereby the data is generated: figure 1 illustrates one of the scenarios known as selection on the observables.

Figure 1 Setting 1: Selection on the observables



This condition occurs when Z as shown in the figure (a covariate) is correlated with u (an error) but furthermore Z is uncorrelated with v. Also under this data generation process the following is also imposed as shown in Figure 1:

1) Three variables x1,x2,x3 are related to the outcome variable y

2) Z determines treatment assignment w only

3) X3 also affects treatment effect

Using Mplus syntax

1. Specification of setting 1 using the following specification is generated:

$Y = 100 + .5x_1 + .2x_2 – 0.05x_3 + .5w + u$

$W^* = .5z + .1x_3 + v$

Where x_1, x_2, x_3, z and u are random variables, normally distributed with a mean vector of (3 2 10 5 0), standard deviation vector (.5 .6 9.5 2 1), and the following symmetric correlation matrix:

$$r(x_1, x_2, x_3, z, u) = \begin{bmatrix} 1 & & & & \\ .2 & 1 & & & \\ .3 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & .4 & 1 \end{bmatrix}$$

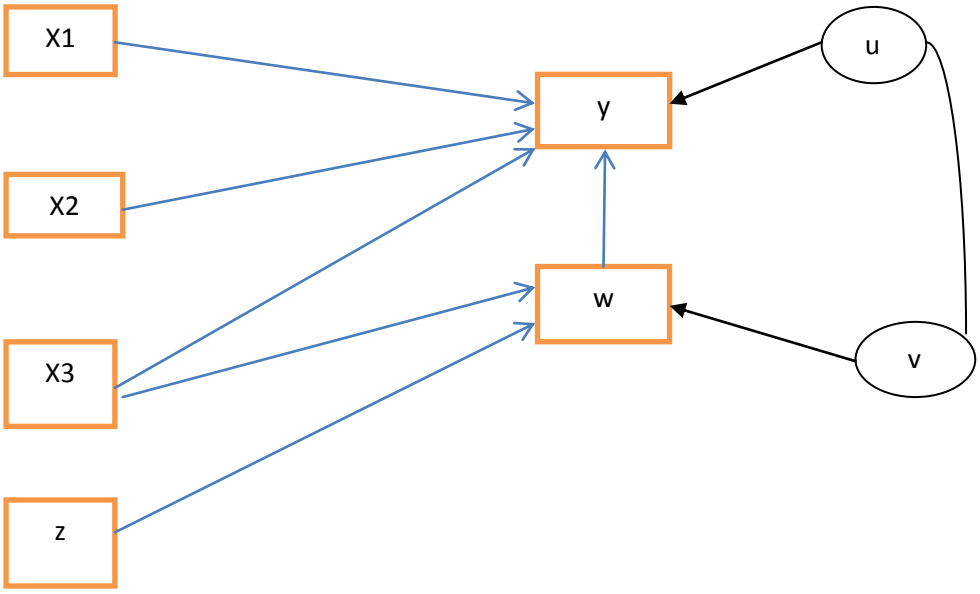
In addition, v is a random variable that is normally distributed with mean zero and variance one; and W = 1, if W* > Median (W*), and W = 0 otherwise.

The above specification creates a correlation between z and u of .4 and correlation between u and v of 0. Thus, the data generation meets the requirement for simulating selection on observables.

The Monte Carlo study generates 10,000 samples for each technique with a size of 500 observations per sample. Under this specification, the true average treatment effect in the population is known in advance that is W = .5

Figure 2 illustrates the other scenario known as selection on the observables.

Figure 2 Setting 2: Selection on the unobservables



This condition occurs when Z as shown in the figure (a covariate) is uncorrelated with u (an error) but furthermore u is correlated with v. Also under this data generation process the following is also imposed as shown in Figure 1:

1) Three variables x1,x2,x3 are related to the outcome variable y
2) Z determines treatment assignment w only
3) X3 also affects treatment effect

Using Mplus syntax

2. Specification of setting 2 using the following specification is generated:

$Y = 100 + .5x_1 + .2x_2 – 0\ .05x_3 + .5w + u$

$W^* = .5z + .1x_3 + v$

$v = 8 + .15E$

Where x_1, x_2, x_3, z, u and E are random variables, normally distributed with a mean vector of (3 2 10 5 0 0), standard deviation vector (.5 .6 9.5 2 11.)

And the following symmetric correlation matrix:

$$r(x_1, x_2, x_3, z, u, E) = \begin{bmatrix} 1 & & & & & \\ .2 & 1 & & & & \\ .3 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & .7 & 1 \end{bmatrix}$$

In additional, 8 is a random variable that is normally distributed with mean zero and variance one; and W = 1, if W* > Median (W*) and W = 0 otherwise.

The above specifications create a correlation between z and u of 0, and a small correlation between u and v of .1. Thus, the data generation meets the requirements for simulating selection on unobservables.

The Monte Carlo study generates 10,000 samples for each technique with a size of 500 observations per sample. Under this specification, the true average treatment effect in the population is known in advance, that is W = .5

Specification of each model in setting 1and setting 2are shown below

1) OLS regression : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z + \tau W$

2) $p(w=1)\ e(x)=1/\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z$ is the estimated propensity score

Using the estimated propensity score I created matched samples using the following in R function called Matchit for nearest neighbor, optimal matching and subclassification to obtain the parameter estimates for x1,x2, and x3 to run the Monte Carlo simulation in mplus:

Findings

Table 1 presents the findings of the Monte Carlo study under the two settings. Under Setting 1, that is the selection on observables, the optimal model performed the best. The classification model follows with nearest neighbor ranked third. The OLS regression came last. On average, OLS regression estimated a treatment effect of which was above the true effect (or an overestimation of 2.9%). Though OLS came last, it is worth noting that OLS works reasonable well in this setting because x3 and Z are the main variables determining selection, Z and u are correlated and both source variables x3 and Z are controlled in the analysis. These conditions are restrictive and may not hold in practice. In a typical application, we may not know that x3 and Z are the major source of selection x3 and Z may not be available or collected; and Z and u may not be correlated.

I now summarized the main findings of the model performances under setting 2. Overall classification model, optimal model and nearest neighbor produced acceptable estimates. The OLS regression model did very poorly. This confirms the danger inherent in using OLS regression to correct for selection bias, particularly when hidden selection bias is present. Furthermore, this finding also indicates that nearest neighbor and optimal findings are slightly sensitive to hidden bias. However, classification is robust under this condition. This goes to

confirm other studies that classification technique balances covariate distribution between a treatment and control group in the mist of s hidden bias.

Table1 Monte Study Comparing Models

	Estimated Average Treatment Effect			
Model	Population Estimates	Average Estimates	Bias (%)	Rank
Setting 1				
OLS regression	0.5	0.5143	2.86	4
Nearest Neighbor	0.5	0.4979	-0.42	3
Optimal Matching	0.5	0.4998	-0.04	1
Classification	0.5	0.4996	-0.08	2
Setting 2				
OLS regression	0.5	0.6178	23.56	4
Nearest Neighbor	0.5	0.5157	3.14	3
Optimal Matching	0.5	0.4917	-1.66	2
Classification	0.5	0.5045	0.9	1

Note: Models are OLS and propensity scores techniques.

Conclusion

One very important thing that most researchers particularly those in quantitative fields know is that OLS regression is not a valid approach to correct selection bias. This is confirmed this study. We can conclude from this study that overall propensity score matching methods are better in correcting selection bias as compared to OLS regression. However there are still challenges imposed by selection bias.

References

Guo, S., **&** Fraser, M. **(**2010**)** . Propensity Score Analysis: Statistical Methods and Applications. *Advanced Quantitative Techniques in the Social Sciences Series,* \ SAGE, NY.

Holland , P.W., & Rubin, B.D (1984) . On Lord's Paradox. Program Statistics Research Project. Education Testing Service, Princeton NY

Hutcheson, G. D. (2011). Ordinary Least-Squares Regression. The SAGE Dictionary of Quantitative Management Research. Pages 224-228.

Rosenbaum, P.R., & Rubin, D.B. (1884) . Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association, Vol. 79, No. 387 (Sep., 1984), pp. 516- 524*

Rosenbaum, P.R., & Rubin, D.B. (1883) . The central role propensity score in observation study. Biometrica.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin