Monte Carlo Simulation Study

Author:
Matthias Lukosch (20-601-050)

1.Setting

i) Parameter of Interest

The parameter of interest in this simulation study is the average treatment effect (ATE). The ATE is defined as

$$ATE = \mathbb{E}\left[Y^1 - Y^0\right] \tag{1}$$

where Y^1 denotes the potential outcome under treatment and Y^0 the potential outcome under non-treatment. It is only possible to estimate an average treatment effect and not an individual treatment effect because we do not observe the counterfactual for an individual. By using the definition of the average treatment effect of those treated and not treated, one can rewrite equation (1) as

$$\begin{aligned} \Leftrightarrow ATE = P(D=1) \cdot \left(\mathbb{E}\left[Y^1|D=1\right] - \mathbb{E}\left[Y^0|D=1\right]\right) \\ + P(D=0) \cdot \left(\mathbb{E}\left[Y^1|D=0\right] - \mathbb{E}\left[Y^0|D=0\right]\right) \end{aligned} \tag{2}$$

By generating data in an experiment or by collecting observational data, some of the elements in (2) are identified by the data as indicated below.

$$\begin{aligned} \Leftrightarrow ATE = \underbrace{P(D=1)}_{observable} \cdot \left(\underbrace{\mathbb{E}\left[Y|D=1\right]}_{identified} - \underbrace{\mathbb{E}\left[Y^0|D=1\right]}_{counterfactual} \right) \\ + \underbrace{P(D=0)}_{observable} \cdot \left(\underbrace{\mathbb{E}\left[Y^1|D=0\right]}_{counterfactual} - \underbrace{\mathbb{E}\left[Y|D=0\right]}_{identified} \right) \end{aligned} \tag{3}$$

However, data alone cannot identify the ATE. Thus, one needs further identifying assumptions that define a research design. To be able to identify the ATE in this study, the identifying assumption which form the so-called 'selection-on-observables' research design are used.

ii) Identification Strategy

As it is typical with observational data, individuals/observations with the same background characteristics are usually not randomly assigned to the treatment or control group. Thus, it would be misleading to simply compare the average outcome between treated and controls due to selection bias. Nevertheless, even in the case with observational data, it is possible to invoke further assumptions to remove selection bias and to identify the ATE. The assumptions needed are:

1. Conditional Independence Assumption (CIA)

2. Common Support Condition (CSC)
3. Exogeneity of Confounders
4. Stable Unit Treatment Value Assumption (SUTVA)

Under these assumptions, there are three ways to identify causal effects (Proof is not stated as it can be found in the lecture notes): By outcome regressions, by inverse probability weighting (IPW) or by combining both (double robustness). In this study, the focus is laid on the first two ways.

$$\text{Outcome regression:} \qquad ATE = \mathbb{E}\left[\mu(1,x) - \mu(0,x)\right] \tag{4}$$

$$\text{IPW:} \qquad ATE = \mathbb{E}\left[\frac{DY}{p(x)} - \frac{(1-D)Y}{1-p(x)}\right] \tag{5}$$

$$\mu(1,x) = \mathbb{E}(Y|D=1, X=x)$$
$$\mu(0,x) = \mathbb{E}(Y|D=0, X=x)$$
$$p(x) \equiv \text{propensity score}$$

iii) Estimators

The estimators used to estimate the ATE are OLS and IPW as given below.

$$\text{OLS:} \qquad \hat{ATE} = \hat{\beta}_0 \tag{6}$$

$$\text{IPW:} \qquad \hat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d_i y_i}{\hat{p}(x_i)} - \frac{(1-d_i)y_i}{1-\hat{p}(x_i)}\right] \tag{7}$$

Furthermore, it is assumed that there is no effect heterogeneity related to covariates and for OLS it is additionally assumed that the true data generating process is a linear model.

2.Simulation Design

i) DGP1: The first data generating process is set as simple as possible and aims to show the weaknesses of the IPW estimator. The process equation is given by

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_i + u_i \qquad \text{i = 1,..N} \tag{8}$$

where y_i is a continuous dependent variable, d_i a dummy variable indicating the treatment status, x_i a continuous confounder, and u_i a homoskedastic error term. The latter is normally distributed with mean zero and constant variance of 50. The observations of the confounder x are drawn from a normal distribution with mean 0 and standard deviation 40. The dummy variable is engineered such that the treatment assignment depends on the value of the confounder x.[1] This can be formulated as

$$d_i = \mathbb{1}(x_i + e_i > 0) \qquad \text{where } e_i \sim N(0, 50^2). \tag{9}$$

[1] This procedure can be motivated by a simple example: Assume the treatment to be the admission to university which typically depends on individuals' characteristics (e.g. grades, financial situation) given by the confounders. To reduce complexity, the DGP1 only contains one of these characteristics.

The number of observations to be drawn from the defined population is N = 200 and the number of simulations is set to 500. The true coefficients are assumed to be $\beta_0 = 30$, $\beta_1 = 5$, and $\beta_2 = 8$.
For the coefficient estimation of DGP1, the OLS regression function is correctly specified as a linear model and both OLS and IPW do not neglect the relevant confounder. Thus, it is reasonable to expect the OLS estimator to perform well in this setting as there are no violations of the typically imposed OLS assumptions. Indeed, one would expect the OLS estimator to be BLUE since the Gauß-Markov-Theorem applies. Regarding the IPW estimator, the estimate of the propensity score should be sufficiently far away from 0 and 1 such that the estimator does not explode. However, the drawn samples are relatively small as well as the number of simulations and there is only one confounder specified such that it might happen to get estimates of the propensity score close to 0 or 1 by chance. This might lead to exploding estimates of the ATE which might have a substantial impact on the mean and variance of the distribution of the estimates. The latter effect could potentially be mitigated by drawing larger samples and doing more simulations.

ii) DGP2: This data generating process aims to showcase the performance of OLS and IPW in the case of imperfect multicollinearity between confounders. The process equation is given by

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + u_i \qquad \text{i = 1,..N} \tag{10}$$

where y_i is a continuous dependent variable, d_i a dummy variable, x_{1i-3i} confounders and u_i a homoskedastic error term with mean zero. Two of the confounders are drawn from a multivariate normal distribution with moments given below.

$$\mathbb{E}[X_1] = 4 \qquad \mathbb{E}[X_2] = 9 \qquad Cov(X_1, X_2) = \begin{pmatrix} 300 & 80 \\ 80 & 150 \end{pmatrix} \tag{11}$$

X_3 can be approximately represented as a linear combination of the remaining two confounders.

$$x_{3i} = 0.8x_{2i} + 0.2x_{1i} + e_i \qquad \text{where } e_i \sim N\left(0, 500^2\right) \tag{12}$$

The dummy variable is set to depend on the values of the confounders X_{1i} and X_{2i} as given below.

$$d_i = \mathbb{1}(x_{1i} + e_i > 4 \ \& \ x_{2i} + e_i > 9) \qquad \text{where } e_i \sim N(0, 50^2). \tag{13}$$

In each of the 500 simulations, 500 observations are drawn from the population. The true coefficients are assumed to be $\beta_0 = 30$, $\beta_1 = 5$, $\beta_2 = 8$, $\beta_3 = 45$, and $\beta_4 = 0$.
Since OLS uses an irrelevant confounder that can be approximately represented as a linear combination of the other two confounders to estimate the coefficients of the DGP2, one would expect the OLS estimator to become less efficient. However, one would still expect the OLS estimator to be unbiased. Regarding the IPW, the imperfect multicollinearity should not have a large impact on the performance of the IPW since there is no need to specify a parametric relationship between the Y's, X's, and d's to estimate the ATE.

iii) DGP3: This data generating process aims to show what happens to the consistency of the OLS

and IPW estimators if the CIA assumption is violated. The process equation is given by

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_{1i} + \beta_3 x_{2i} + u_i \qquad \text{i = 1,..N} \tag{14}$$

where y_i is a continuous dependent variable, d_i a dummy variable, x_{1i} and x_{2i} confounders, and u_i a homoskedastic error term. The error term is distributed as in DGP1. The observations of the confounders are drawn from a multivariate normal distribution with the same moments as under DGP2. The dummy variable d_i is again engineered as in DGP1. The number of observations drawn from the population is equal to 500 and in total 500 samples are drawn from the population. The true coefficients are assumed to be $\beta_0 = 30$, $\beta_1 = 5$, $\beta_2 = 8$, and $\beta_3 = 45$.
For the coefficient estimation, it is assumed that the confounder x_{1i} is not observed and therefore not specified in the OLS regression as well as in the Probit regression of the IPW estimator.
Regarding the expectation on the performance of the estimators, it is reasonable to state that both estimators will not be able to identify the ATE as the CIA does not hold in this setting. The OLS estimator is likely to be biased and in theory, this omitted-variable bias does not vanish if the sample size goes to infinity.

3. Results

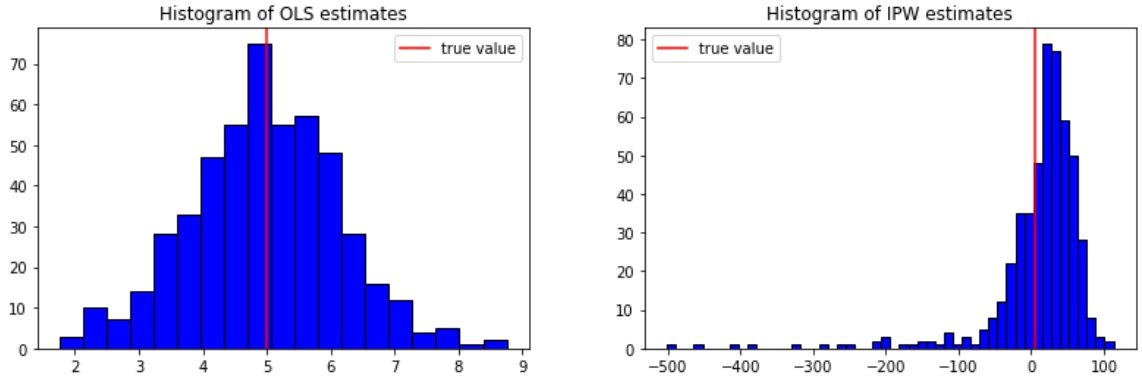
i) Results DGP1
To assess the performance of the OLS and IPW estimator, the following performance measures are calculated: Bias, variance, and mean-squared error (MSE). The values of the latter are given in Table 1.

Table 1: *Performance measures of OLS and IPW in DGP1 setting.*

	OLS	IPW
Bias	-0.0315	4.8181
Variance	1.3547	4461.6475
MSE	1.3557	4484.8616

As indicated by all performance measures, OLS dominates IPW in this setting. This result is in line with the expectations stated under 2i). as the framework of DGP1 is ideal for the OLS estimator and especially due to the low sample sizes less ideal for the IPW estimator. The histograms in Figure 1 illustrate the unbiasedness of OLS and the biasedness of the less efficient IPW.

Figure 1: *Histograms of IPW and OLS estimates in DGP1 setting.*



ii) Results DGP2

The performance measures of the OLS and IPW estimator in the setting of DGP2 are given in Table 2.

Table 2: *Performance Measures of OLS and IPW in DGP2 setting.*

	OLS	IPW
Bias	71.51	-0.6341
Variance	$8.097 \cdot 10^6$	521.63
MSE	$8.103 \cdot 10^6$	522.03

As expected, imperfect multicollinearity increases the variance of the OLS estimates substantially such that the IPW is more efficient compared to OLS in this setting. However, theory suggests that imperfect multicollinearity does not lead to biased OLS estimates which is surprisingly the case in this simulation.

iii) Results DGP3

The performance measures of the OLS and IPW estimator in the setting of DGP3 are given in Table 3.

Table 3: *Performance measures of OLS and IPW in DGP3 setting.*

	OLS	IPW
Bias	62.5151	62.7986
Variance	122.1806	123.1697
MSE	4030.3183	4066.8339

As indicated above, both estimators perform equally worse in this setting where the CIA as a fundamental identification assumption is violated. Both estimators are biased and entail large variances. This result is also in line with the expectation stated under 2iii).