Econ 184b, Econometrics, Assignment 3

The raw .rmd file for this assignment are available at: https://raw.githubusercontent.com/f1kidd/Econ184/main/Assignment/Assignment_3.Rmd

Note: For the math problems, you can either (1) type your solution and compile it with RMarkdown; (2) type your solution use another word-processing software and convert it into a PDF file; (3) write your solution on paper, scan it and make it into a legible PDF file. TAs can decide, at their discretion, that the submitted file is illegible and thus give zero credit to a question or the entire problem set. If you type your solutions (options 1 and 2 above), you will get **a bonus equal to 5% of your performance** on the problem set. No credit will be given if you only report the final answers without showing intermediate steps whenever appropriate.

For the programming problems, you need to submit both the compiled pdf/html file and the RMarkdown (.rmd) file on LATTE. You will get **an additional bonus equal to 5% of your performance** if your RMarkdown file is completely reproducible with minimal altercation (install packages, etc.). That is, our team (or anyone else) should be able to recompile your Rmarkdown file and reach *the same* result. You need to explicitly write out your answers - just showing programming outputs will receive zero credit.

You will be receiving a total of 15% bonus if you submit one single pdf/html compiled directly from Rmarkdown, along with the .rmd file.

Question 1:

A survey of 328 registered voters is conducted, and the voters are asked to choose between candidate A and candidate B. Let p denote the fraction of voters in the population who prefer candidate A, and let $\hat{p}$ denote the fraction of voters in the sample who prefer candidate A.

You are interested in constructing a confidence interval for p. Someone suggested that you can use $[\hat{p}-0.1, \hat{p}+0.1]$ as a confidence interval.

a. What is the coverage probability of this confidence interval if p=0.6? (What is $p(\hat{p} - 0.1 \leq p \leq \hat{p} + 0.1)$ when p=0.6?)
b. If you observe that $\hat{p} = 0.5$, can you reject the hypothesis that $H_0 : p = 0.6$?

Question 2:

You want to study whether or not academic achievement is related among different subjects. For example, do students who perform well on reading also tend to perform well on math? To answer this question, you collected data on 420 school districts in California. The sample is divided into two groups:

- High-reading group is defined as school districts with reading score above 655. Let μ_1 denote the population mean of math scores in this group. In the sample, this group consists of 259 school districts. For this group, the sample average of the math score is 664.11 and the standard deviation of the math score is 14.28.
- Low-reading group is defined as school districts with reading score below 655. Let μ_2 denote the population mean of math scores in this group. In the sample, this group consists of 161 school districts. For this group, the sample average of the math score is 636.02 and the standard deviation of the math score is 10.13.

a. construct a 95% confidence intervals for $\mu_1 - \mu_2$

b. can we conclude that the two groups (defined by the reading score) have different math scores? Use 1% significance level. Please write out the null and alternative hypothesis, and use both the t-test and the p-value approach to test the hypothesis.

Question 3:

A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam, while others have 120 minutes. Each student is randomly assigned one of the examination times, based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i-th student. Let X_i measure the amount of time that the student has to complete the exam: $X_i = 1$ means "90 minutes" and $X_i = 0$ means "120 minutes". Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose that we estimated that $\hat{\beta}_1 = 10$. Consider the following statements.

(1) On average, students who get 90 minutes score 10 points lower than students who get 120 minutes.
(2) On average, students who get 90 minutes score 10 points higher than students who get 120 minutes.
Which one of the above statements is true? Explain your answer.

Question 4:

You want to study whether expenditure per student can predict the educational outcome. To answer this question, you decided to run the following regression:

$$\text{reading.score} = \beta_0 + \beta_1 * \text{expenditure} + u$$

In the dataset CASchools.csv (available from "https://raw.githubusercontent.com/f1kidd/Econ184/main/Data/CASchools.csv"), "expenditure per student" is measured by the "expenditure" variable and "reading.score" is measured by the "read" variable. Run the above regression in R and answer the following questions:

a. Compute β_0, β_1 and SE(β_1).
b. Construct a 95% confidence interval for β_1.
c. Should we conclude that expenditure per student can predict the reading score? Use 1% significance level. Please write out H_0 and H_1 and use the p-value approach.

Your code could look like this:

```
library(tidyverse)
ca_school_url = "https://raw.githubusercontent.com/f1kidd/Econ184/main/Data/CASchools.csv"
data = read_csv(ca_school_url)
# run regression
# generate regression diagnostics
# compute confidence interval
```

Question 5

Generate a set of new data from the model

$$test.score = \beta_0 + \beta_1 * \text{expenditure} + u$$

with the following parameter: $\beta_0 = 686$, $\beta_1 = 0.006$, and n=420.

- Generate "expenditure" from N(5312,401956). (So $\mu_{expenditure} = 5312$)
- Generate $u = 0.83 * \sqrt{|expenditure - \mu_{expenditure}|} * Z$, where Z~N(0,1) is independent from expenditure.

a. Is the model homoskedastic?
b. The homoskedastic standard error for β_1 is 0.0014359 (from SW equation 5.22). With $\beta_1 = 0.006$, what is the range that contains 95% of $\hat{\beta}_1$ given this standard error?

c. Use 5,000 simulations, calculate the probability that the actual $\hat{\beta}_1$ falling into the range calculated in part (b).
d. The hetetoskedasticity-robust standard error for β_1 is 0.0020346. what is the range that contains 95% of $\hat{\beta}_1$ given this standard error?
e. Use the same simulation above, calculate the probability that the actual $\hat{\beta}_1$ falling into the range calculated in part (d).

Your code could look like the following:

```
n=420
N=5000
beta0 = 686
beta1 = 0.006

for(i in 1:N){
# generate expenditure from the distribution

# generate u as a function of expenditure

# generate test score as a function of expenditure and u

# run regression

# calculate coverage probability
}
```