MEM5220 - R Econometrics Evaluation

YOUR NAME HERE

This first R econometrics self-evaluation assignment is focused on data cleaning, data manipulation, plotting and estimating and interpreting simple linear regression models. You can use **any** additional packages for answering the questions.

Packages I have used to solve the exercises:

```
library(tidyverse)
library(stargazer)
library(huxtable)
library(broom)
library(lmtest)
library(sandwich)
library(car)
```

Note:

- This assignment has to be solved in this R markdown 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 you feedback!

The answer to the first question is already provided. Please proceed with the rest of the questions in a similar way.

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

Caschool exercises

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")
```

What are the dimensions of the `Caschool` dataset?

A:

```
dim(Caschool)
```

```
[1] 420  17
```

Question:

Display the structure of the Caschool dataset. Which variable has values encoded as factors?

A:

Question:

Provide a summary statistic of the data

A:

Question:

What are the names of the variables in the dataset?

A:

Question:

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

A:

Question:

Summarize the mean number of students grouped by county.

A:

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:

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:

Question:

Let us test whether 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**. You can plot either in base R or use ggplot2.

A:

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 positive or negative?

A:

Question:

Now, plot the regression line for the model we have just estimated. Again, you can use either base R or ggplot2-style.

A:

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:

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:

Question:

In question 10, 12 and 13, you have fitted 3 models. Report the regression results, the number of observations, the Akaike information criterion and the model fit (adj. R^2) in formatted table regression output table. You can use for example the **stargazer** or the **huxtable** package. Which model fits the data best?

A:

Wage1 excercises

Wooldridge Source: These are data from the 1976 Current Population Survey.

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

Question:

Estimate the following model:

$$log(wage) = \beta_0 + \beta_1(married \times female) + \beta_3 educ + \beta_4 exper + beta_5 exper^2 + \beta_6 tenure + \beta_7 tenure^2 + u \tag{4}$$

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:

Question:

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

A:

Question

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

$$wage = \beta_0 + \beta_1 female + \beta_3 educ + \beta_4 exper + u \tag{5}$$

Adjust for heteroscedasticity if necessary.

A:

Collinearity exercises

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

Question:

Perform 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:

Question:

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

A:

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:

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:

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:

Question:

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

A:

Question:

Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y = c(y,6)
```

Re-fit the linear model using the new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier?

A: