Homework 1: R / Stata Practice

Prof. Tzu-Ting Yang

March 26, 2026

Instruction

- You should submit Homework 1 before 4/22.
- There is no late submission (zero points after the deadline).
- Homework 1 accounts for 15 points in your final grade.
- This homework will help you practice basic data skills and start working on your term paper.
- You may use **either Stata or R**. Submit **two files**:
  1. **Empirical analysis record (HTML)**: A single HTML file documenting your entire empirical workflow from loading the data through the regression analysis.
     - **Stata**: Write code and narrative in a `.stmd` file and compile with `markstat using "filename", bundle`.
     - **R**: Write code and narrative in an `.Rmd` file and knit to HTML via "Knit to HTML" in RStudio or `rmarkdown::render()`.
  2. **Answer sheet (PDF)**: A PDF answering all questions (1–12), including tables and figures.
- Recommended R packages: `haven`, `dplyr`, `ggplot2`, `fixest`, `rmarkdown`.
- Upload files here: https://www.dropbox.com/request/vE8QChJE7x3Ww9kQPuJ8
- File name format: `StudentID_YourName_HW1.html` and `StudentID_YourName_HW1.pdf`.

Empirical Analysis Record

Submit an HTML file recording your complete empirical workflow — from loading the data to the analyses in Questions. The HTML should include:

- All code you executed (data loading, cleaning, visualization, regression).
- Output from each code block (tables, figures, regression results).
- Brief narrative text explaining what each step does.

Hint (Stata): Create a `.stmd` file with markdown text and Stata code blocks, then compile:

```
markstat using "your_file", bundle
```

The `bundle` option embeds all figures directly into the HTML.

Hint (R): Create an `.Rmd` file with YAML header `output: html_document` and embed R code in ```` ```{r} ```` chunks. Render with:

```
rmarkdown::render("your_file.Rmd")
```

Question 1 Research Question

Describe the research question you plan to investigate in your term paper. Clearly state:

(a) What is the **research question**? Why this question is interesting?

(b) What is your **treatment variable**?

(c) What is your **outcome variable**?

Question 2 Data Introduction

Describe the **data source**, the **unit of observation** (e.g., individual, household, firm, region), the **type of data** (e.g., cross-sectional, panel, repeated cross-section), and the time period and geographic coverage if applicable.

Question 3 Examine Data

(a) Check whether your **outcome variable** has any missing values. Report the number of missing observations (if any) and discuss whether they could affect your analysis.

(b) Identify the variable(s) serving as the **unique ID** for each observation. Check whether this ID contains duplicate values and explain your findings.

Hint (Stata):

- Missing values: `misstable summarize outcomevar`, `inspect outcomevar`
- Duplicate IDs: `isid idvar` or `duplicates report idvar`
  For panel data: `xtset id time` or `isid id time`

Hint (R):

- Missing values: `sum(is.na(df$outcomevar))`
- Duplicate IDs: `sum(duplicated(df$idvar))`
  For panel data: `anyDuplicated(df[, c("id", "time")])`

Question 4 Sample Construction

(a) Describe the steps you took to construct your **analysis sample**. For each step, state the selection criterion (e.g., age restriction, non-missing outcome) and report the number of observations remaining after applying it.

(b) Describe the final sample size and sample period

Hint: Useful commands include:

- **Create variables** — Stata: `generate`, `egen` R: `mutate()`
- **Recode values** — Stata: `recode`, `replace` R: `case_when()`, `ifelse()`
- **Keep/drop observations** — Stata: `keep`, `drop` R: `filter()`, `select()`
- **Merge/append** — Stata: `merge`, `append` R: `left_join()`, `bind_rows()`
- **Reshape** — Stata: `reshape` R: `pivot_longer()`, `pivot_wider()`
- **Collapse** — Stata: `collapse` R: `group_by()` + `summarise()`

Question 5 Distribution of a Variable

Create a graph displaying the **distribution** of a variable you are interested in. Specify this graph as Figure 1 in your answer sheet. Briefly explain your findings.

Hint (Stata): `histogram varname, xtitle("...") ytitle("...")`
Hint (R): `ggplot(df, aes(x = var)) + geom_histogram() + labs(x = "...", y = "...")`

Question 6 Relationship Between Two Variables

Create a graph displaying the **relationship between two variables** of interest (e.g., your treatment and outcome). Specify this graph as Figure 2 in your answer sheet. Briefly explain your findings.

Hint (Stata): `graph twoway (scatter y x) (lfit y x)`
Hint (R): `ggplot(df, aes(x = x, y = y)) + geom_point() + geom_smooth(method = "lm")`

Question 7 Simple Regression

Use regression to investigate the relationship between your treatment and outcome variable **without controlling for any covariates**. Write down the regression equation, define the variables, and explain your finding.

Hint (Stata): `reghdfe y treat`
Hint (R): `lm(y ~treat, data = df)`

Question 8 Causal Interpretation and Omitted Variable Bias

Do you think the estimate from Question 7 represents a **causal effect**? Is there potential **omitted variable bias (selection bias)**? Provide a concrete example of an omitted variable that affects both the treatment and the outcome, and explain the direction of the bias.

Question 9 OVB Formula

Use the **omitted-variable bias (OVB) formula** to discuss how the omitted variable you identified in Question 8 biases your estimate. Does it lead to an **overestimation** or **underestimation** of the causal effect? Show the formula and explain each component.

Question 10 Adding Control Variables

List the **control variables** you plan to include. Add them to the regression and compare the results before and after (focusing on changes in the treatment coefficient). What do you find?

Hint (Stata): `reghdfe y treat controls`
Hint (R): `lm(y ~treat + controls, data = df)`

Question 11 Bad Controls

Explain why the control variables you added in Question 10 are **not bad controls** — that is, why they are not themselves caused by the treatment variable.

Question 12 Fixed Effects

Add **fixed effects** to your regression. Explain which fixed effects you include and why they help control for confounding factors. Compare results before and after adding fixed effects and discuss the impact on the treatment coefficient.

Hint (Stata): `reghdfe y treat controls, absorb(fe1 fe2)`
Hint (R): `fixest::feols(y ~treat + controls | fe1 + fe2, data = df)`