POLI706: Advanced Methods of Political Analysis

Problem set 3

Exercise 1

Load `data_assignment3.rdata` into your RStudio. Explain whether the following data sets are cross-sectional, time-series, pooled cross-sectional, or panel data. Using `glimpse()` or `str()` to explore the data sets.

a. `data1`

b. `data2`

c. `data3`

Exercise 2

Compare and contrast observational studies and randomized experiments. Discuss the strengths and weaknesses of each in terms of causal inference. Recap the second and third weeks.

Exercise 3

```r
set.seed(1224)  # Set seed for reproducibility

# Simulating observational data (non-random)
n <- 100  # Number of participants
income <- rnorm(n, mean = 50000, sd = 10000)  # Generating random income
age <- rnorm(n, mean = 40, sd = 10)  # Generating random age
group <- ifelse(income > 50000, 1, 0)  # Assign group based on income
 ↪  (observational)

# Outcome depends on both group and income
outcome <- 5 + 0.3 * income + 2 * group + rnorm(n)

# Creating the data frame
obs_data <- data.frame(group, income, age, outcome)

# Simulating randomized data (random assignment)
group_random <- sample(0:1, n, replace = TRUE)  # Randomly assigning to two
 ↪  groups
```

```r
# Outcome still depends on income and group, but group is randomized
outcome_random <- 5 + 0.3 * income + 2 * group_random + rnorm(n)

# Creating the randomized data frame
rand_data <- data.frame(group_random, income, age, outcome_random)
```

a. Run a linear regression model with `group` and `income` on `outcome` for the observational data, and with `group_random` and `income` on `outcome_random` for the randomized data, respectively.

b. Compare how well the group variable predicts the outcome in both datasets and explain what are the differences between two linear regression models using `obs_data` and `rand_data`.

c. Visualize the differences between observational and randomized data using boxplots.