Assignment 2: Applied Regression

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.

1.4 Question 4: Interaction Effects

Consider the model: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + \varepsilon$

a) What is the marginal effect of X on Y? Show how it depends on Z.

b) Suppose $\beta_1 = 0.5$, $\beta_3 = -0.1$, and Z ranges from 0 to 10. At what value of Z does the effect of X become zero?

c) Why is it incorrect to interpret β_1 as "the effect of X" in this model?

2 Applied Analysis: European Social Survey

For this assignment, 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 5: Data Exploration and Bivariate Regression

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 scatter plot of income (x-axis) versus support for redistribution (y-axis), using jittering to show the distribution.

c) Estimate a bivariate regression with support for redistribution as the outcome and household income decile as the predictor. Print a summary of the results using `broom::tidy()`.

d) 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?

2.2 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).

2.3 Question 7: Interactions

a) Estimate a model that interacts income with gender (e.g., `gincdif ~ hinctnta * gndr + eduyrs + agea`). Print the results.

b) What is the marginal effect of income on redistribution preferences for men? For women?

c) Use `marginaleffects::plot_predictions()` to visualize how the relationship between income and redistribution support differs by gender. Save the plot.

d) Discuss your findings in a comment: does the income-redistribution relationship differ between men and women? Is the difference substantively meaningful?

2.4 Question 8: Presenting Results

a) Using `modelsummary()`, create a table that presents all three models side by side: bivariate, multiple regression, and interaction model.

b) Create a coefficient plot using `modelsummary::modelplot()` comparing the three models.

c) In a comment, describe how the income coefficient changes across the three models and what this tells us about the relationship between income and redistribution preferences.

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
- Save any plots to files (e.g., using `ggsave()`)
- Run without errors from top to bottom