

Problem Set 1

Statistics and Econometrics

Deadline: 11am, 4 November 2020

General Guideline

What we are looking for in the assignments is a demonstration that you can understand the econometrics and statistics questions and can solve them with R or conceptually. That means effective programming to get correct results is needed, but at the same time, clear explanations of economics/business concepts in well presented reports are equally important when assessing your work. In particular, you will be marked for successful (correct) programming (not the style of coding), good understanding of related concepts, and clear interpretations and explanations of results.

Please submit a pdf or html file converted from R markdown/notebook after you program in R.

Question 1

Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved performance of students in Year 6. (fyi, “in schools in England Year 6 is the sixth year after Reception. It is the sixth full year of compulsory education, with children being admitted who are aged 10 before 1 September in any given academic year. It is also the final year of Key Stage 2 in which the National Curriculum is taught.” source: Wikipedia)

1. Suppose you can collect observational data on several thousands of Year-6 students in England. You can obtain the size of their class and a standardized test score taken at the end of Year 6. Would you expect any correlation between class size and test score?
2. Suppose you find a negative correlation between class size and test score. Does it necessarily indicate that smaller class sizes cause better performance? Explain.

Question 2

The data in `fertil2.RData` were collected on women living in the Republic of Botswana in 1988. The variable *children* refers to the number of living children. The variable *electric* is a binary indicator equal to one if the woman’s home has electricity, and zero if not. The variable *heduc* refers to the husband’s years of education.

1. Find the smallest and largest values of *children* in the sample. What is the average of *children*?
2. Find the average of husband’s years of education in the sample. How many observations are used to compute this average? Explain.
3. Create a graph to examine the relationship between *children* and *heduc*. Comment.
4. What percentage of women have electricity in the home?
5. Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find.
6. Estimate the simple regression model

$$children = \beta_0 + \beta_1 electric + u,$$

and report your results.

7. Does this simple regression necessarily capture a causal relationship between the number of children and the presence of electricity in the home? Explain.