# Gov 2001: Problem Set 4

**Introduction to Linear Regression**

Spring 2026

---

**Due:** Friday, April 10, 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:** OLS mechanics, properties (unbiasedness, consistency), residuals, $R^2$, interpretation

**Readings:** Blackwell Ch. 5–7; Aronow & Miller §4.1; Angrist & Pischke §3.1

---

## Question 1: The OLS Estimator (25 points)

This question builds your intuition for what OLS does and why it works.

### Part A: Deriving OLS

Consider the simple linear regression model:

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

OLS minimizes the sum of squared residuals:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

(a) (5 points) Take the first-order conditions with respect to $\beta_0$ and $\beta_1$. Show that they yield:

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

(b) (5 points) From the first equation, show that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

What does this tell you about the OLS regression line? (Hint: think about what point the line must pass through.)

(c) (5 points) Using the result from (b), show that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}$$

### Part B: Verify with Simulation

(d) (10 points) **R Simulation:** Verify the OLS formulas.

```r
set.seed(2001)
n <- 100

# Generate data: Y = 2 + 3*X + error
X <- rnorm(n, mean = 5, sd = 2)
epsilon <- rnorm(n, mean = 0, sd = 4)
Y <- 2 + 3*X + epsilon

# Calculate OLS estimates "by hand"
# beta1_hat = Cov(X,Y) / Var(X)
# beta0_hat = Ybar - beta1_hat * Xbar

# Your code should:
# 1. Calculate beta1_hat using the covariance/variance formula
# 2. Calculate beta0_hat using the mean formula
# 3. Compare to R's lm() function
# 4. Verify that residuals sum to zero
# 5. Verify that X and residuals are uncorrelated
```

## Question 2: Properties of OLS (25 points)

This question explores the finite-sample and asymptotic properties of OLS.

### Setup

The data generating process is:

$$Y_i = 5 + 2X_i + \varepsilon_i$$

where $X_i \sim N(3, 4)$ and $\varepsilon_i \sim N(0, 9)$, with $X$ and $\varepsilon$ independent.

(a) (5 points) Under standard OLS assumptions, is $\hat{\beta}_1$ unbiased for $\beta_1$? State which assumptions are needed for unbiasedness.

(b) (5 points) What is the variance of $\hat{\beta}_1$? Express it in terms of $\sigma^2$ (the error variance) and $\sum(X_i - \bar{X})^2$.

(c) (15 points) **R Simulation:** Verify unbiasedness and explore the sampling distribution.

```r
set.seed(2001)
n_sims <- 5000
n <- 50

# True parameters
beta0_true <- 5
beta1_true <- 2
sigma <- 3  # SD of epsilon

# Storage for estimates
beta1_estimates <- numeric(n_sims)

for (sim in 1:n_sims) {
  # Generate X (fixed in repeated samples, or random)
  X <- rnorm(n, mean = 3, sd = 2)

  # Generate errors and Y
  epsilon <- rnorm(n, mean = 0, sd = sigma)
  Y <- beta0_true + beta1_true * X + epsilon

  # Estimate OLS
  fit <- lm(Y ~ X)
  beta1_estimates[sim] <- coef(fit)[2]
}

# Your code should:
# 1. Calculate mean(beta1_estimates) - is it close to 2?
# 2. Calculate sd(beta1_estimates) - sampling variability
# 3. Create a histogram of beta1_estimates
# 4. Overlay a normal distribution (CLT prediction)
# 5. Repeat with n = 500 - how does the distribution change?
```

## Question 3: Interpreting Regression Output (25 points)

A researcher studies the relationship between campaign spending (in $100,000s) and vote share (% of votes received) for 435 U.S. House candidates. She estimates:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.2       1.4     27.3    <0.001
spending        2.1       0.3      7.0    <0.001
```

```
Residual SE: 12.5 on 433 df
Multiple R-squared: 0.102
```

(a) (4 points) Interpret the coefficient on spending (2.1) in one sentence. Be precise about units.

(b) (4 points) A campaign consultant says: "This proves that spending more money causes candidates to get more votes." Evaluate this claim. What would need to be true for this interpretation to be valid?

(c) (4 points) Construct a 95% confidence interval for the spending coefficient. Interpret it.

(d) (4 points) Interpret the $R^2$ of 0.102. What does this tell us about the model?

(e) (4 points) A journalist writes: "The $R^2$ of only 10% shows this model is bad." Respond to this claim. In what sense might a "low" $R^2$ still indicate an important finding?

(f) (5 points) **R Simulation:** Explore what $R^2$ means.

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

# Generate data with known R^2
# Y = a + b*X + epsilon
# R^2 = Var(b*X) / Var(Y)
#     = b^2 * Var(X) / (b^2*Var(X) + Var(epsilon))

# Create data with approximately R^2 = 0.10
X <- rnorm(n, mean = 5, sd = 3)
# To get R^2 ~ 0.10 with b = 2.1:
# Need Var(epsilon) such that b^2*Var(X)/(b^2*Var(X) + Var(e)) = 0.10
# => Var(e) = 9 * b^2 * Var(X) = 9 * 4.41 * 9 = 357
epsilon <- rnorm(n, mean = 0, sd = sqrt(357))
Y <- 38.2 + 2.1*X + epsilon

# Your code should:
# 1. Run lm(Y ~ X) and check R^2
# 2. Create a scatterplot with regression line
# 3. Comment: with R^2 = 0.10, how does the plot look?
# 4. Would you say spending "matters"?
```

## Question 4: Residual Analysis (25 points)

This question explores what residuals tell us about model fit.

### Setup

A researcher models the relationship between GDP per capita (X, in $1000s) and life expectancy (Y, in years) for 150 countries:

$$\text{LifeExp}_i = \beta_0 + \beta_1 \cdot \text{GDP}_i + \varepsilon_i$$

(a) (5 points) After running the regression, the researcher plots residuals against GDP and sees a clear pattern: residuals are positive for low and high GDP countries, but negative for middle-income countries (a U-shape).

What does this pattern suggest about the linear model? What might be a better specification?

(b) (5 points) The researcher also notices that residual variance seems larger for low-GDP countries than for high-GDP countries. What assumption might be violated? What are the consequences for OLS estimates and standard errors?

(c) (15 points) **R Simulation:** Explore residual diagnostics.

```r
set.seed(2001)
n <- 150

# Generate GDP (skewed toward lower values)
GDP <- rexp(n, rate = 0.1) + 1  # Range roughly 1-50

# True relationship is nonlinear: log transform
LifeExp <- 50 + 10*log(GDP) + rnorm(n, 0, 3)

# Fit linear model (misspecified)
fit_linear <- lm(LifeExp ~ GDP)

# Your code should:
# 1. Create residual vs. fitted plot
#    - Is there a pattern?
# 2. Create residual vs. GDP plot
#    - Same pattern?
# 3. Fit correct model: lm(LifeExp ~ log(GDP))
# 4. Compare residual plots for both models
# 5. Compare R^2 for both models

# Also calculate:
# - Sum of residuals (should be ~0)
# - Correlation of residuals with X (should be ~0)
```

## 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 8–9: regression as CEF approximation, OLS mechanics, properties, and interpretation.*