University of St. Gallen
Kevin Hardegger (12-758-78)

Data Analytics 2

Self-Study Project

1. Intro

In this project I test the performance of two different estimation methods namely the Inverse Probability Weighting (IPW) and Doubly Robust estimation (DR). The two estimation methods are applied on three different data generating processes in a Monte Carlo simulation which is repeated one hundred times. I analyze how performance varies with increasing sample sizes starting from 100 up to 1000 with increments of 50 resulting in 19 different sample sizes and results.

1.1 Parameter of Interest

The parameter of interest is the average treatment effect (ATE). Due to the missing true counterfactual, we have four necessary assumptions that must hold for our ATE to be correctly estimated:

i. **Conditional Independence Assumption** (**CIA):** We observe all variables X that affect treatment D and (potential) outcome Y. If D is randomly assigned or the information for assignment is within observation CIA should hold. This project will violate this assumption in the third and last data generating process.

ii. **Common Support:** Observed covariates X have to take similar values for both treated (D=1) and untreated (D=0) groups. There should be no value of X that can only be realized in one group.

iii. **Exogeneity of Confounders:** Variables X mustn't be dependent of outcome Y.

iv. **Stable Unit Transfer Value** (**SUTVA**): Potential outcome of a unit's treated is not affected by assignment of treatment to other units. Furthermore, there are no different levels of the treatment.

1.2 Estimator

We apply two different estimators Inverse Probability Weighting and Doubly Robust estimation. IWP and DR are similar in implementing propensity score matching. In other words, the estimators evaluate the probability of observations having been assigned with the treatment by grouping the units according to similar covariate values and uses this probability as a weight. The Doubly Robust is an augmented version which introduces a regression mode as well. Thus, DR contains two estimation models and is more robust; it should be able to deliver proper estimations even if one model is not correctly specified. Hence the name doubly robust. The formula for calculating the ATE estimations are as follow:

Estimator A (IPW):

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

Estimator B (DR):

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

1.3 Data Generating Processes

The data generating process results in observations for a population size of 1000.
Main specifications are:
Outcome: $Y = D\alpha + X\beta + u,\ where\ \ u{\sim}N(0{,}3)$
True ATE: $\alpha =\ 2$
Beta Vector: $\beta = (1, 5, 8),\quad where\ first\ entry\ is\ intercept$
Mean Vector: $\mu = (5, -5)$
Variance-Covariance matrix: $Cov = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}$

DGP1:
In the first DGP all assumptions hold. Treatment is randomly assigned ($condition \sim N(0{,}1)$): ($D = 1 \mid condition\ \geq\ 0$). We assume DR to perform better because the models are correctly specified and DR applies OLS in its estimation, which should be able to help correctly identifying ATE, in addition to the propensity score matching.

DGP2:
In the second DGP treatment is dependent on X and thus non-random ($condition\ =\ X\beta\ +\ e$): ($D = 1| \ condition \geq 0$). To ensure common support we apply a large error term ($e\ {\sim}N(0{,}20)$). We assume IPW to perform better, due to the non-randomness in treatment assignment leading to an overestimation of ATE in the naïve OLS approach in DR. Yet, since DR is a robust estimation method, in other words, even though the regression may be incorrect, the propensity model is correctly specified. Thus, the ATE estimation of DR doesn't necessarily have to be far off.

DGP3:
In the last DGP treatment we introduce confounder V which both affects outcome Y and treatment D ($condition\ =\ V$). By this CIA is heavily violated and leads to biased estimators:
($D = 1 \mid condition\ \geq 0$), $\quad Y = D\alpha + X\beta + V\ +\ u, \qquad where\ \ V\ {\sim}N(-1{,}3), u{\sim}N(0{,}3).$
Since CIA is violated, we expect heavy omitted variable bias for both estimators.

2. Results

2.1 Data Generating Process 1:

Table 1: Results DGP1

	Mean ATE	Mean SE	Mean Bias	Mean MSE
Estimator A (IPW)	1.857	0.307	-0.143	0.144
Estimator B (DR)	1.928	0.412	-0.072	0.204

In the first DGP the results are mixed. The ATE estimation of DR (1.928) is nearer to the true ATE (2). However, the standard error for DR (0.412) is larger indicating suboptimal precision. This can also be observed in ***Figure 2***, compared to ***Figure 1*** the estimates are more scattered on the left and right tail. The reason for this can be observed when comparing ***Figure a*** and Figure b as we can determine that DR starts with much higher variance for lower sample sizes. ***Figure a*** also indicates that MSE of IPW doesn't follow the variance as smoothly compared to the MSE of DR, which indicates more volatility. Nevertheless, with increasing sample sizes the results of both IPW and DR converge.

Figure 1: Histogram DGP1 & IPW

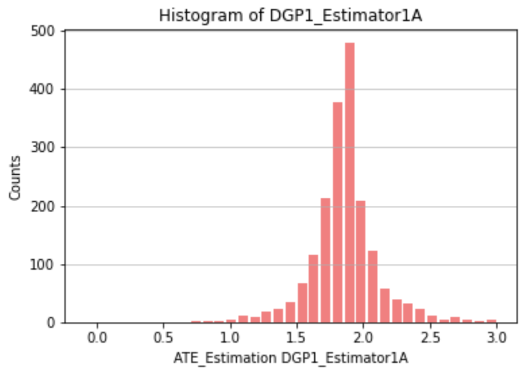


Figure 2: Histogram DGP1 & DR

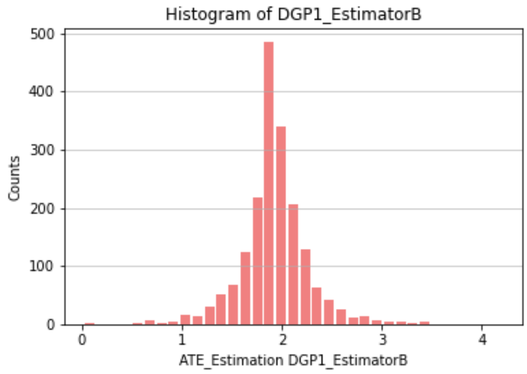


Figure a: Performance DGP1 & IPW

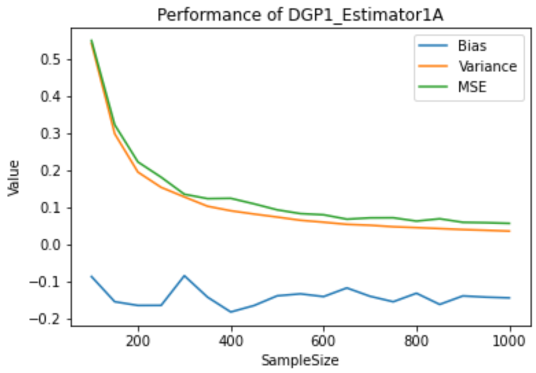
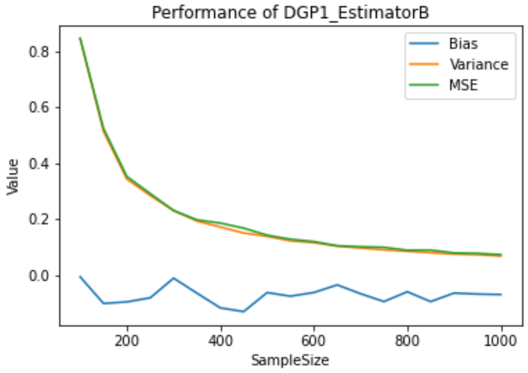


Figure b: Performance DGP1 & DR



2.2 Data Generating Process 2:

Table 2: Results DGP2

	Mean ATE	**Mean SE**	**Mean Bias**	**Mean MSE**
Estimator A (IPW)	4.401	2.875	2.401	14.755
Estimator B (DR)	4.439	2.937	2.439	15.488

In the second DGP both estimators possess very close end results with IPW having a very slight edge with a lower mean ATE (4.401). The similarity of the estimates is clearly visible in ***Figure 3*** and ***Figure 4.*** This parallelism can also be seen in the performance in **Figure c** and **Figure d.** Like in DGP1 we can identify higher variance for DR in lower sample sizes. However, the results for both quickly converge with increasing sample sizes, which shows the strength of the double robustness of DR. According to these results, the differences of the two models are neglectable and not significant.

Figure 3: Histogram DGP2 & IPW

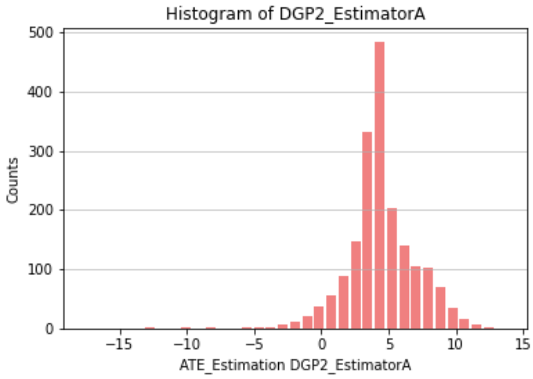
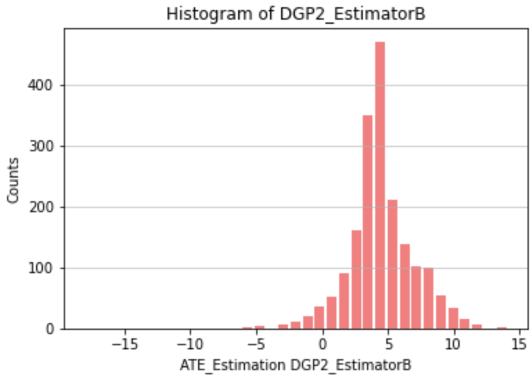


Figure 4: Histogram DGP2 & DR



b) Performance

Figure c: Performance DGP2 & IPW

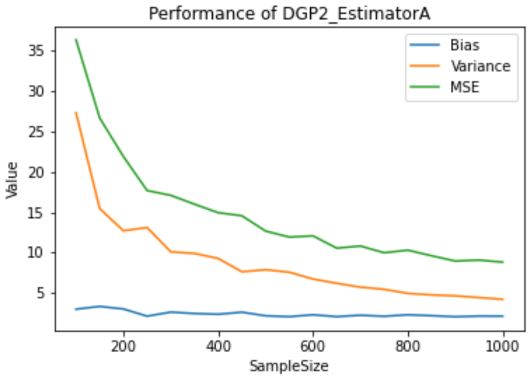
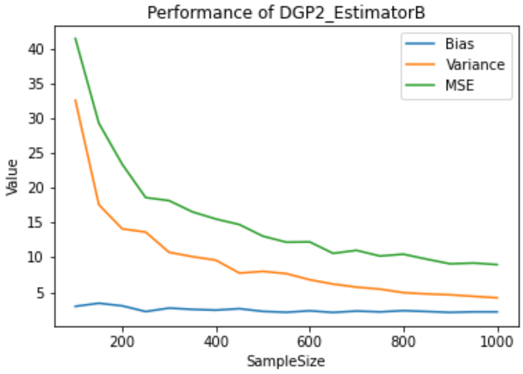


Figure d: Performance DGP2 & DR



2.3 Data Generating Process 3

Table 3: Results DGP3

	Mean ATE	Mean SE	Mean Bias	Mean MSE
Estimator A (IPW)	6.894	0.393	4.894	24.144
Estimator B (DR)	6.557	0.485	4.557	21.051

In the last DGP DR performs better. Similarly, to DGP2 the shape of the figures for both estimators are nearly identical. It's clear that since the models are incorrectly specified and thus CIA violated, our estimators absorb huge bias due to omitted variable bias caused by the missing confounder V. The mean estimated ATE thus settles far above the true ATE (between 6.557 to 6.894). This is also represented in **Figure e** and **Figure f**, as even with increasing sample sizes, the bias remains high and does not seem to improve at all even after reaching the whole population. Still, DR seems to be able to respond slightly to this problem with its augmentation and performs better than IPW.

Figure 5: Histogram DGP3 & IPW

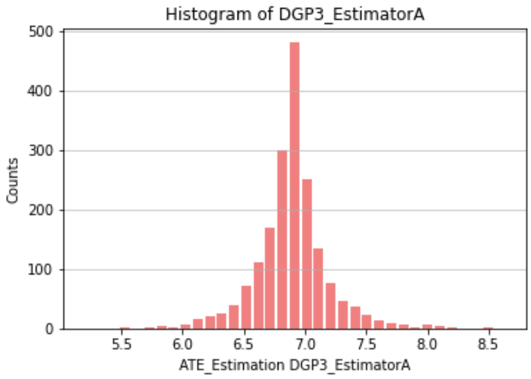


Figure 6: Histogram DGP3 & DR

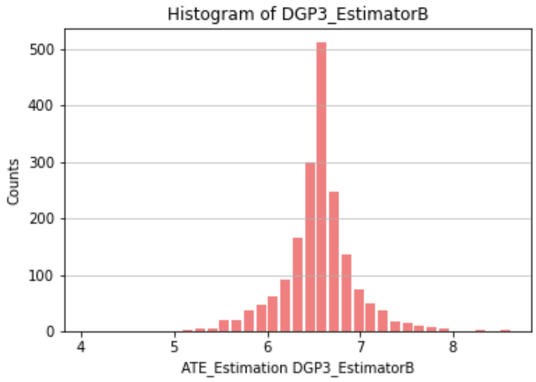


Figure e: Performance DGP3 & IPW

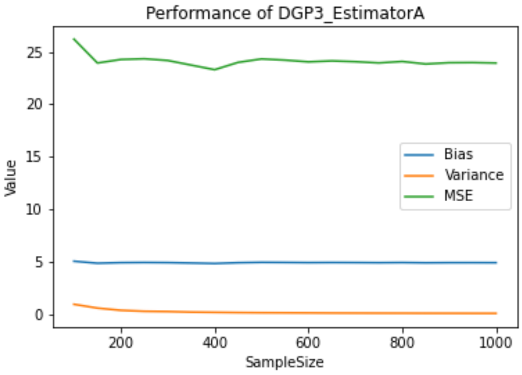
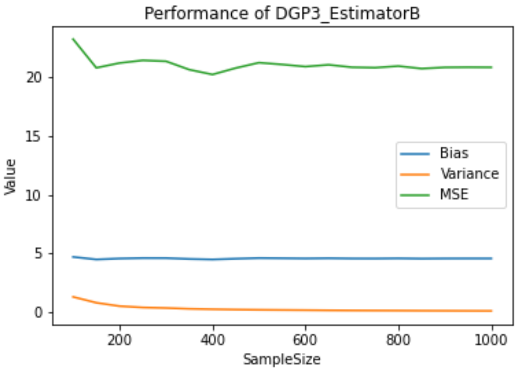


Figure f: Performance DGP3 & DR



In conclusion, IPW and DR behave quite similarly. Even though the augmentation in DR should reduce variability this has only happened in the last DGP. It has also become clear that DR suffers from lacking precision with small sample sizes which is caused by the robustness of the estimation method. Nevertheless, when given a large enough sample size I would recommend the Doubly Robust estimator, as most often than not, models aren't correctly specified when conducting research and DR has shown the ability to counter the negative impact of such a flaw.