ECMA 31330 - Problem Set 1

Bruno Aravena M.

Due date: January 14th 3.30pm

Instructions: For every assignment, you will be asked to submit a document where you write up your answers and insert the figures/ tables requested for the assignment.

You will also be required to submit the code where you code your simulations and produce your answers, figures, tables, etc. You can write your code in any programming language you want, but I will be submitting solutions and presenting examples in TA sessions in R.

You are allowed to work in pairs, but I would insist that you stick to the same pair for the quarter, though you can work with someone else for your final project.

1 Introduction to Simulations

The following exercise will ask you to go through a basic simulation.

Consider the following (simple) model, which we will take as given.

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i \tag{1}$$

Assume $\beta_1 = 3, \beta_2 = 2, \beta_3 = -2$ and let x_k and u to be standard normals of length N.

Note that for each simulation s you will redraw these vectors and construct/simulate y accordingly. For simplicity, throughout the pset, we will focus in the estimation of β_1.

1. In the regressions below we will not use u as a regressor. Comment on what role does u play in this setup, why it makes sense not to include it in the regression and what would occur if we did.

2. Draw 100 samples of N = 100 observations. Run OLS on each simulation and report (plot) the distribution of $\hat{\beta}_1$ you obtain. Include β_1 in your plot.

3. Draw 100 samples of N = 5 observations. Run OLS on each simulation and report (plot) the distribution of $\hat{\beta}_1$ you obtain. Include β_1 in your plot.

4. Plot both distributions together, comment on the differences. Do you recover β_1?: What can you say about the bias and variance of these two estimators ?

5. To make this relationship clearer, we now want to repeat the process for all values of N between 5 and 200. Write a loop that draws 100 samples for each value of N. Store all the $\hat{\beta}_{1_{n,s}}$ in a matrix.

6. Plot the average $\hat{\beta}_1$, as well as those at the 5th and 95th percentile for each value of N. *(Hint: To compute values for each row or column of a matrix you might want to use the function* `apply`).

7. Discuss your results above. What do you see and how does that connect to the theory. What property of the OLS estimator is represented here?

8. Go back to the case with N = 100 (Q 1.2). Adapt your code such that you also store the 95% confidence interval of each OLS regression. How many of these confidence intervals contain the true value of β_1? Should this result surprise us?

9. (Optional but recommended: Plot these 100 confidence intervals in the same plot and distinguish those that include and do not include the true value of β_1)

2 Increasing the number of regressors

Consider now a researcher that does not know the DGP described in equation 1. They have access to three data sets: i) a "short" one, which contains (x_1, x_2, x_3, y); ii) a "medium" containing $(x_1, x_2, x_3, ..., x_{25}, y)$; and iii) a "long" one, which contains $(x_1, x_2, x_3, ..., x_{50}, y)$.

Another researcher suggests that having more data is always better and that as a result the wider dataset strictly dominates. The researchers agree to settle the issue by running simulations.

1. Write a loop that for each value of n between 5 and 100, draws 100 samples –including all x_k potential regressors – and stores $\hat{\beta}_{1_{n,s}}$ using the "short" dataset. *(Hint: in a way you have already done this before, except that now you need to draw more potential regressors.)*

2. Include in your loop code that estimates $\hat{\beta}_{1_{n,s}}$ using all data from the "medium" and "long" datasets. You should expect some of these regressions to not be computed. Note that given that we are trying to compare estimators, estimators must be computed from the same data for a given (n, s) draw.

3. Plot (on one figure) the performance of these three estimators across N values between 5 and 100. You should have the "medium" and "long" estimators only for the N values for which it makes sense. Use different

colors for each estimator. *(Hint: You are trying to summarize a distribution of estimators for each N value: you might want to revisit your code from question 1.6)*

4. Comment on the bias and variance of these 3 estimators.

5. Is it true that "more data is always better"? Discuss the statement based on your results. *(1 - 2 paragraphs)*

6. Can you think of strategies/good practices that might help us avoid this type of issues?