Assignment 3: Binary Outcomes

Applied Quantitative Methods for the Social Sciences II

Carlos III–Juan March Institute, Spring 2026

Instructions:

- **Deadline**: **February 26, before class**
- Submit your work as a `.R` file called `ps3.R` in your GitHub repository
- Use comments in your R code to answer conceptual questions and explain your analysis
- For model interpretation, use the `marginaleffects` package
- 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.

1.1 Question 1: The Linear Probability Model

a) What is the linear probability model (LPM)? How does the interpretation of the coefficient β_1 differ from standard OLS with a continuous outcome?

b) Describe two problems with the LPM. For each, explain why it is a problem and when it matters most.

c) Under what circumstances might the LPM be "good enough" despite its limitations?

1.2 Question 2: Logistic Regression

a) Write down the logit model. What is being modeled as a linear function of X?

b) Explain in your own words why we cannot interpret logit coefficients as changes in probability.

c) If a logit model estimates $\beta_1 = 0.8$, what is the odds ratio? Interpret it in words.

1.3 Question 3: Marginal Effects

a) What is an average marginal effect (AME)? How does it differ from a marginal effect at representative values?

b) Why are AMEs often similar to LPM coefficients? When might they diverge?

c) A researcher reports only the log-odds coefficients from a logit model. Why is this insufficient for understanding the substantive findings?

2 Applied Analysis: Voter Turnout

For this assignment, you will analyze voter turnout using data from the American National Election Study (ANES) or a similar dataset on political participation.

You can use the ANES data available through the `poliscidata` package, or download it from: https://electionstudies.org/data-center/

```
install.packages("poliscidata")
library(poliscidata)
data(nes)
```

Key variables for analysis (variable names may vary depending on data source):

- `voted`: Whether the respondent voted (binary: 0/1)
- `age`: Age in years
- `educ`: Education level (years or categories)
- `income`: Household income (categories or continuous)
- `pid`: Party identification strength
- `female`: Gender indicator

2.1 Question 4: Data Exploration

a) Load and prepare the data. Create a binary turnout variable if needed. Report the overall turnout rate and sample size.

b) Create a bar chart showing turnout rates by education level. Comment on the pattern.

c) Create a table showing turnout rates by age group (e.g., 18–29, 30–44, 45–64, 65+). Discuss the pattern in a comment.

2.2 Question 5: Linear Probability Model

a) Estimate an LPM with turnout as the outcome and age, education, income, and gender as predictors. Print a summary of the results.

b) Interpret the coefficient on education. What does it mean in terms of probability?

c) Check how many observations have predicted probabilities outside $[0, 1]$. Report the minimum and maximum predicted values.

2.3 Question 6: Logistic Regression

a) Estimate a logit model with the same predictors as the LPM. Print the results.

b) Report the odds ratios for all predictors using `exp(coef(model))`. Interpret the odds ratio for education.

c) Calculate the average marginal effects using `marginaleffects::avg_slopes()`. Compare these to the LPM coefficients—how similar are they?

2.4 Question 7: Predicted Probabilities and Visualization

a) Using the logit model, calculate the predicted probability of voting for:

- A 25-year-old woman with low education and low income
- A 55-year-old man with high education and high income

Report both point estimates and 95% confidence intervals.

b) Use `marginaleffects::plot_predictions()` to plot the predicted probability of voting across values of education, holding other variables at their means. Save the plot.

c) Create a similar plot showing predicted probabilities across age for different education levels (hint: use `condition = c("age", "educ")`). Save the plot and discuss the patterns in a comment.

2.5 Question 8: Model Comparison

a) Using `modelsummary()`, create a table comparing the LPM and logit models side by side.

b) Briefly discuss in a comment: For this dataset, do the LPM and logit models lead to different substantive conclusions? When might the differences matter?

3 Submission

Commit your `ps3.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