Monte Carlo Study

OLS vs. Doubly Robust Estimator

Sandro Gassner

01.05.2022

1 Setting

This self-study is intended to compare the two estimation methods of Ordinary Least Squares (OLS) and Doubly Robust Estimator (DRE) by means of a Monte Carlo simulation study. This means we define a data generating process (DGP) based on random variables where we know the true parameters. Then we repeatedly draw a sample using random number generators and apply the two estimators mentioned. Finally, we compare the estimated parameters of interest to the known truth and evaluate the performance of the estimators.

1.1 Parameters of Interest

Our variable of interest in this simulation study is the average treatment effect (ATE). The ATE is the average change in the outcome variable for individuals who received a treatment, compared to the average change in the outcome variable for those who did not receive the treatment. This can be written as $ATE = E(Y^1 - Y^0)$ where Y^1 is the potential outcome of the treated individual and Y^0 is the potential outcome of the untreated individual (control group). If we condition the outcome on different covariates to avoid potential selection bias problems, the ATE can be written as:

$$ATE = E[E\,[Y \mid X = x, D = 1] - E[Y \mid X = x, D = 0]]$$

1.2 Identification Strategy

The underlying research design of this self-study is called selection-on-observables. This means, all confounding variables can be observed. A confounding variable is a variable that is not the outcome or the treatment variable but affects the outcome of the study. To avoid a potential bias, we must control for such confounders. To obtain valid results from our estimations, the following identification assumptions must hold:

- **Conditional Independence Assumption (CIA):**
  The potential outcomes are conditionally independent of treatment for any given values of the confounding variables: $Y^0, Y^1 \perp\!\!\!\perp D \mid X = x;\ \forall\, x \in \mathcal{X}$
- **Common Support:**
  For any given value of the confounding variables, a unit could potentially be observed in both treatments: $0 < P(D = 1 \mid X = x) < 1;\ \forall\, x \in \mathcal{X}$
- **Exogeneity of Confounders:**
  The confounding variables are not influenced by the treatment in a way that is related to the outcome variables: $X^d = X^{1-d}$
- **Stable Unit Treatment Value Assumption (SUTVA):**
  The observed outcomes in one treatment state correspond to the potential outcomes of that state for the participants in that state: $Y = DY^1 + (1 - D)Y^0$

1.3 Estimators

Like already mentioned, we want to compare Ordinary Least Squares (OLS) to Doubly Robust Estimation (DRE). Subsequently, a short explanation of these two estimation approaches.

1.3.1 Ordinary Least Squares (OLS)

OLS is a method for estimating the unknown parameters in a linear regression model. It chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable in the given dataset and those predicted by the linear function of the independent variable.

Under the assumptions that the effects are homogeneous and that the conditional outcomes depend linearly on confounders, OLS is a consistent estimator of the ATE.

$$E(X \mid D = d, X = x) = \mu(d, x) = d\alpha_0 + x\beta_0$$
$$\Rightarrow ATE = E[\mu(1, x) - \mu(0, x)] = \alpha_0$$

1.3.2 Doubly Robust Estimator (DRE)

Doubly Robust Estimation is a way of combining the two estimation approaches of propensity score weighting and linear regression in a way we do not have to rely on either of them. This means, it only requires one of the models to be correctly specified. To do this, we need to follow the subsequent procedure:

- Step 1: Estimate the propensity score (probability of being treated given a set of observed covariates) by, e.g., probit or logit model: $\widehat{p(x_i)}$
- Step 2: Estimate the outcome equation by parametric model: $\widehat{\mu(1, x_i)}$, $\widehat{\mu(0, x_i)}$
- Step 3: Calculate the estimated ATE by the following formula:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\mu(1, x_i)} - \widehat{\mu(0, x_i)} + \frac{d_i\left[y_i - \widehat{\mu(1, x_i)}\right]}{p(x_i)} - \frac{(1 - d_i)\left[y_i - \widehat{\mu(0, x_i)}\right]}{1 - p(x_i)}\right.$$

The first part of the formula hereby corresponds to a simple OLS regression and the second part corresponds to the propensity score part, which also can be seen as a correction term. If OLS exactly estimates the outcome variable, the correction term becomes zero and wipes out the relevance of the propensity score model as $y_i - \widehat{\mu(1, x_i)} = 0$ and $y_i - \widehat{\mu(1, x_i)} = 0$. On the other hand, if the OLS model is mis-specified and the propensity score model is correctly specified, the correction term becomes unequal to zero and wipes out the relevance of the OLS model. This way either the OLS or the propensity score model can be mis-specified and DRE still delivers consistent results.

2 Simulation Design

The simulation study is based on three different data generating processes, that randomly generate datasets with 1'000 observations. Afterwards, the estimators mentioned above were applied to estimate the ATE. This procedure was performed 1'000 times with the following data generating processes:

2.1 Data Generating Processes (DGPs)

2.1.1 DGP 1

In the first data generating process all identifying assumptions must hold and OLS should perform better than DRE. In particular, the data was generated in the following way:

$$y = 4 + 3d + 5x_1 - 2x_2 - x_3 + u$$

such that:

$$d = \begin{cases} 1 \ if \ 5x_1 - 2x_2 + 0.5x_3 > 0 \\ 0 \ if \ 5x_1 - 2x_2 + 0.5x_3 \leq 0 \end{cases} \qquad x_3 \in \{0; 1\}$$

$$[x_1, x_2] \sim N(\mu, \Sigma); \qquad u \sim N(0, 5)$$

$$\mu = (2 \quad 3); \ \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad E[ATE] = 3$$

For this data generating process the outcome variable depends linearly on the three covariates x_1, x_2 and x_3 and the binary treatment variable d. The covariates x_1 and x_2 represent a multivariate normal distribution. This means, that the covariates x_1 and x_2 are correlated. The covariate x_3 is a randomly generated number between 0 and 1 and the error term u is normally distributed with an expected value of 0 and a standard deviation of 5. The treatment variable d is affected by all three confounders x_1, x_2 and x_3. Based on these parameters, we know that the true average treatment effect is 3.

2.1.2 DGP 2

In the second data generating process all identifying assumptions must hold and DRE should perform better than OLS. For this example, I went for the following data generating process:

$$y = dy^1 + (1 - d)y^0$$

such that:

$$d = \begin{cases} 1 \ if \ 0.4 + 0.2x_1 - 2x_2 + x_3 > 0 \\ 0 \ if \ 0.4 + 0.2x_1 - 2x_2 + x_3 \leq 0 \end{cases} \qquad x_2 \sim N(0, 2)$$

$$y^1 = -2 + 2x_1 + x_2 + 2x_3 + u \qquad x_3 \sim U(1, 3)$$

$$y^0 = 0 + u \qquad u \sim N(0, 2)$$

$$x_1 \in \{0; 1\} \qquad E[ATE] = E[y_1 - y^0] = 3$$

Like in DGP 1, the outcome variable is linearly affected by the three covariates x_1, x_2 and x_3. This time, the covariates are not correlated to each other, x_1 represents a random number between 0 and 1, x_2 is normally distributed with a mean of 0 and a standard deviation of 2 and x_3 is uniformly distributed between 0 and 2. The binary treatment variable is defined by an indicator function that depends linearly on x_1, x_2 and x_3 and the error term u is normally distributed with an expected value of 0 and a standard deviation of 5. The true average treatment effect in this example is also 3.

2.1.3 DGP 3

In the third data generating process one identifying assumption should be violated. Hereby, the data was generated as follows:

$$y = 2 + 3d + 3x_1 - 2x_2 + x_3 + u$$

such that:

$$d = \begin{cases} 1 \ if \ 2x_1 - x_2 - x_3 > 0 \\ 0 \ if \ 2x_1 - x_2 - x_3 \leq 0 \end{cases} \qquad x_3 \sim N(0,2)$$

$$x_1 \sim binomial(1, 0.5) \qquad u \sim N(0,2)$$

$$x_2 \in \{0; 1\} \qquad E[ATE] = 3$$

Again, the three covariates x_1, x_2 and x_3 affect the outcome and the treatment variable linearly. The covariate x_1 is equal to 1 with a probability of 50 percent, x_2 is a random number between 0 and 1 and x_3 is normally distributed with a mean of 0 and a standard deviation of 2. Like in the previous examples, the true ATE is equal to 3. Important: For the estimation of the treatment effect, I decided to omit the variable x_2 in this DGP. This means, I ran my OLS and DRE functions without specifying x_2 as an exogenous variable.

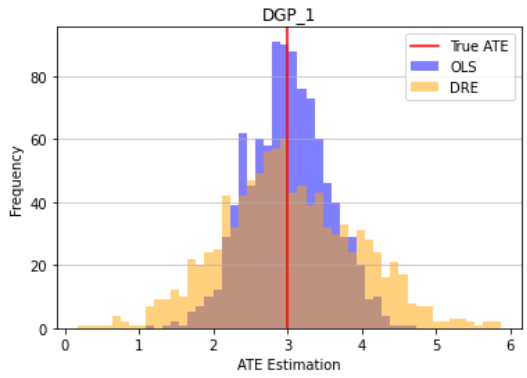
2.2 Expectations

I expect OLS to run better than DRE for the first DGP, since it is a linear model with a homogeneous treatment effect and there is only little common support. The opposite is true for DGP 2 since the treatment effect is not homogeneous anymore in this example. This could potentially lead to a situation, where DRE delivers better results. In the third and last DGP, both OLS and DRE are likely to perform badly, since I omitted a confounder, which is affecting the outcome as well as the treatment variable. This represents a clear violation of the identification assumption of conditional independence (see chapter 1.2).

3 Results

The results of the 1'000 simulations are listed below. For each data generating process I am going to compare the performance measures (bias, variance and mean squared error) of the two estimators and give an interpretation of the results.

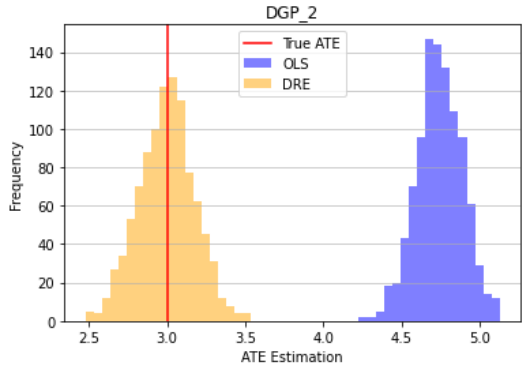
3.1 DGP 1



	OLS	DRE
Bias	0.0072	0.0180
Variance	0.2894	0.8371
MSE	0.2894	0.8374

Like expected, OLS delivered better estimates than DRE in this example. We can observe this in all three performance measures. The bias of OLS is about 2.5 times lower than that of DRE. Also, the variance of 0.2894 is much lower compared to the variance of 0.8371 for DRE and the MSE is lower for OLS. We can also observe this difference in the histogram, where the distribution of the estimated ATEs is much narrower around the true ATE (red line) for OLS (blue) than for DRE (orange). Therefore, OLS might be the better choice for estimating linear models with a homogeneous treatment and little common support.

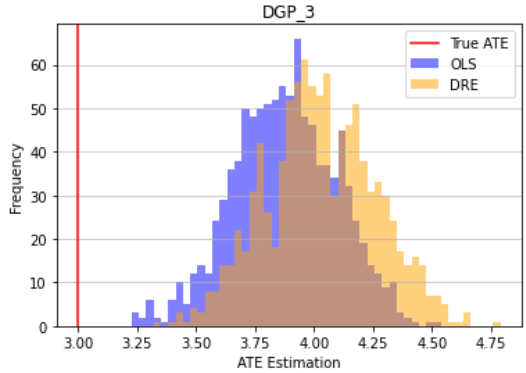
3.2 DGP 2



	OLS	DRE
Bias	1.7426	0.0048
Variance	0.0217	0.0311
MSE	3.0584	0.0311

In this case, we see a very large difference between our two estimators. Although the variance does not differ largely, the bias of 1.7426 of OLS seems huge compared to the bias of 0.048 for DRE. This can also be seen in the very large MSE of OLS. The histogram reveals that OLS consistently overestimated the average treatment effect. These results are in line with my expectations in chapter 2.2. Hence, for treatment effects that are not homogeneous, DRE will most likely be the better choice.

3.3 DGP 3



	OLS	DRE
Bias	0.8779	1.0301
Variance	0.0479	0.0541
MSE	0.8187	1.1153

As already mentioned, in this final DGP I violated the identification assumption of conditional independence by omitting the covariate x_2. As stated in the table and the plot above, I was right to believe that both estimators will perform badly. Although OLS performed slightly better, both estimators overestimated the average treatment effect. We can see this in the bias of 0.8779 for OLS and 1.0301 for DRE respectively. Therefore I was able to confirm that it is crucial to make sure that the identification assumptions hold.