Homework 1

STUDENT NAME

Due 6/27 11:59pm

NOTE: If you would like to answer the written portions using RMarkdown, simply type your answers below the questions. If you would like to print out the pdf to do the written portions, you should add `\vspace{50mm}` to add space in the knitted pdf for written work (adjust the 50mm as necessary). Alternatively, you can do the written portions on a separate paper.

Part 1: Theory

1.

Suppose you have a sample of 10 draws $\{x_1, x_2, \ldots, x_10\}$ from the distribution of a random variable X and you would like to estimate its mean, μ_X. Your friend, who hates even numbers, suggests that you take the average of the all your draws with an odd i, meaning $\frac{1}{5}(x_1 + x_3 + x_5 + x_7 + x_9)$.

1.1

Is this an unbiased estimator for the mean? Show why or why not.

1.2

Name a specific reason why the sample average of the whole sample, $\bar{x}$, is a better estimator for the mean.

2.

You would like to estimate the model $y = \beta_0 + \beta_1 x + u$ using data you've collected $\{x_i, y_i\}$. Show a complete derivation (using any method) of how to calculate the Ordinary Least Squares estimators $\hat{\beta_0}$ and $\hat{\beta_1}$ using your data.

3.

Suppose you want to know how city mandatory recycling programs affect the amount of litter on the streets. You have collected data on which cities in California have mandatory recycling programs and how much litter was present on a random sample of their streets.

3.1

State the *ceteris paribus* thought experiment that underlies this question about recycling programs.

3.2

Suppose you find that, on average, cities with mandatory recycling programs have 2 pounds less of litter per road mile than cities without such a program. Is this convincing evidence that mandatory recycling programs reduce litter? Why or why not?

4

In a large university class, the conditional expectation of the final exam score (in percentage points) with respect to hours spent studying was $\mathbb{E}(Score|Hours) = 71 + 2 \cdot Hours$

4.1

The average hours a student spent studying in this class was 8 hours. What was the average final exam score?

4.2

A student named Bob studies for 8 hours. Does that mean his exam score is the answer you gave in the previous problem? Why or why not?

Part 2: Application

Preliminaries

Load the `tidyverse` package and the `world_bank_data.csv` file. (If you are doing this assignment on your own computer for the first time, you will need to install the package with `install.packages("tidyverse")`.)

This data comes from the World Bank Development Indicators for 2012. It includes the variables GDPCapita (GDP per capita measured in thousands of USD), RenewEnergyPct (the share of renewable sources in total energy consumption), and GGTonnesCapita (gigatons of CO2-equivalent greenhouse gas emissions per capita) for 123 countries.

1.

Using the `summarise()` function, generate custom summary statistics that tell you the mean, median, 5th percentile, 95th percentile, and number of observations for GGTonnesCapita.

2.

Plot the data with GDPCapita on the x axis and GGTonnesCapita on the y axis. Make sure your plot has a nice title and axis labels.

3.

Manually (i.e. not using the `lm()` or any other regression estimation function) calculate the Ordinary Least Squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the regression equation $GGTonnesCapita = \beta_0 + \beta_1 \cdot GDPCapita + u$. Then check your work by estimating the parameters with the `lm()` function.

4.

Now use the `lm()` function to estimate the regression equation $\log(GGTonnesCapita) = \beta_0 + \beta_1 \cdot \log(GDPCapita) + u$

5.

What do the values you estimated for $\hat{\beta}_1$ in the previous two problems *mean* in terms of how GDP per capita is associated with GHG emissions per capita? Interpret both.