Data Analytics II: Causal Econometrics
Simulation Study

Jonas Schmitten

May 2021

Setting

In this simulation study I decided to explore the limitations and properties of the ordinary least squares estimator (OLS), and inverse probability weighting estimator (IPW) under different data generating processes. The aim is to estimate the average treatment effect (ATE) with both estimators using three different data generating processes and a Monte Carlo simulation. Furthermore, I compare the variance of the estimators to compare efficiency and to analyse convergence in both small and large samples.

The identification strategy is based on the conditional independence assumption (CIA) which rests on the assumptions that the potential outcomes are conditionally independent of treatment (e.g. no confounding observables), that for any value of the confounding variables we can observe the treatment or not (common support), treatment does not influence the confounding variables (exogeneity), and that the potential outcome of one observation should be unaffected by the treatment assignment to another (SUTVA). The first two data generating processes fulfill those conditions.

As for the estimators, the OLS is estimated as $\hat{\beta} = (X'X)^{-1} X'Y$ where X refers to the matrix of exogenous variables (including the treatment) while Y is the outcome variable. I chose the OLS as a benchmark given its well-understood properties and limitations. It is important to note that one of the main assumptions of the OLS linearity in the coefficients. IPW allows us to calculate statistics such as the ATE from a pseudo-population different from the population from which the data was collected, i.e., if the assignment of treatment is disproportional (not randomly). It applies an alternative weighting scheme to counter bias as opposed to unweighted estimators such as OLS and may improve efficiency. The estimation is done in two steps. First, I estimate the propensity score using a logit model. Then, the ATE is calculated using the following formula:

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

where d_i is the treatment variable, x_i the matrix of exogenous variables, y_i the outcome variable, and $\widehat{p(x_i)}$ the propensity score computed via the logit model in the first step.

The performance is measured by the variance of the estimator while the parameter of interest is the ATE, which basically measures how much the treatment affected the potential outcome, on average, i.e., the difference of coefficient means between treatment and control group.

Simulation Design

I generate three different data generating processes with common pre-specified means, covariance matrix, and betas which I define as:

$$\mu_{ij} = (12, 0)$$

$$\sum = \begin{bmatrix} 9 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\beta_i = (5000, 100, 300, 1000)$$

The first two data generating processes are created from multivariate normal exogenous variables and a binomial treatment variable as follows:

$$\mathbb{E}(y) = \beta_0 + \beta_1 \cdot \vartheta + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

$$y_i = \beta_0 + \beta_1 \cdot \vartheta + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \epsilon_i$$

where the error term is deterministic with

$$\epsilon_i = 80$$

and $\vartheta \sim B(n, p)$ the treatment variable for different sample sizes n and $p = 0.5$ for the first and $p = 0.1$ for the second data generating process, with p indicating the probability of receiving treatment. The binomial probabilities for treatment assignment were chosen to test if the IPW outperforms the OLS in situations were alternative weighting is more important, i.e., the treatment is unequally assigned. For both data generating processes, I would expect OLS to outperform in small sample sizes, with a lesser difference in large samples, while the IPW should outperform in large samples for $p = 0.1$.

Finally, I introduce a third data generating process which violates the common support assumption. To do this, I scale up one exogenous variable after the treatment is assigned in the following way:

$$x_1 = \begin{cases} x_1 \cdot c, & \text{if } x_0 = 1 \\ x_1, & \text{otherwise} \end{cases}$$

where c is a positive scaling constant chosen in a way to avoid perfect separation. This allows me to *stretch* the data and break the common support assumption of the identification strategy. One can easily verify this by plotting the propensity scores of the IPW estimator conditional on the treatment, which would show a high concentration of propensity scores around 0 and 1 (with a *valley* in between). It is important to note here, it is not possible to fully separate the two groups, otherwise the logit model used to calculate the propensity scores in the first step would not work due to perfect separation. For instance, assigning the treatment conditional on the mean of one of the variables would not work. Generally, I would expect the IPW to do worse given the use of maximum likelihood estimation (MLE) and the blowing up of the variance as a result of low propensity scores.

Results

I ran a Monte Carlo simulation with 100 iterations per sample size. To compare also small and large sample properties, the sample size increases from 100 to 5000 in increments of 100. I chose a starting sample size of 100 to avoid issues resulting from a singular matrix (treatment column is 0, leads to 0 determinant), which then does not allow us to compute the inverse and get the OLS coefficients. For each sample size the ATE and variance are then computed and graphed to compare both estimators across data generating processes and sample sizes. For simplicity, the main focus is on the variance in the graphs.

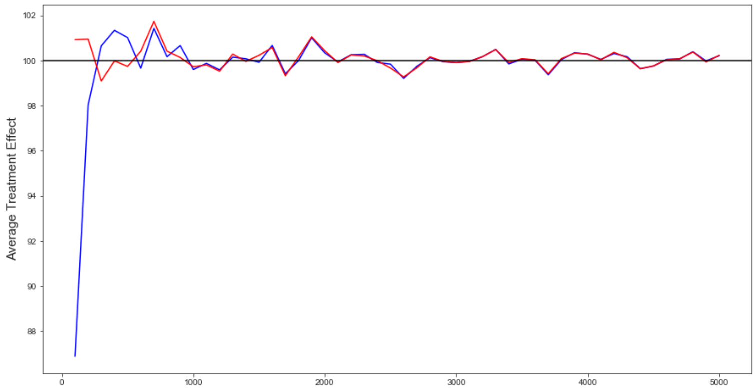

Figure 1: Average Treatment Effect OLS and IPW $\vartheta \sim B(n, 0.5)$

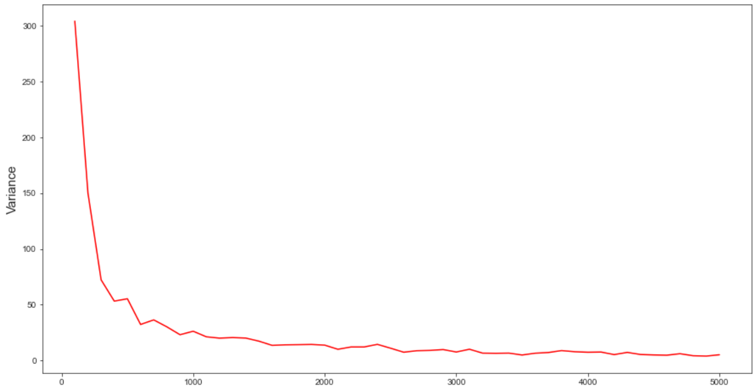

Figure 2: Variance OLS $\vartheta \sim B(n, 0.5)$

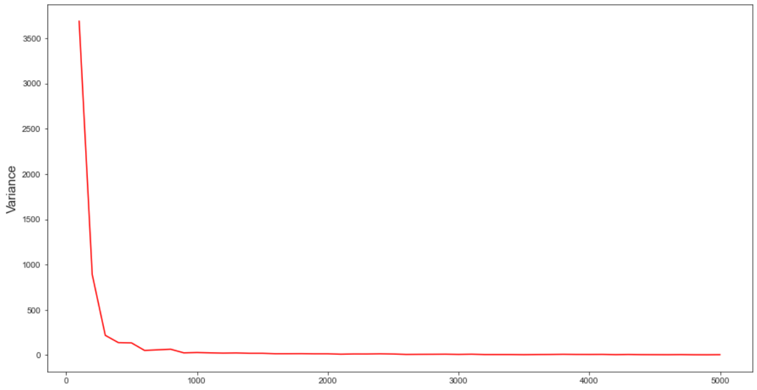

Figure 3: Variance IPW $\vartheta \sim B(n, 0.5)$

The ATEs under the first data generating process are graphed in the first figure. It is obvious that the OLS in red performs closer to the black horizontal (correct ATE under assumption that $\beta = 100$) on smaller sample sizes. However, IPW and OLS perform similar in terms of ATE once the sample size increases. In terms of variance in figures two and three, the OLS performs better than the IPW up until $n \sim 2,500$ when both estimators are similar, with OLS only slightly outperforming. Unsurprisingly, both variances decrease fast as the sample size increases while IPW has a larger variance in small samples.

For the second data generating process, the assignment probability is decreased to $p = 0.1$. I expected that IPW would perform better than OLS given the unequal assignment. However, OLS outperforms again in terms of variance for small and slightly for large samples (figure four). It may be that the OLS outperforms in both settings given the linearity of the data generating processes with IPW having no such assumption. Interestingly, though, the variance of IPW starts off much higher than in the first data generating process and seems to decrease much quicker than previously.

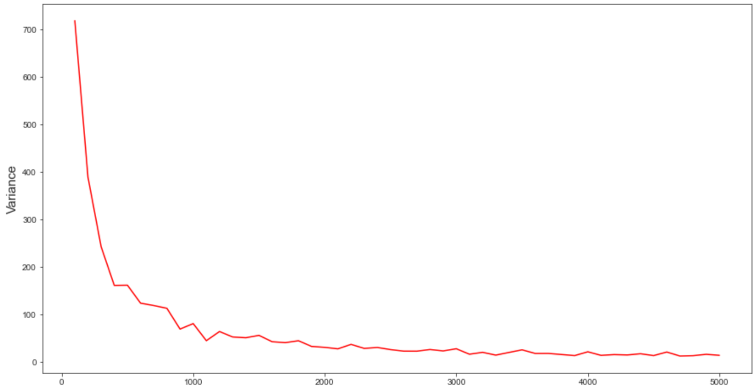

Figure 4: Variance OLS $\vartheta \sim B(n, 0.1)$

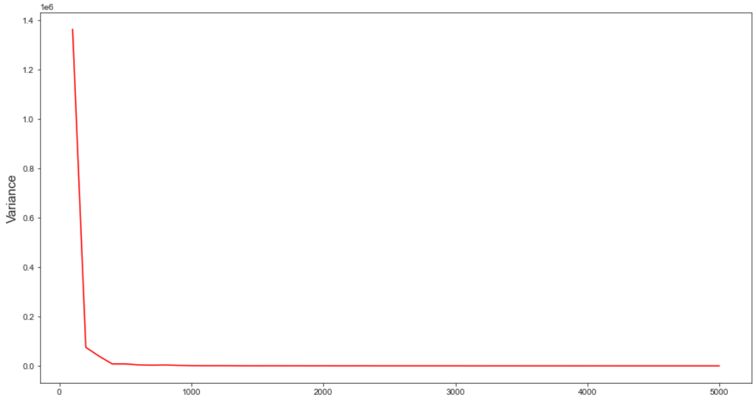

Figure 5: Variance IPW $\vartheta \sim B(n, 0.1)$

For the third data generating process, the common support assumption is violated. I only present the variance and ATE of the IPW here because OLS did not change much. This may be the result of the general robustness of the OLS estimator or fact that it does not depend on propensity scores which are a function of the treatment variable. As for the IPW, ATE and variance are shown in figures six and seven, respectively. The variance does not decrease and exhibits random increases throughout different sample sizes while the ATE remains significantly above the true ATE line. This is in line with expectations given that the maximum likelihood blows up the variance as a result of the propensity scores clustered around the extremes.

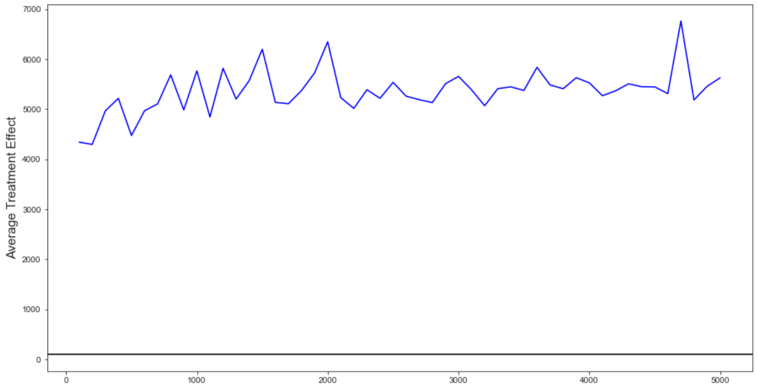

Figure 6: ATE IPW with violation and $\vartheta \sim B(n, 0.5)$

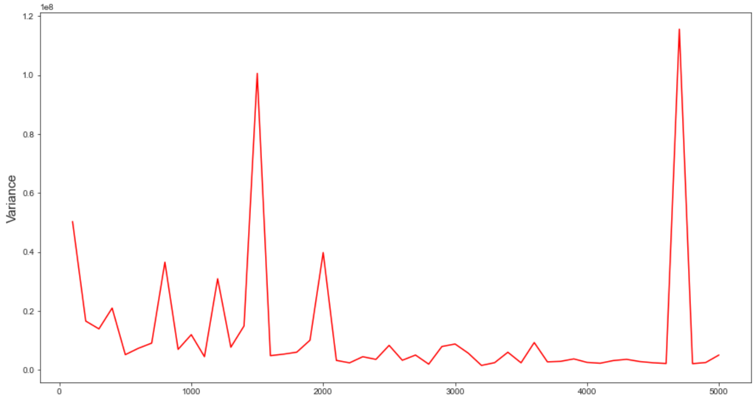

Figure 7: Variance IPW with violation and $\vartheta \sim B(n, 0.5)$

To conclude, OLS performs much better on smaller sample sizes than IPW while the difference decreases significantly fast with increasing sample sizes. In the first two data generating processes, the estimators perform similar in large samples. However, in the last data generating process the IPW fares much worse which is most likely the result of clustered propensity scores which blow up the variance through the MLE.