MEM5220 - Microeconometrics Self-Evaluation 1

Taltech - DEF

YOUR NAME HERE

24 March, 2021

Preface

This first R econometrics self-evaluation assignment is focused on data cleaning, data manipulation, plotting and estimating and interpreting simple linear regression models.
Some packages I have used to solve the exercises:

```
library(tidyverse)
library(modelsummary)
library(broom)
library(lmtest)
library(sandwich)
library(car)
```

You can use **any** additional packages for answering the questions.

Note:

- This assignment has to be solved in this R Markdown document and you should be able to "knit" the document without errors.
- Fill out your name in "yaml" - block on top of this document
- Use the R markdown syntax:
  - Write your code in code chunks
  - Write your explanations including the equations in markdown syntax
- If you have an error in your code use `#` to comment the line out where the error occurs but do not delete the code itself. I want to see your coding errors so I can give feedback!

For more information on using R Markdown for class exercises see https://ntaback.github.io/UofT_STA130/Rmarkdownforclassreports.html

---

You will be working with the dataset `Caschool` from the **Ecdat** package, the dataset `Wage1` from the **wooldrige** package. In the last two exercises, you will be working with a simulated data.

Try to answer by questions by including a code chunk and then a written answer. The answer to question 1.1 serves as a template. Please proceed with the rest of the questions in a similar way.

You can use any plotting package but the figures should have a research project style quality (eg. axis labels, figure legend, figure title and if necessary figures notes).

---

1 Caschool exercises

1.1 Question:

Load the dataset `Caschool` from the **Ecdat** package.

The Caschooldataset contains the average test scores of 420 elementary schools in California along with some additional information.

```
# install.packages("Ecdat")
library("Ecdat")
data("Caschool", package = "Ecdat")
```

1.2 Question:

What are the dimensions of the `Caschool` dataset?

A:

```
dim(Caschool)
```

```
[1] 420  17
```

The dataset has 420 rows and 17 columns.

1.3 Question:

Does the `Caschool` dataset contain missing observations?

A:

1.4 Question:

Display the structure of the `Caschool` dataset. Which variable are encoded as factors?

A:

1.5 Question:

Provide a summary statistic of the data.

A:

1.6 Question:

What are the names of the variables in the dataset?

A:

1.7 Question:

How many unique observations are available in the variable "county"

A:

1.8 Question:

Summarize the mean number of students grouped by county.

A:

1.9 Question:

Calculate the log of average income from of the Caschool dataset. Call the variable **logavginc** and add this variable to the dataset. Then, plot a histogram of the average income vs. a histogram of log average income. What do you observe?

A:

1.10 Question:

We want to create now a subset of counties that have the ten highest district average income and that have the ten lowest district average income. Call this subset *Caschool_lowhighincome*.

Hint: One way is the create two subsets (eg. Cascholl_highincome and Caschool_lowincome and the use the `rbind()` function to bind them together.).

A:

1.11 Question:

Let us test wether a high student/teacher ratio will be associated with higher-than-average test scores for the school? Create a scatter plot for the full dataset (*Caschool*) for the variables **testscr** and **str**.

A:

1.12 Question:

Suppose a policymaker is interested in the following linear model:

$$testscr = \beta_0 + \beta_1 str + u \tag{1}$$

Where *testscr* is the average test score for a given school and *str* is the Student/Teacher Ratio (i.e. the average number of students per teacher).

Estimate the specified linear model. Is the estimated relationship between a school's Student/Teacher Ratio and its average test results postitive or negative?

A:

1.13 Question:

Now, plot the regression line for the model we have just estimated.

A:

1.14 Question:

Let us extend our example of student test scores by adding families' average income to our previous model:

$$testscr = \beta_0 + \beta_1 str + \beta_2 avginc + u \tag{2}$$

A:

1.15 Question:

Assume know that "str" depends also on the value of yet another regressor, "avginc". Estimate the following model. Compare the sign of the estimate of β_2 and β_3. Interpret the results.

$$testscr = \beta_0 + \beta_1 str + \beta_2 avginc + \beta_3 (str \times avginc) + u \tag{3}$$

A:

1.16 Question:

You have fitted 3 specifications for the Caschool example. Report the regression results of equation 1, 2 and 3, in a formatted table regression output table. Discuss the model fit and model selection.

A:

2 Wage1 excercises

Wage data: These are data from the 1976 Current Population Survey. Source of the data is Wooldrige. Familiarize yourself with the dataset if necessary.

```
# install.packages("wooldridge")
library("wooldridge")
data("wage1", package = "wooldridge")
```

2.1 Question

First, estimate the following model and test again for heteroscedasticity.

$$wage = \beta_0 + \beta_1 female + \beta_3 educ + \beta_4 exper + u \tag{4}$$

A:

2.2 Question:

Test for heteroscedasticity test in the estimated regression of the wage1 dataset. Do we reject homoscedasticity for all reasonable significance levels? Adjust for heteroscedasticity by using refined White heteroscedasticity-robust SE.

A:

2.3 Question:

Now, estimate the following model:

$$log(wage) = \beta_0 + \beta_1 (married \times female) + \beta_3 educ + \beta_4 exper + beta_5 exper^2 + \beta_6 tenure + \beta_7 tenure^2 + u \tag{5}$$

1. What is the reference group in this model?
2. Ceteris paribus, how much more wage do single males make relative to the reference group?
3. Ceteris paribus, how much more wage do single females make relative to the reference group?
4. Ceteris paribus, how much less do married females make than single females?
5. Do the results make sense economically. What socio-economic factors could explain the results?

A:

2.4 Question

Create a regression table showing the results from equation 4 and 5. Show a specification where the SE have not been adjusted for heteroscedasticity and another specification where the SE have been adjusted for heteroscedasticity.

A:

3 Collinearity exercises

This exercise focuses on the **collineartiy** problem.

3.1 Question:

Run the following commands in R:

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100)/10
y <- 2 +2*x1 + 0.3 *x2 +rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

A:

3.2 Question:

What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

A:

3.3 Question:

Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

A:

3.4 Question:

Now fit least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

A:

3.5 Question:

Now fit least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

A:

3.6 Question:

Do the results from the previous questions contradict each other? Explain your answer.

A:

4 Simulation exercises

4.1 Question:

The probability that a baby is girl or boy is approximately 48.8% or 51.2%, respectively, and these do not much very much across the world. Suppose that 400 babies are born in a hospital in a given year. How many will be girls?

Set a seed (eg. `set.seed(123)`) to make the result reproducible!

A:

4.2 Question:

Simulate the process 1000 times and plot the distribution. Indicate the mean in the distribution plot.

A:

4.3 Question:

In the previous exercise we simulated a discrete probability model. Now, we will simulate a mixed discrete/ continuous model.

In the United States 52% of the adults are women and 48% are men. The heights of the men are approximately normally distributed wit mean 69.1 inches and standard deviation 2.9. Women have a mean height of 63.7 inches and a standard deviation of 2.7.

Generate the height of one randomly chosen adult (random adult means that this can either be a man or a women). Don't forget to set a seed. How tall is that person? What gender does that random person probably have?

A:

4.4 Question:

Now, simulate the distribution of the average height by generating 1000 draws. Plot the distribution of the average height of those 10 adults.

A:

4.5 Question:

Finally, instead of estimating the average height of 10 people, simulate the same model and extract the maximum height of 10 people. Plot the distribution.

A: