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

Assignment 5: Panel Data I

Instructions:

- Deadline: **March 19, 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 (Presidential Approval)

In this lab, we analyze the relationship between unemployment and presidential approval ratings using U.S. state-level panel data. The goal is to understand how fixed effects models differ from pooled OLS and what each estimator actually identifies. Recall from the lecture that panel data allow us to control for unobserved time-invariant differences across units — a key advantage over simple cross-sectional regressions.

Download the data here:

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

Load it with `readstata13::read.dta13("presidential_approval.dta")`. Key variables:

- `State` — state name
- `StCode` — state numeric ID
- `Year` — year
- `PresApprov` — percent positive presidential approval
- `UnemPct` — state unemployment rate
- `South` — southern state (1 = yes, 0 = no)

1.1 Setup and data exploration

a) Load the dataset. How many unique states and years are in the data? Use `length(unique())` or `n_distinct()` to check. Is the panel balanced (i.e., does every state appear the same number of times)?

b) Compute summary statistics for `PresApprov` and `UnemPct` using `summary()` or `modelsummary::dat` Then plot `PresApprov` over `Year` for a few selected states (e.g., California, Texas, New York) to visualize the panel structure:

```
library(dplyr)
library(ggplot2)

df_sub = df %>%
  filter(State %in% c("California", "Texas", "NewYork"))

ggplot(df_sub, aes(x = Year, y = PresApprov, color = State)) +
  geom_line() +
  theme_minimal() +
  labs(x = "Year", y = "Presidential approval (%)", color = "State")
```

In a comment, describe the trends. Do states move together over time?

c) Create a scatter plot of `PresApprov` (y-axis) against `UnemPct` (x-axis) across all state-year observations. Add a regression line with `geom_smooth(method = "lm")`. In a comment, describe the cross-sectional relationship: does higher unemployment seem to be associated with lower or higher approval ratings?

1.2 Pooled OLS

a) Estimate a pooled OLS model regressing presidential approval on unemployment: `m_pooled = lm(PresApprov ~ UnemPct, data = df)`. Report the results using `summary()` or `modelsummary()`. In a comment, interpret the coefficient on `UnemPct`: what does it say about the relationship between unemployment and approval?

b) Add `South` as a control: `m_pooled2 = lm(PresApprov ~ UnemPct + South, data = df)`. Does controlling for whether a state is in the South change the coefficient on `UnemPct`? In a comment, explain why or why not.

c) In a comment, reflect on the limitations of pooled OLS for this type of data. What kinds of unobserved, time-invariant differences across states might bias the estimate of the unemployment effect? Give two or three concrete examples.

1.3 Entity fixed effects

a) Estimate a model with state fixed effects using `fixest`:

```
library(fixest)
m_fe = feols(PresApprov ~ UnemPct | State, data = df)
```

Report the results alongside the pooled OLS model in a single `modelsummary()` table. How does the coefficient on `UnemPct` change compared to pooled OLS?

b) In a comment, explain what the state fixed effects are absorbing. Note that the `South` variable drops out of the model — why can't it be estimated when state fixed effects are included? What does this imply about any variable that does not vary within a state over time?

c) What does the coefficient on `UnemPct` now identify? In a comment, explain the intuition: the state FE estimator compares approval ratings *within* the same state across different years, rather than across different states. How does this differ from the pooled OLS interpretation?

1.4 Two-way fixed effects

a) Add year fixed effects to absorb common time shocks (e.g., national economic conditions, wars, presidential scandals) that affect all states simultaneously:

```
m_twfe = feols(PresApprov ~ UnemPct | State + Year, data = df)
```

b) Compare all three models in a single `modelsummary()` table with standard errors clustered by state:

```
modelsummary(
  list("Pooled OLS" = m_pooled, "State FE" = m_fe, "Two-Way FE" = m_twfe),
  vcov = ~State,
  stars = TRUE,
  gof_map = c("r.squared", "nobs"))

```

c) In a comment, discuss what the year fixed effects are controlling for. Does adding them change the coefficient on `UnemPct`? If so, what does that suggest about the role of common time trends in driving the relationship between unemployment and approval?

2 Part 2: Take-Home (Teaching Evaluations)

We now turn to a classic question in higher education research: does giving higher grades lead to better teaching evaluations? This is known as the "grade inflation" or "grading leniency" problem. Using instructor-level panel data, we will estimate whether instructors who give more A grades receive higher student evaluations — and whether fixed effects change our conclusions.

Download the data here:

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

Load it with `readstata13::read.dta13("teaching_evals.dta")`. Key variables:

- `Eval` — average course evaluation (5-point scale)
- `Apct` — percent of students receiving an A or A- in the course
- `Enrollment` — number of students enrolled
- `Required` — dummy variable: 1 if the course is required
- `InstrID` — unique identifier for each instructor
- `CourseID` — unique identifier for each course
- `Year` — academic year

2.1 Data exploration

a) How many unique instructors and courses are in the data? Use `n_distinct(df$InstrID)` and `n_distinct(df$CourseID)`. What is the average number of observations (course-year pairs) per instructor? In a comment, note whether this looks like a short or long panel.

b) Create a scatter plot of `Eval` (y-axis) against `Apct` (x-axis). Add a regression line with `geom_smooth(method = "lm")`. In a comment, describe the cross-sectional relationship between grading generosity and evaluations: is it positive or negative? Does this pattern surprise you?

2.2 Pooled OLS baseline

a) Estimate a pooled OLS model with all three regressors:
`m1 = lm(Eval ~ Apct + Enrollment + Required, data = df)`. Report the results using `summary()` or `modelsummary()`. In a comment, interpret the coefficient on `Apct`: a one-percentage-point increase in the share of A grades is associated with how much of a change in evaluation scores?

b) In a comment, explain why the OLS estimate of `Apct` might be biased. What unobserved characteristics of instructors could simultaneously drive both grading generosity and evaluation scores? Give at least two concrete examples. Is the expected bias upward or downward?

2.3 Fixed effects models

a) Estimate a model with instructor fixed effects, and a two-way model adding year fixed effects:

```
library(fixest)
m_instr = feols(Eval ~ Apct + Enrollment + Required | InstrID, data = df)
m_twfe  = feols(Eval ~ Apct + Enrollment + Required | InstrID + Year, data = df
   )
```

b) Compare all three models (m1, m_instr, m_twfe) in a single table with standard errors clustered by instructor:

```
modelsummary(
  list("Pooled OLS" = m1, "Instructor FE" = m_instr, "Two-Way FE" = m_twfe),
  vcov = ~InstrID,
  stars = TRUE,
  gof_map = c("r.squared", "nobs"))
```

c) Interpret the coefficient on Apct in the instructor-FE model (m_instr). In a comment, explain what the instructor fixed effect is controlling for. Is the FE coefficient on Apct larger or smaller than in the pooled OLS? What does this tell us about the direction of omitted variable bias in the pooled OLS estimate — are more lenient graders systematically better or worse evaluators in terms of their unobserved characteristics?

2.4 Random effects and the Hausman test

a) Estimate a random effects model using plm. The random effects model assumes that the unobserved instructor-level heterogeneity is uncorrelated with the regressors:

```
library(plm)
pdata = pdata.frame(df, index = c("InstrID", "CourseID"))
m_re  = plm(Eval ~ Apct + Enrollment + Required,
            data = pdata, model = "random")
```

b) Run the Hausman test to assess whether fixed or random effects is more appropriate. The Hausman test checks whether the random effects assumption (no correlation between unobservables and regressors) holds:

```
m_fe_plm = plm(Eval ~ Apct + Enrollment + Required,
               data = pdata, model = "within")
phtest(m_fe_plm, m_re)
```

c) In a comment, interpret the Hausman test result. What is the null hypothesis? Is it rejected? Based on the test and on the substantive reasoning from the previous sub-section, should you prefer the fixed effects or the random effects estimator for this dataset? Explain in 3–5 sentences.

3 Data Sources

Both datasets are available at the course GitHub repository:

- Presidential approval: github.com/franvillamil/AQM2/tree/master/datasets/presidential_appr
- Teaching evaluations: github.com/franvillamil/AQM2/tree/master/datasets/teaching_evals

4 Submission

Commit your file to your GitHub repository before the deadline. Put it in a separate folder, e.g. `assignment5`. 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