HUL315: Econometrics Methods

Assignment 5

Indian Institute of Technology Delhi
Maximum Marks: 10 Marks

Instructions:

1. Deadline for submission is **20th April, 2024**.
2. You need to upload your assignment on Moodle in the following format (Name_Entry Number).
3. Assignments have to be submitted only in PDF generated using LaTeX.
4. You are **not** required to submit the codes along with the assignment.

---

1. This assignment is the continuation of the last surprise test in class. For your reference, I append the question here. This assignment is the continuation of the last surprise test in class. For your reference, I append the question here :
   To test the effectiveness of a job training program on the subsequent wages of workers, we specify the model

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

   where train is a binary variable equal to unity if a worker participated in the program. Think of the error term u as containing unobserved worker ability. If less able workers have a greater chance of being selected for the program, and you use an OLS analysis, what can you say about the likely bias in the OLS estimator of β_1 ?

   Many of you have answered the question theoretically/intuitively. However, there is a practical way of comprehending the potential bias due to omitted variables. We shall construct our own data assuming a (true) model. Then verify how omitting a relevant variable biases the estimates. Note that simply excluding a variable does not cause potential bias in OLS estimates; it has to be correlated with any of the included explanatory variables. Follow the steps outlined here. You may use any programming language.

   - Construct two variables x_1 and x_2 (using any random number generator) with sample size of 500 (i.e. two vectors of size 500). Now, construct a third variable, x_3 which is corelated with either x_1 or x_2 or both. You may assume a model like $x_3 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + v$. First, you generate the random variable v from a Gaussian distribution (mean zero and some constant variance) and then use the previously drawn x_1 and x_2 and v to construct x_3. You are free to choose your parameter values δ_0, δ_1, δ_2. Note down the values of the parameters. You shall play around these parameter values later.
   - Generate the values of the dependent variable, y . Assume a true model:

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

     First, randomly draw u from a Gaussian distribution with mean and some constant variance. Then generate y using some specific values of the parameters β_0, β_1, β_2 and β_3. Once you have the data on y, x_1, x_2 and x_3 , you will be able test how close your OLS estimates are to the true parameter values. Test it!
   - Now we shall verify the omitted variable bias. Run the following regression omitting the variable x_3 (which is correlated with x_1 and x_2 by construction).

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

     Do OLS estimates come closer to the true values? What are the directions of the biases? Can you redo the same exercise changing the sign of the true parameter values of δs and βs. What can you say about the likely bias in the class quiz problem?

- In this last part of the exercise, we shall use a proxy variable for x_3. Suppose x_3 is unobservable. A valid proxy z_3, must satisfy the condition, $E[x_3|x_1, x_2, z_3] = E[x_3|z_3]$. To do this, we have to reconstruct the independent variables x_1, x_2 and x_3 again (and also the y as per our true model). First, construct a random variable z_3. Then generate x_3 as some linear function of z_3 and x_1, x_2 as functions of x_3. For example,

$$x_1 = \theta_{10} + \theta_{11} x_3 + \epsilon_1$$

$$x_2 = \theta_{20} + \theta_{21} x_3 + \epsilon_2$$

$$x_3 = \theta_{30} + \theta_{31} z_3 + \epsilon_3$$

Note the x_1 and x_2 are correlated with x_3 (as before). However, if you condition on z_3, then the mean of x_3 no longer depends on x_1 and x_2. Once you generate the data (i.e y, x_1, x_2, x_3, z_3), regress y on x_1 and x_2 only (omitting x_3). Verify the omitted variable bias again. Now use z_3 as a proxy for x_3, i.e regress y on x_1, x_2 and z_3. Does it help reducing the bias? What happens when you violate the condition for the proxy i.e $E[x_3|x_1, x_2, z_3] \neq E[x_3|z_3]$? Check it by constructing x_3 as a function of x_1 or x_2 or both (in addition to z_3) i.e,

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