Assignment 5: Panel Data I – Part 2 (Teaching Evaluations)

Applied Quantitative Methods for the Social Sciences II

Spring 2026

```
library(readstata13)
library(ggplot2)
library(fixest)
library(plm)
library(modelsummary)

df = read.dta13("https://raw.githubusercontent.com/franvillamil/AQM2/refs/heads/master/datasets/teaching_e
```

1. Data exploration

a) Panel dimensions:

```
length(unique(df$InstrID))
```

```
## [1] 48
```

```
length(unique(df$CourseID))
```

```
## [1] 254
```

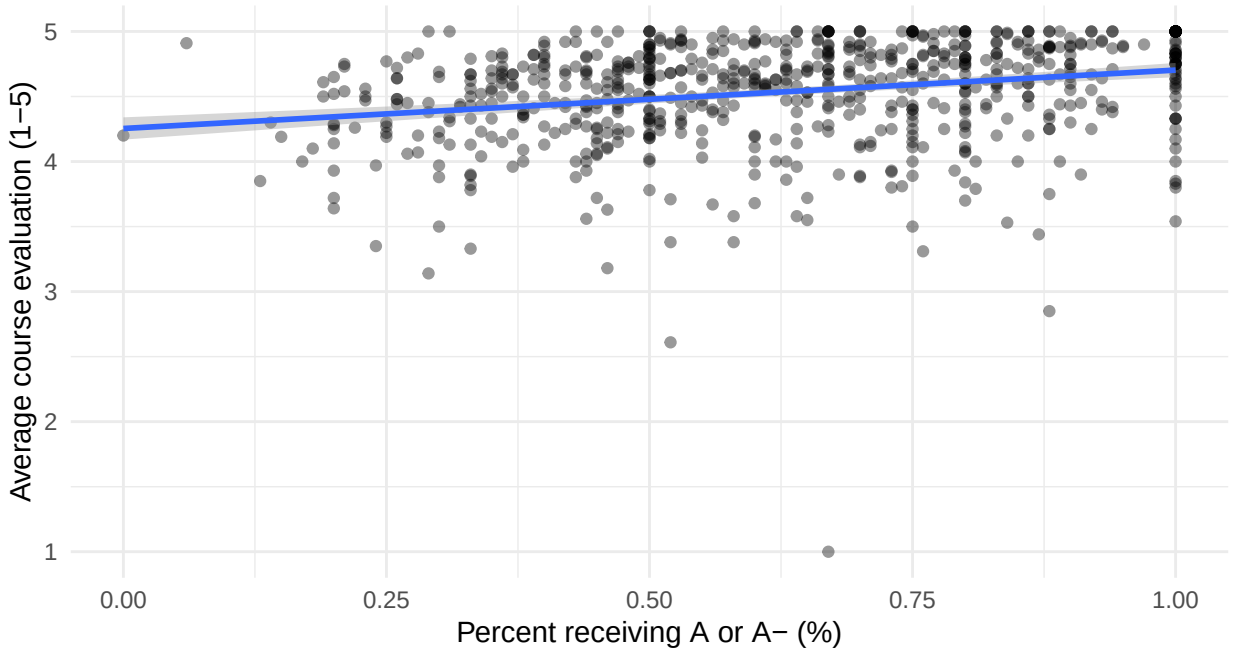
```
nrow(df) / length(unique(df$InstrID))
```

```
## [1] 17.52083
```

The panel has relatively few instructors observed over several courses and years. The average number of observations per instructor is substantially above 1, making this a moderately long panel at the instructor level. Because a single instructor may teach multiple courses in the same year, the appropriate panel index for `plm` uses both `InstrID` and `CourseID`, not `InstrID` and `Year`.

b) Scatter plot of evaluations against grading generosity:

```
ggplot(df, aes(x = Apct, y = Eval)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  theme_minimal() +
  labs(x = "Percent receiving A or A- (%)", y = "Average course evaluation (1-5)")
```



The cross-sectional relationship between `Apct` and `Eval` is positive: instructors who give more A grades tend to receive higher evaluations. This pattern is consistent with the grade-inflation hypothesis — that lenient grading buys better evaluations — but it could also reflect that both grading and evaluations are driven by unobserved instructor quality (e.g., more talented teachers give more deserved A grades and also teach better).

2. Pooled OLS baseline

a) Pooled OLS with all controls:

```
m1 = lm(Eval ~ Apct + Enrollment + Required, data = df)
summary(m1)
```

```
##
## Call:
## lm(formula = Eval ~ Apct + Enrollment + Required, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5782 -0.1919  0.0872  0.2780  0.6804
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  4.3382783  0.0548821  79.047 < 0.0000000000000002 ***
## Apct         0.3590967  0.0725780   4.948           0.00000094 ***
## Enrollment  -0.0002407  0.0005231  -0.460               0.6455
## Required    -0.1217797  0.0541220  -2.250               0.0248 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3858 on 706 degrees of freedom
```

	Pooled OLS	Instructor FE	Two-Way FE
(Intercept)	4.338***		
	(0.093)		
Apct	0.359**	0.306***	0.318***
	(0.126)	(0.062)	(0.064)
Enrollment	-0.000	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
Required	-0.122+	-0.148*	-0.151*
	(0.062)	(0.066)	(0.068)
R2	0.078	0.422	0.429
Num.Obs.	710	710	710

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
##   (131 observations deleted due to missingness)
## Multiple R-squared:  0.07752,    Adjusted R-squared:  0.0736
## F-statistic: 19.78 on 3 and 706 DF,  p-value: 0.000000000002561
```

The coefficient on `Apct` gives the association between a one-percentage-point increase in the share of students receiving an A grade and the change in average evaluation score, pooling across all instructors and years. Larger enrollment and required courses are often associated with lower evaluations, which is consistent with the course-level context: students are less satisfied in large required courses.

b) The OLS estimate of `Apct` is likely upward biased. Instructors who are genuinely excellent may both earn better evaluations and give grades that students feel are deserved — producing a spurious positive correlation between `Apct` and `Eval`. Similarly, instructors who are more charismatic or popular may give higher grades to maintain that reputation, inflating both `Apct` and `Eval`. In both cases, unobserved instructor characteristics (talent, charisma, conscientiousness) drive both variables simultaneously, leading us to overstate the causal effect of grading leniency on evaluations.

3. Fixed effects models

a–b) Instructor fixed effects and two-way fixed effects:

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

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

c) The instructor fixed effects control for all time-invariant differences across instructors: their baseline teaching quality, personality, subject area, grading philosophy, and any other stable characteristic. Comparing `Apct` coefficients across models reveals the direction of omitted variable bias in pooled OLS. If the FE coefficient on `Apct` is smaller (in absolute value or in magnitude of the positive effect) than the pooled OLS coefficient, it means that the

pooled estimate was inflated by unobserved instructor quality — i.e., better instructors give more A grades and also receive better evaluations for reasons unrelated to grading leniency. Within the same instructor over time, increases in the share of A grades are associated with a different (typically smaller) change in evaluations than what pooled OLS suggests.

4. Random effects and the Hausman test

a) Random effects model via `plm`:

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

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = Eval ~ Apct + Enrollment + Required, data = pdata,
##     model = "random")
##
## Unbalanced Panel: n = 46, T = 2-32, N = 710
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.09961 0.31561 0.659
## individual    0.05165 0.22726 0.341
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2993  0.6520  0.7100  0.6852  0.7461  0.7616
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.3428 -0.1273  0.0501  0.0001  0.1953  0.8229
##
## Coefficients:
##                Estimate  Std. Error z-value             Pr(>|z|)
## (Intercept)  4.38169956  0.06334753 69.1692 < 0.00000000000000022 ***
## Apct         0.32707573  0.07017082  4.6611          0.000003145 ***
## Enrollment  -0.00080130  0.00051917 -1.5434             0.122732
## Required    -0.14407018  0.05364144 -2.6858             0.007236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    170.74
## Residual Sum of Squares: 70.168
## R-Squared:      0.58904
## Adj. R-Squared: 0.5873
```

```
## Chisq: 95.0148 on 3 DF, p-value: < 0.000000000000000222
```

b) Hausman test:

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

```
##
##  Hausman Test
##
## data:  Eval ~ Apct + Enrollment + Required
## chisq = 4.9169, df = 3, p-value = 0.178
## alternative hypothesis: one model is inconsistent
```

c) The null hypothesis of the Hausman test is that the random effects estimator is consistent — equivalently, that the unobserved instructor-level heterogeneity is uncorrelated with the regressors (`Apct`, `Enrollment`, `Required`). If the test statistic is significant ($p < 0.05$), we reject this null and conclude that the RE assumption is violated: instructor-specific unobservables are correlated with the covariates, and the FE estimator is preferred because it remains consistent under this correlation. The substantive reasoning from the previous section already pointed in this direction: unobserved instructor quality plausibly drives both grading behavior and evaluation scores, violating the RE assumption. Whether or not the formal test is significant, the FE estimator is the more defensible choice here, as it controls for all time-invariant instructor characteristics and directly addresses the endogeneity concern. The RE estimator would only be appropriate if we were willing to assume that instructors' grading decisions are uncorrelated with their unobserved characteristics — an implausible assumption in this setting.