Problem Set 2: Applied Regression (I)

Applied Quantitative Methods for the Social Sciences II

Carlos III–Juan March Institute, Spring 2026

Instructions:

- **Deadline**: **February 19, before class**
- Submit your work as a `.R` file called `ps2.R` in your GitHub repository
- Use comments in your R code to answer conceptual questions and explain your analysis
- You are encouraged to work together, but each person must submit their own code

1 Conceptual Questions

Answer these questions using comments in your R script. You don't need to run any code for this section—just write your answers as comments.

1.1 Question 1: Conditional Expectations

Consider a study examining the effect of education on income.

a) Explain in your own words what a conditional expectation function (CEF) is. Write down the CEF for income given education.

b) Why is regression considered an approximation to the CEF? Under what conditions is the regression exactly equal to the CEF?

c) A researcher finds that, on average, people with a college degree earn $20,000 more than people with only a high school diploma. Is this a descriptive or causal statement? Explain.

1.2 Question 2: Omitted Variable Bias

a) State the omitted variable bias formula and explain each component.

b) In the education-income example, suppose "ability" is an omitted variable that affects both education and income. If ability is positively related to both education and income, what is the sign of the omitted variable bias? Is the effect of education on income over- or under-estimated?

c) A researcher argues: "I cannot measure ability, but I can control for test scores as a proxy." Discuss the limitations of this approach.

1.3 Question 3: Good and Bad Controls

For each of the following scenarios, identify whether the proposed control variable is a good control, a bad control (post-treatment), or a collider. Explain your reasoning.

a) **Research question:** Effect of job training on wages. **Proposed control:** Current occupation.

b) **Research question:** Effect of smoking on lung cancer. **Proposed control:** Family history of cancer.

c) **Research question:** Effect of education on income. **Proposed control:** Being employed (yes/no).

d) **Research question:** Effect of democracy on economic growth. **Proposed control:** Colonial history.

2 Applied Analysis: European Social Survey

For this problem set, you will use data from the European Social Survey (ESS). The ESS is a cross-national survey that collects data on attitudes, beliefs, and behavior patterns of diverse populations in Europe.

You can download the data from: https://www.europeansocialsurvey.org/data/

Alternatively, use the `essurvey` package in R:

```
install.packages("essurvey")
library(essurvey)
set_email("your@email.com")  # Register at ESS website first

# Download ESS Round 10 (2020-2022)
ess <- import_rounds(10)
```

We will examine the determinants of support for redistribution. Key variables include:

- `gincdif`: Government should reduce income differences (1–5 scale, 5 = strongly agree)
- `hinctnta`: Household income decile (1 = lowest, 10 = highest)
- `eduyrs`: Years of education
- `agea`: Age in years
- `gndr`: Gender (1 = male, 2 = female)
- `cntry`: Country

2.1 Question 4: Data Exploration

a) Select a subset of countries (at least 3) and prepare the data for analysis. Remove missing values and recode variables as needed. Report sample sizes by country.

b) Create a histogram of the redistribution variable (`gincdif`). Describe the distribution of preferences in a comment.

c) Create a scatter plot of income (x-axis) versus support for redistribution (y-axis), using jittering to show the distribution. Describe the relationship you observe.

2.2 Question 5: Bivariate Regression

a) Estimate a bivariate regression with support for redistribution as the outcome and household income decile as the predictor. Print a summary of the results.

b) Interpret the coefficient on income in a comment. What is the predicted difference in redistribution support between someone in the lowest income decile and someone in the highest?

c) Calculate and print the 95% confidence interval for the income coefficient. Is the effect statistically significant? Is it practically significant?

2.3 Question 6: Multiple Regression

a) Estimate a model that includes income, education (years), age, and gender. Print a summary of the results.

b) Compare the coefficient on income in this model to the bivariate model. Does it change? In what direction? Explain what this suggests about the role of the control variables.

c) Interpret each of the coefficients in the multiple regression model (in comments).

2.4 Question 7: Country Differences

a) Add country fixed effects to your model (include country as a factor variable). How do the results change?

b) Calculate the predicted support for redistribution for a 40-year-old woman with median income and 14 years of education in each of the countries you selected.

c) Discuss in a comment: Why might there be country-level differences in support for redistribution even after controlling for individual characteristics?

2.5 Question 8: Thinking Causally

Answer these questions in comments:

a) Can we interpret the coefficient on income as the causal effect of income on redistribution preferences? What would we need to assume?

b) Design a hypothetical experiment that would allow you to estimate the causal effect of income on redistribution preferences. What are the practical and ethical challenges?

c) List at least three potential confounders that might bias the relationship between income and redistribution preferences in the ESS data.

3 Submission

Commit your `ps2.R` file to your GitHub repository before the deadline. Make sure your repository is public so I can access it.

Your R script should:

- Be well-organized with clear section headers (using comments)
- Include all code needed to reproduce your analysis
- Include your answers to conceptual questions as comments
- Run without errors from top to bottom