Gov 2001: Problem Set 5

Multiple Regression, OVB, and Inference

Spring 2026

Due: Friday, April 24, 2026, 11:59 PM Eastern

Submit: PDF to Canvas (we recommend R Markdown or Quarto)

Total: 100 points

Instructions:

- Include all R code and output for simulation problems.
- You may collaborate with classmates, but write your own solutions and list collaborators.
- **Do not use AI assistants (ChatGPT, Claude, Copilot, etc.) on this problem set.** Work with each other instead. The struggle is where learning happens.
- Remember: 70% of your grade comes from in-class exams. Use problem sets to *learn*, not just to get answers.

Topics: Multiple regression, omitted variable bias, interactions, robust standard errors, F-tests

Readings: Blackwell Ch. 6–7; Aronow & Miller §4.2; Angrist & Pischke §3.2

Question 1: Omitted Variable Bias (30 points)

This question explores one of the most important concepts in applied regression: what happens when we omit a relevant variable.

Setup

The true data generating process for wages is:

$$\text{Wage}_i = 10 + 2 \cdot \text{Educ}_i + 3 \cdot \text{Ability}_i + \varepsilon_i$$

where:

- Educ_i: years of education (observable)

- Ability_i: natural ability (unobservable)
- ε_i: random error, independent of everything

Suppose ability and education are positively correlated: $\text{Cov}(\text{Educ}, \text{Ability}) > 0$.

(a) (6 points) Write down the omitted variable bias formula. If we regress Wage on Education alone, what is the expected bias in our estimate of the education coefficient?

Specifically, show that:

$$\hat{\beta}^{\text{short}}_{\text{Educ}} = \beta_{\text{Educ}} + \beta_{\text{Ability}} \cdot \frac{\text{Cov}(\text{Educ}, \text{Ability})}{\text{Var}(\text{Educ})}$$

(b) (4 points) Given that ability has a positive effect on wages ($\beta_{\text{Ability}} = 3 > 0$) and ability is positively correlated with education, will the short regression coefficient be biased upward or downward? Explain the intuition.

(c) (8 points) **R Simulation:** Demonstrate OVB.

```r
set.seed(2001)
n <- 1000

# Generate data
Ability <- rnorm(n, mean = 0, sd = 1)
# Education is correlated with Ability
Educ <- 12 + 2*Ability + rnorm(n, mean = 0, sd = 2)
# Wage depends on both
epsilon <- rnorm(n, mean = 0, sd = 5)
Wage <- 10 + 2*Educ + 3*Ability + epsilon

# Your code should:
# 1. Run the "short" regression: Wage ~ Educ
# 2. Run the "long" regression: Wage ~ Educ + Ability
# 3. Compare the education coefficients
# 4. Calculate the OVB formula: beta_Ability * Cov(Educ,Ability)/Var(Educ
  )
# 5. Verify: short coefficient = long coefficient + OVB
```

(d) (6 points) A colleague suggests: "Just include more control variables and the OVB will disappear." Evaluate this advice. Under what conditions does adding controls reduce OVB? When might it make things worse?

(e) (6 points) Another colleague suggests using years of education as an *instrument* for ability, since they're correlated. Why is this a terrible idea?

Question 2: Interaction Terms (25 points)

A researcher studies the effect of a job training program on earnings. She estimates the following model:

$$\text{Earnings}_i = \beta_0 + \beta_1 \cdot \text{Treatment}_i + \beta_2 \cdot \text{Female}_i + \beta_3 \cdot (\text{Treatment}_i \times \text{Female}_i) + \varepsilon_i$$

The results are:

Variable	Coefficient	Std. Error
Intercept	35,000	1,200
Treatment	4,500	800
Female	−5,000	1,100
Treatment × Female	2,000	1,500

(a) (5 points) Interpret each of the four coefficients. Be precise about what comparison each one represents.

(b) (4 points) What is the estimated treatment effect for men? What is the estimated treatment effect for women? Show your calculations.

(c) (4 points) Test whether the treatment effect differs significantly between men and women. Set up the null and alternative hypotheses and conduct the test.

(d) (4 points) A journalist summarizes: "The program helps women more than men." Based on your answer to (c), is this claim statistically supported at $\alpha = 0.05$?

(e) (8 points) **R Simulation:** Explore the interaction model.

```
set.seed(2001)
n <- 500

# Generate data
Female <- rbinom(n, 1, 0.5)
Treatment <- rbinom(n, 1, 0.5)

# True effects:
# - Baseline (male, control): 35000
# - Treatment effect for men: 4500
# - Female penalty: -5000
# - Extra treatment effect for women: 2000
Earnings <- 35000 + 4500*Treatment - 5000*Female +
            2000*Treatment*Female + rnorm(n, 0, 8000)

# Your code should:
# 1. Estimate the model with interaction
# 2. Calculate treatment effect for men and women
# 3. Test whether interaction is significant
# 4. Create a visualization showing the four group means
```

Question 3: Robust Standard Errors (20 points)

This question explores why and when we need robust (heteroskedasticity-consistent) standard errors.

Setup

Consider the regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The "classical" OLS standard error for $\hat{\beta}_1$ assumes homoskedasticity:

$$SE_{\text{classical}}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}}$$

where $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-2}$ is the residual variance.

(a) (5 points) Explain what "heteroskedasticity" means and give a real-world example where you might expect it.

(b) (5 points) If heteroskedasticity is present:

- Are OLS coefficient estimates still unbiased?
- Are classical standard errors still correct?
- What are the consequences for hypothesis tests and confidence intervals?

(c) (10 points) **R Simulation:** Compare classical and robust SEs.

```r
set.seed(2001)
n_sims <- 1000
n <- 200

# Storage
reject_classical <- 0
reject_robust <- 0

for (sim in 1:n_sims) {
  # Generate X
  X <- runif(n, 1, 10)

  # Generate Y with heteroskedastic errors
  # Error variance increases with X
  epsilon <- rnorm(n, mean = 0, sd = X)  # SD = X
  Y <- 5 + 0*X + epsilon  # TRUE beta1 = 0

  # Fit model
  fit <- lm(Y ~ X)

  # Classical test: does t-stat reject H0: beta1 = 0?
  t_classical <- summary(fit)$coefficients[2, 3]
  if (abs(t_classical) > 1.96) reject_classical <- reject_classical + 1

  # Robust test (use sandwich package)
  library(sandwich)
  library(lmtest)
```

```r
  robust_test <- coeftest(fit, vcov = vcovHC(fit, type = "HC1"))
  t_robust <- robust_test[2, 3]
  if (abs(t_robust) > 1.96) reject_robust <- reject_robust + 1
}

# Type I error rates
cat("Classical SE rejection rate:", reject_classical/n_sims, "\n")
cat("Robust SE rejection rate:", reject_robust/n_sims, "\n")
cat("Nominal level: 0.05\n")
```

Report your findings. Which standard error maintains the correct Type I error rate under heteroskedasticity?

Question 4: F-tests and Model Comparison (25 points)

A researcher estimates two models of congressional voting:

Restricted Model:

$$\text{Vote}_i = \beta_0 + \beta_1 \cdot \text{Party}_i + \varepsilon_i$$

Unrestricted Model:

$$\text{Vote}_i = \beta_0 + \beta_1 \cdot \text{Party}_i + \beta_2 \cdot \text{Ideology}_i + \beta_3 \cdot \text{Seniority}_i + \varepsilon_i$$

The results are:

	Restricted	Unrestricted
R^2	0.45	0.62
Residual SS	550	380
df	433	431

(a) (5 points) What null hypothesis does the F-test for comparing these models test? Write it out in terms of the parameters.

(b) (6 points) Calculate the F-statistic:

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k-1)}$$

where q is the number of restrictions (added variables), RSS_r is the restricted residual sum of squares, and RSS_u is the unrestricted.

(c) (4 points) With $q = 2$ and $df_2 = 431$, the critical value at $\alpha = 0.05$ is approximately 3.0. Do you reject the null hypothesis? What do you conclude about the additional variables?

(d) (4 points) What is the relationship between the F-test and R^2? Show that an equivalent formula is:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k - 1)}$$

(e) (6 points) **R Simulation:** Verify F-test behavior.

```
set.seed(2001)
n <- 435

# Generate data where added variables have NO effect
Party <- rbinom(n, 1, 0.5)
Ideology <- rnorm(n)
Seniority <- rpois(n, lambda = 5)

# Vote depends ONLY on Party
Vote <- 50 + 20*Party + rnorm(n, 0, 15)

# Fit both models
fit_r <- lm(Vote ~ Party)
fit_u <- lm(Vote ~ Party + Ideology + Seniority)

# Your code should:
# 1. Conduct the F-test using anova()
# 2. Verify: under H0, we should reject ~5% of the time
# 3. Simulate 1000 datasets and check rejection rate

# Also: add a REAL effect of Ideology and see rejection rate increase
```

Submission Checklist

Before submitting, verify:

- ☐ All analytical work shows clear steps
- ☐ All R code runs without errors
- ☐ Simulation results are compared to analytical answers
- ☐ Collaborators are listed (if any)

This problem set covers material from Weeks 10–12: multiple regression, omitted variable bias, interaction effects, robust inference, and model testing.