Applied Quantitative Methods for the Social Sciences II
MA in Social Sciences, UC3M, Spring 2026

Assignment 3: Binary Outcomes

Instructions:

- Deadline: **February 26, before class**
- Submit your work in a separate folder in your GitHub repository
  - You can include only the R file or additional ones (e.g. pdf with results)
- **Always use comments** in your R code – and use them to answer questions
- You are encouraged to work together, but each person must submit their own code
- Plan is to start Part 1 in class and complete Part 2 at home
- I'll upload a solution file to the website after next class

Contents

1 Part 1: In-Class (ANES Voter Turnout)

In this lab, we analyze voter turnout using data from the American National Election Study (ANES). You will practice estimating and comparing the linear probability model and logistic regression, computing marginal effects, and plotting predicted probabilities. Follow along in class.

1.1 Setup and data preparation

Download the data (and related files) here:

- github.com/franvillamil/AQM2/tree/master/datasets/anes

a) Load the dataset. Key variables to define:

   - `voted` — whether the respondent voted (0/1, our outcome)
   - `age` — age in years
   - `female` — gender indicator (1 = female, 0 = male)
   - `education` — years of education
   - `income` — household income (standardized scale)
   - `party_id` — strength of party identification (1–7 scale)

b) Drop observations with missing values on any of these variables. How many observations remain?

c) Compute the overall turnout rate (proportion of `voted == 1`) and print summary statistics for all variables.

1.2 Exploratory visualization

a) Create a bar chart showing the turnout rate by education level (hint: compute the mean of `voted` for each value of `education`, then use `geom_col()`).

b) In a comment, describe the pattern. Does turnout increase with education?

1.3 Linear probability model

a) Estimate an LPM with `voted` as the outcome and `age`, `education`, `income`, and `female` as predictors:
   `lpm = lm(voted ~ age + education + income + female, data = df)`.

b) Print the results using `broom::tidy()`.

c) Interpret the coefficient on `education` in a comment. What does it mean in terms of probability?

d) Check the predicted probabilities: how many are below 0 or above 1? Report the minimum and maximum predicted values.

1.4 Logistic regression

a) Estimate a logit model with the same predictors:
`logit = glm(voted ~ age + education + income + female,`
`family = binomial, data = df)`.

b) Print the results using `broom::tidy()`.

c) Compute the odds ratios using `exp(coef(logit))`. Interpret the odds ratio for `education` in a comment.

d) Verify that all predicted probabilities are between 0 and 1.

1.5 Comparing LPM and logit

a) Compute average marginal effects for the logit model using
`marginaleffects::avg_slopes(logit)`.

b) Compare the AMEs to the LPM coefficients. How similar are they? Discuss in a comment.

c) Create a table with `modelsummary()` showing the LPM and logit side by side. Use robust standard errors for the LPM:
`modelsummary(list("LPM" = lpm, "Logit" = logit), vcov = list("robust", NULL))`.

1.6 Predicted probabilities

a) Use `plot_predictions(logit, condition = "education")` to plot the predicted probability of voting across education levels. Save the plot.

b) Create a second plot showing predicted probabilities across `age` for men and women separately:
`plot_predictions(logit, condition = c("age", "female"))`.

c) In a comment, describe the patterns. How does the effect of age differ from the effect of education?

1.7 Presenting results

a) Create a coefficient plot comparing the LPM and logit models using `modelplot()`.

b) Save the plot.

c) In a comment: for this dataset, do the LPM and logit lead to different substantive conclusions? When might the differences matter?

2 Part 2: Take-Home Exercises (STAR — High School Graduation)

We return to the STAR experiment data from Assignment 2, but now focus on a **binary outcome**: whether students graduated from high school. This lets you practice binary outcome methods on a dataset you already know.

Use the same `star.csv` from the course page. Key variables for this assignment:

- `classtype`: class type (1 = small, 2 = regular, 3 = regular + aide)
- `race`: student race (1 = White, 2 = Black, 3 = Asian, 4 = Hispanic, 5 = Native American, 6 = Other)
- `yearssmall`: years spent in a small class (0–4)
- `hsgrad`: graduated from high school (0/1, our outcome)

2.1 Data preparation

a) Load `star.csv` and create the same factor variables as in Assignment 2: `classtype` with labels `"Small"`, `"Regular"`, `"Regular+Aide"`, and `race` with labels `"White"`, `"Black"`, etc.

b) Create a binary variable `small` equal to 1 if the student was in a small class and 0 otherwise.

c) Drop observations with missing values on `hsgrad`. How many observations remain?

d) Compute the high school graduation rate overall and by class type. In a comment, describe the differences.

2.2 LPM and logit

a) Estimate an LPM predicting `hsgrad` from `small`:
`lpm1 = lm(hsgrad ~ small, data = df)`.

b) Estimate a logit model with the same predictor:
`logit1 = glm(hsgrad ~ small, family = binomial, data = df)`.

c) Interpret the LPM coefficient on `small`: what is the estimated difference in graduation probability between small and non-small classes?

d) Compute the AME from the logit using `avg_slopes(logit1)`. How does it compare to the LPM coefficient?

2.3 Adding controls

a) Estimate both LPM and logit with controls:

```
lpm2 = lm(hsgrad ~ small + race + yearssmall, data = df)
logit2 = glm(hsgrad ~ small + race + yearssmall,
family = binomial, data = df).
```

b) Compare the coefficient on `small` between the bivariate and controlled models. Does it change much? What does this tell you about the randomization?

c) Interpret the coefficient on `yearssmall` from the logit model. Use `avg_slopes()` to convert to a marginal effect.

2.4 Predicted probabilities

a) Using the controlled logit model, compute predicted graduation probabilities for:

- A White student in a small class with 3 years in small classes
- A Black student in a regular class with 0 years in small classes

Use `predictions(logit2, newdata = datagrid(...))`. Report the estimates and 95% CIs.

b) Plot predicted graduation probabilities across `yearssmall` for small vs. non-small classes: `plot_predictions(logit2, condition = c("yearssmall", "small"))`. Save the plot.

2.5 Interactions

a) Does the small class effect on graduation differ by race? Estimate:
`logit3 = glm(hsgrad ~ small * race + yearssmall,`
`family = binomial, data = df)`.

b) Use `avg_slopes(logit3, variables = "small", by = "race")` to compute the marginal effect of `small` separately for each racial group.

c) In a comment, discuss: is the small class effect larger for some groups than others?

2.6 Presenting results and discussion

a) Create a table with `modelsummary()` comparing all four models (LPM bivariate, LPM controlled, logit bivariate, logit controlled). Use robust SEs for the LPM models.

b) Create a coefficient plot with `modelplot()`.

c) In a comment (5–10 sentences), discuss:

- What does the STAR data suggest about the effect of small class sizes on high school graduation?
- How do the LPM and logit results compare? Do they tell a similar or different story?
- Why is this experimental evidence more credible than an observational study?

3 Submission

Commit your file to your GitHub repository before the deadline. Put it in a different folder, e.g. `assignment3`. 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 and interpretations as comments
- Save any plots to files (e.g., using `ggsave()`)
- Run without errors from top to bottom