# OLS Derivation and Mechanics

## Gov 2001: Quantitative Social Science Methods I

Week 9, Lecture 17

Spring 2026

# For Today

## Required Reading

- Blackwell, Chapter 6 (pp. 119–138)
- Aronow & Miller, §4.1.1–4.1.3 (pp. 143–151)

Today: The full derivation of OLS and its algebraic properties.

# Roadmap

1. The optimization problem
2. Deriving the OLS formula
3. The normal equations
4. Algebraic properties
5. Numerical example

# Part I: The Optimization Problem

# The Setup

We have $n$ observations: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

We want to fit a line: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$

**Question:** How do we choose $\hat{\alpha}$ and $\hat{\beta}$?

**Answer:** Minimize the sum of squared prediction errors.

# Sum of Squared Residuals

For any candidate line $Y = a + bX$, the **residual** for observation $i$ is:

$$e_i(a, b) = Y_i - a - bX_i$$

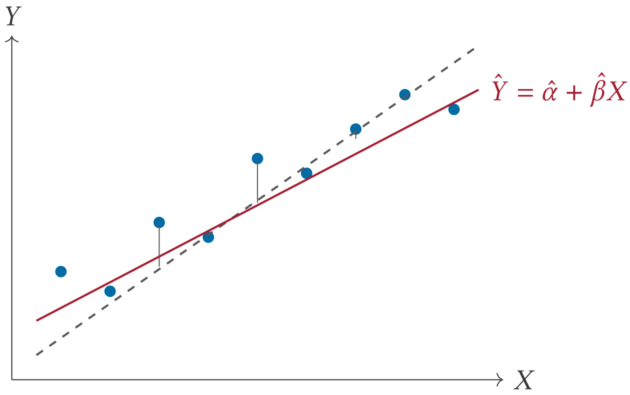
The **Sum of Squared Residuals** (SSR):

$$SSR(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

**OLS Problem:**

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

# Visualizing the Problem




OLS finds the line that makes $\sum e_i^2$ smallest.

# Part II: Deriving the OLS Formula

# The Calculus Approach

To minimize $SSR(a, b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$:

**Step 1:** Take partial derivatives with respect to $a$ and $b$.
**Step 2:** Set them equal to zero (first-order conditions).
**Step 3:** Solve the system of equations.

**Note:** SSR is a convex quadratic function, so the solution is a minimum.

# First-Order Condition for $\alpha$

$$\frac{\partial SSR}{\partial a} = \frac{\partial}{\partial a} \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Using the chain rule:

$$\frac{\partial SSR}{\partial a} = \sum_{i=1}^{n} 2(Y_i - a - bX_i) \cdot (-1) = -2 \sum_{i=1}^{n} (Y_i - a - bX_i)$$

Setting equal to zero:

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

This says: **Residuals sum to zero.**

## Solving for $\hat{\alpha}$

From the FOC:

$$\sum_{i=1}^{n} Y_i - n\hat{\alpha} - \hat{\beta} \sum_{i=1}^{n} X_i = 0$$

Divide by $n$:

$$\bar{Y} - \hat{\alpha} - \hat{\beta}\bar{X} = 0$$

Therefore:

$$\boxed{\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}}$$

The intercept ensures the line passes through $(\bar{X}, \bar{Y})$.

# First-Order Condition for $\beta$

$$\frac{\partial SSR}{\partial b} = \sum_{i=1}^{n} 2(Y_i - a - bX_i) \cdot (-X_i) = -2\sum_{i=1}^{n} X_i(Y_i - a - bX_i)$$

Setting equal to zero:

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

This says: **Residuals are uncorrelated with $X$.**

Equivalently:

$$\sum_{i=1}^{n} X_i\hat{e}_i = 0$$

## Solving for $\hat{\beta}$

Substitute $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ into the second FOC:

$$\sum_{i=1}^{n} X_i \left( Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i \right) = 0$$

Simplify:

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) - \hat{\beta} \sum_{i=1}^{n} X_i (X_i - \bar{X}) = 0$$

Using $\sum X_i (Y_i - \bar{Y}) = \sum (X_i - \bar{X})(Y_i - \bar{Y})$:

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

# The OLS Formulas

## OLS Estimators

**Slope:**

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

**Intercept:**

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

These are the **only** formulas you need for simple linear regression.

# Part III: The Normal Equations

# The Normal Equations

The first-order conditions are called the **normal equations**:

**Equation 1:**

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

**Equation 2:**

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

Two equations, two unknowns $\Rightarrow$ unique solution (when $\text{Var}(X) > 0$).

# Why "Normal" Equations?

**Not** because of the normal distribution!

"Normal" here means **perpendicular/orthogonal**.

The normal equations ensure:

- Residuals are orthogonal to the constant vector $(1, 1, \ldots, 1)$
- Residuals are orthogonal to the $X$ vector

This is a **projection** interpretation: OLS projects $Y$ onto the space spanned by $\{1, X\}$.

# Part IV: Algebraic Properties

# Properties of OLS Residuals

Define fitted values and residuals:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i, \qquad \hat{e}_i = Y_i - \hat{Y}_i$$

**Property 1:** $\displaystyle\sum_{i=1}^{n} \hat{e}_i = 0$

*Residuals sum to zero.*

**Property 2:** $\displaystyle\sum_{i=1}^{n} X_i \hat{e}_i = 0$

*Residuals are uncorrelated with $X$.*

Both follow directly from the normal equations.

## More Properties

**Property 3:** $\sum_{i=1}^{n} \hat{Y}_i \hat{e}_i = 0$
*Fitted values are uncorrelated with residuals.*

**Proof:** Since $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$:

$$\sum_i \hat{Y}_i \hat{e}_i = \hat{\alpha} \sum_i \hat{e}_i + \hat{\beta} \sum_i X_i \hat{e}_i = 0 + 0 = 0$$

**Property 4:** $\bar{\hat{Y}} = \bar{Y}$
*Mean of fitted values equals mean of $Y$.*

**Proof:** $\bar{\hat{Y}} = \hat{\alpha} + \hat{\beta}\bar{X} = \bar{Y}$

## Variance Decomposition

**Total Sum of Squares:**

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

**Explained Sum of Squares:**

$$SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

**Residual Sum of Squares:**

$$SSR = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

**Key result:** $SST = SSE + SSR$

# Proving the Decomposition

Start with:

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) = (\hat{Y}_i - \bar{Y}) + \hat{e}_i$$

Square both sides and sum:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i \hat{e}_i^2 + 2\sum_i (\hat{Y}_i - \bar{Y})\hat{e}_i$$

The cross term vanishes:

$$\sum_i (\hat{Y}_i - \bar{Y})\hat{e}_i = \sum_i \hat{Y}_i\hat{e}_i - \bar{Y}\sum_i \hat{e}_i = 0 - 0 = 0$$

Therefore: $\boxed{SST = SSE + SSR}$

## $R^2$: Coefficient of Determination

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

**Interpretation:** Fraction of variance in $Y$ "explained" by the regression.

**Range:** $0 \leq R^2 \leq 1$

- $R^2 = 0$: $\hat{Y}_i = \bar{Y}$ for all $i$ (horizontal line)
- $R^2 = 1$: $Y_i = \hat{Y}_i$ for all $i$ (perfect fit)

# $R^2$ and Correlation

**Key result:** In simple linear regression,

$$R^2 = r_{XY}^2$$

where $r_{XY}$ is the sample correlation between $X$ and $Y$.

**Proof sketch:**

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = r_{XY} \cdot \frac{s_Y}{s_X}$$

After algebra: $R^2 = r_{XY}^2$

This only holds for **simple** regression (one $X$ variable).

# Part V: Numerical Example

# Example: Study Hours and Exam Scores

(Political science applications use the same mechanics: campaign spending $\longrightarrow$ vote share, media exposure $\longrightarrow$ attitudes, aid $\longrightarrow$ development.)

| Student | $X_i$ (Hours) | $Y_i$ (Score) | $X_i - \bar{X}$ | $Y_i - \bar{Y}$ | $(X_i - \bar{X})(Y_i - \bar{Y})$ |
|---|---|---|---|---|---|
| 1 | 2 | 65 | $-3$ | $-10$ | 30 |
| 2 | 3 | 70 | $-2$ | $-5$ | 10 |
| 3 | 4 | 72 | $-1$ | $-3$ | 3 |
| 4 | 5 | 75 | 0 | 0 | 0 |
| 5 | 6 | 80 | 1 | 5 | 5 |
| 6 | 7 | 82 | 2 | 7 | 14 |
| 7 | 8 | 81 | 3 | 6 | 18 |
| Sum | 35 | 525 | 0 | 0 | 80 |
| Mean | 5 | 75 | | | |

$n = 7$, $\bar{X} = 5$, $\bar{Y} = 75$

## Computing $\hat{\beta}$

Numerator:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 80$$

Denominator:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = (-3)^2 + (-2)^2 + \cdots + 3^2 = 28$$

Slope:

$$\hat{\beta} = \frac{80}{28} = \frac{20}{7} \approx 2.86$$

**Interpretation:** Each additional hour of study is associated with about 2.86 more points on the exam.

# Computing $\hat{\alpha}$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 75 - \frac{20}{7} \times 5 = 75 - \frac{100}{7} \approx 60.71$$

**The estimated regression:**

$$\hat{Y} = 60.71 + 2.86X$$

**Interpretation of intercept:**
A student who studies 0 hours is predicted to score about 60.7 points.
(But this extrapolates beyond our data—use caution!)

## Fitted Values and Residuals

| Student | $X_i$ | $Y_i$ | $\hat{Y}_i$ | $\hat{e}_i = Y_i - \hat{Y}_i$ |
|---|---|---|---|---|
| 1 | 2 | 65 | 66.43 | $-1.43$ |
| 2 | 3 | 70 | 69.29 | 0.71 |
| 3 | 4 | 72 | 72.14 | $-0.14$ |
| 4 | 5 | 75 | 75.00 | 0.00 |
| 5 | 6 | 80 | 77.86 | 2.14 |
| 6 | 7 | 82 | 80.71 | 1.29 |
| 7 | 8 | 81 | 83.57 | $-2.57$ |
| Sum | | | | $\approx 0$ |

Check: $\sum \hat{e}_i = 0$ (up to rounding)

# Computing $R^2$

**SST:**

$$SST = \sum_i (Y_i - \bar{Y})^2 = 100 + 25 + 9 + 0 + 25 + 49 + 36 = 244$$

**SSR:**

$$SSR = \sum_i \hat{e}_i^2 \approx 2.04 + 0.51 + 0.02 + 0 + 4.59 + 1.65 + 6.61 \approx 15.43$$

$R^2$**:**

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{15.43}{244} \approx 0.937$$

About 94% of the variance in exam scores is "explained" by study hours.

# Part VI: Important Distinctions

# Residuals vs. Errors

**Population error** (unobserved):

$$\varepsilon_i = Y_i - \alpha^* - \beta^* X_i$$

**Sample residual** (observed):

$$\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$$

**Key differences:**

- $\varepsilon_i$ uses true (unknown) parameters
- $\hat{e}_i$ uses estimated parameters
- $\mathbb{E}[\varepsilon_i | X_i] = 0$ by construction of BLP
- $\sum \hat{e}_i = 0$ and $\sum X_i \hat{e}_i = 0$ by construction of OLS

# Estimation vs. Prediction

**Estimation:** Learning about population parameters.
What is $\beta^*$? $\Rightarrow$ Use $\hat{\beta}$

**Prediction:** Guessing $Y$ for a given $X$.
What is $Y$ when $X = 6$? $\Rightarrow$ Use $\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot 6$

**Prediction for new observation:**

$$\hat{Y}_{new} = \hat{\alpha} + \hat{\beta} X_{new}$$

(Uncertainty in prediction involves both estimation error and inherent variability.)

# Summary

**Today we covered:**

1. OLS minimizes $\sum(Y_i - a - bX_i)^2$
2. The formulas:
$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$
3. Normal equations: $\sum \hat{e}_i = 0$, $\sum X_i \hat{e}_i = 0$
4. Variance decomposition: $SST = SSE + SSR$
5. $R^2 = 1 - SSR/SST = r_{XY}^2$

# Looking Ahead

**Next lecture:** Properties of OLS

- Under what conditions is OLS unbiased?
- What is the sampling variance of $\hat{\beta}$?
- Gauss-Markov theorem: OLS is BLUE
- Standard errors and inference

**Key question:** We know *what* OLS does. Now we ask *how well* it does it.

OLS finds the line that minimizes
the sum of squared residuals.

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

This is the sample analog of the BLP:
$\beta^* = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$