Day 4 Pset

Cecilia Sui and all other TAs

Day 4 Outline

1. Descriptive Statistics
2. Introduction to `ggplot`
3. Simple linear regression

Descriptive Statistics and Simple linear regression

This section touches on how to run linear regression briefly. Don't worry if you do not understand the meaning of every item in your summary table for your model. We will learn more about how to interpret them in QPM I and II.

1. Install and load the **faraway** package. Load the **gavote** dataset. Study the dataset using the help() function. Create a new variable called undercount by calculating the percentage of ballots that were not counted into votes. What's the range and quantiles of this new variable? Draw a histogram to illustrate the distribution of this variable.
2. Create a new variable **perGore** for the percentage of votes for Gore. Use **plot(col1, col2)** to create a scatter plot for the columns: **perGore** and **perAA**. What do you see? How can you interpret the plot?
3. Let's run a linear regression with undercount as the response and perAA as the predictor. (Please refer back to the lecture notes for the example we did.) Summarize the regression results and describe what you see. Can you get the coefficients?
4. Run a linear regression with undercount as the response variable and both perGore and perAA as predictors. Summarize the results and describe what has changed. Can you guess why we observe such change?
5. Use one of the previous linear regression object. Run visual diagnosis using `plot()` function.

ggplot2 Basics

1. Find all the variables that negatively correlates with TFR from the **worldTFR** dataset. Can you offer a reasoning of the relationships?
2. Pick two variables from Q1 and use the package ggplot2 to plot them and TFR separately, where the two variables you choose should be on your x-axis, and TFR should be on the y-axis. Add appropriate labels. Compare the two plots and explain the differences.
3. Create a new dataframe named df1, such that the dataframe has two columns: Year and avg_TFR. The column Year should list the years with no duplicates, and the column TFR should store the average TFR of all countries during that specific year. Use the package ggplot2 to visualize the relationship between Year and avg_TFR. Describe what you see.

Linear regression

1. Find a dataset you think interesting. Load and do basic cleanings (dealing with NA values, check the data types, etc.).
2. Run a single linear regression, and report the results in Tables and Figures.