ECMA 31330 - Problem Set 2

Bruno Aravena M.

Due date: January 23rd 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.

This means you will submit a .pdf where you answer the questions and include your final figures, and a script of code (e.g. .R) where I can see your work.

This problem set will be focused on subset selection. I would recommend watching Professor Bonhomme's lecture from last Tuesday if you missed it. Section 6.1 of the James et al. textbook might also prove useful.

1 Subset selection and simulations

In this question you will run a simulation to study a subset selection problem with s = 1. Consider the following model, unknown to the researcher, which we will take as given.

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + u_i \tag{1}$$

Assume that $\beta_1 = 4$; $\beta_2 = \beta_3 = \beta_4 = 2$. Note that this is a departure from the lecture case where we only had one $\beta_k \neq 0$, which illustrated the Sparsity condition. *Will this departure suppose a problem for the method?*

1. Let N = 500. Let x_k be standard normals of size/length N independent of each other. u (also of size/length N) will be drawn from a normal of mean 0 and a standard deviation of 2.

   Run 500 simulations in which you use subset selection (with s=1) to try to identify the *"best regression"* (given s=1, this is equivalent to identifying which regressor j induces the largest R^2). For each simulation you should store a $\hat{j}$ and its associated coefficient $\hat{\beta}_j$.

2. During lecture we saw a trick to simplify the comparison of R^2 across regressions using the β's instead. Can we use that trick in this setting ? Why / Why not ?

3. In how many (what fraction) of these regressions is the researcher able to correctly identify the most relevant regressor? Should this surprise us considering our departure from the class version of the sparsity condition?

4. Repeat this exercise in question 1. but now using the following variance/covariance matrix between the x's:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \gamma & \gamma \\ 0 & \gamma & 1 & \gamma \\ 0 & \gamma & \gamma & 1 \end{bmatrix}$$

   Note that $\gamma \in [-1, 1]$ and that γ was 0 in the previous problem, **we will now set it to 1/2.** Importantly, this is also a departure from the class example. Again: For each simulation, you should store a $\hat{j}$ and its associated coefficient $\hat{\beta}_j$.

5. In how many (what fraction) of these regressions is the researcher able to correctly identify the most relevant regressor? Should this surprise us considering our departures from the class assumptions?

6. *(Optional)* Can you relate this result to the Omitted Variable Bias?

7. Suppose that we want our subset selection method to choose the correct regressor at least 95% of the time. A researcher argues that won't occur unless the $x's$ are perfectly uncorrelated ($\gamma = 0$). Do you agree? Are there values of $\gamma > 0$ such that subset selection still *"works"* (chooses the correct regressor at least 95% of the time) ? Try at least 3 different values of γ to get an idea and report on your findings. (you don't need to run 500 simulations on each γ, but certainly having a few would help you get more precision on what you find).

8. Another researcher argues that the problem is not necessarily the correlation structure but the lack of sparsity. They say:

   "Suppose a world with the same variance/covariance matrix and $\gamma = 1/2$ but where there is more sparsity: where the distance between β_1 and the other β_k is higher. If that distance is large enough subset selection would still work."

   Do you agree? For simplicity keep fixed $\beta_2 = \beta_3 = \beta_4 = 2$. Are there values of β_1 such that subset selection would still *"work"* (choose the correct regressor at least 95% of the time) ? Try at least 3 different values of β_1 to get an idea and report on your findings. (you don't need to run 500 simulations on each β_1, but certainly having a few would help you get more precision on what you find).

2 Subset selection and data: growth regressions

For this question we will be working with data from Sala-i-Martin's paper I Just Ran Two Million Regressions" (AER, May 1997) In particular, you might want to look at the end of the paper where the author describes the data. I also found this resource that might complement the data description.

DATA: You can find the data here

1. Implement subset selection with s = 1. Given that there are 62 regressosr you will need 62 regressions. You might want to think how to avoid doing this manually. Report on your result: what variable is chosen ?

2. During lecture we saw a trick to simplify the comparison of R^2 across regressions using the β's instead. Can we use that trick in this setting ? Why / Why not ?

3. If we wanted to implement subset selection with s = 2 naively we would need to run 62*61/2 = 1891 regressions. While still far from Sala-i-Martin's impressive number that would be a bit of a hassle. Can you propose a way such that you only need to run 61 additional regressions to implement subset selection with s = 2 ? *Hint: Look at the book section referenced at the beginning.*

4. Implement your strategy above and report on your selection.

5. What can you say about the 2 variables that were selected: Does the selection make intuitive sense ? Would you claim those are the two variables that "matter most" (largest β in true model) for growth ? What assumptions would you need to make that claim ?

6. If you make those assumptions: Can you make causal claims about the β you estimate?