HUL315 Assignment-5

Amaiya Singhal 2021CS50598

Objective

In this assignment we are trying to practically figure out the bias caused due to omitted variables. We assume a true model and construct a variable that is dependant on other variables. Then we try to perform OLS on the generated data (using our true model) and analyse the effect of omitted variables as well the effect of using proxy variables.

Model

We construct our own true model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Coefficient	Value
β_0	2.8
β_1	1.2
β_2	6.7
β_3	4.3

Variables x_1 and x_2 are randomly initialised using a random number generator in python and have a sample size of 500. A third variable x_3 is constructed using the following model:

$$x_3 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + v$$

Coefficient	Value
δ_0	0.5
δ_1	2
δ_2	1.5

Both u and v are generated using a normal distribution with mean 0 and variance 1. Using these and the randomly generated x_1 and x_2, we generate x_3 using the above equation and then we generate y using the true model.

Part 1: Regressing over the True Model

In this part we run OLS on all the variables x_1, x_2 and x_3. The results we obtained are as follows:

Coefficient	Actual	Predicted
β_0	2.8	2.815197686373722
β_1	1.2	1.2008089989831205
β_2	6.7	6.700001502365922
β_3	4.3	4.299734481450287

We observe that the coefficient estimates are very close to the true values.

Part 2: Omitted Variable Bias

In this part we omit the variable x_3 and regress using the following equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Coefficient	Actual	Predicted
β_0	2.8	2.6780145311833286
β_1	1.2	9.979993748512737
β_2	6.7	13.005755828661748

Observations:

In this case we observe that the OLS estimates get farther from the true values of the parameters and there is a positive bias for each estimate except for β_0 where we get a negative bias.

Changing Signs of δs and βs

Now we flip the sign of the original δs and βs and again run regression as done above.

Coefficient	Value
δ_0	-0.5
δ_1	-2
δ_2	-1.5

Including x_3

Coefficient	Actual	Predicted
β_0	-2.8	-2.8586880020066587
β_1	-1.2	-1.2062620549013445
β_2	-6.7	-6.687338288467572
β_3	-4.3	-4.299644033346681

In this case again we observe that the regression with all variables gives estimates which are very close to the true value of the coefficients.

Excluding x_3

Coefficient	Actual	Predicted
β_0	-2.8	-0.8897878756087648
β_1	-1.2	7.773706010266436
β_2	-6.7	-0.4739158446978757

The results obtained in this part are similar to the case without reversing the signs of the coefficients. Except for β_0 which now has a positive bias, we observe that the biases in the values of β_0 and β_1 both are positive again and also the bias in both is almost the same, irrespective of the sign of βs.

This happens because of the signs of βs and δs. If we do not include a variable in the regression which is correlated with another variable then the coefficient of that variable gets biased in the direction given by the sign of the product of the corresponding β and δ.

Class Question

$$log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u$$

As we have observed above the bias is depending on the values of βs and δ_s. Since worker ability has a negative impact on the variable train and it has positive impact on wage, had we taken into account the unobserved working ability, we would have gotten an estimate that was higher than the value that we have observed when not taking it into account. Hence there is a negative bias in the estimated value of β_1

Part 3: Effect of Proxy Variable

In this part we use a proxy variable for x_3 assuming that x_3 is unobservable. We construct a proxy variable z_3 which satisfies the condition $E[x_3|x_1, x_2, z_3] = E[x_3|z_3]$. We first construct z_3 randomly then generate x_1, x_2 and x_3 as follows:

$$x_1 = \theta_{10} + \theta_{11} x_3 + \epsilon_1$$
$$x_2 = \theta_{20} + \theta_{21} x_3 + \epsilon_2$$
$$x_3 = \theta_{30} + \theta_{31} z_3 + \epsilon_3$$

Coefficient	Value
θ_{10}	2.8
θ_{11}	1.2
θ_{20}	6.7
θ_{21}	4.3
θ_{30}	6.7
θ_{31}	4.3

y is generated using the true model and regression is performed for 3 cases : the true model equation, omitting x_3 and using proxy variable z_3 in place of x_3. The results are as follows:

Coefficient	Actual	With x_3	Without x_3	With z_3
β_0	2.8	2.799134092256281	1.5190558991515624	3.8633557652428863
β_1	1.2	1.2137386040712954	2.681639066775568	2.140957722800522
β_2	6.7	6.773510178318247	9.29416545239102	8.474509623003541
β_3	4.3	4.200238380974042	NA	1.5578265893163916

Observations

- The results with using the true model equation are very close to the actual values of the coefficients. This aligns with our observations in the previous parts.
- Similarly in the case of omitting the variable x_3, our estimates of the coefficients are off from the true values showing omitted variable bias in the positive direction again similar to previous parts.
- What we observe when using the proxy variable z_3 is that the estimates of both β_1 and β_2 are better in this case however still far from the true values. Also β_0 in this case shows positive bias compared to negative bias when using the true model equation.

Violating Proxy Conditions

In this subpart we generate x_3 differently and violate the conditions for the proxy variable i.e. $E[x_3|x_1, x_2, z_3] \neq E[x_3|z_3]$.

$$x_3 = \theta_{30} + \theta_{31} z_3 + \gamma_1 x_1 + \gamma_2 x_2 \epsilon_3$$

We observe the following results:

Coefficient	Actual	With x_3	Without x_3	With z_3
β_0	2.8	2.799134076150949	1.5190558992881702	3.8633557651628507
β_1	1.2	1.3833333558868617	9.99163906677209	9.450957722801832
β_2	6.7	6.86329563439358	13.16416545238911	12.344509623013437
β_3	4.3	4.200238381512463	NA	1.5578265893127536

Observations

- The results with using the true model equation are very close to the actual values of the coefficients. This is expected.
- In the case of omitted variable, the estimates of the coefficients are significantly worse than the previous scenario when we had generated x_3 differently.
- What we observe when using the proxy variable z_3 which does not follow the necessary conditions is that the estimates are only slightly better than the omitted variable case but still far from the values that we obtained when using a valid proxy variable.

Hence we observe the importance of following the necessary conditions for a proxy variable if we want better estimates of our parameters as not following them can lead to much worse results.

Code

- Python Notebook : Link