

OLS as Sample BLP

Gov 2001: Quantitative Social Science Methods I

Week 8, Lecture 16

Spring 2026

For Today

Required Reading

- ▶ Blackwell, Chapter 5 (pp. 99–118)
- ▶ Aronow & Miller, §2.2.4 (pp. 80–88)

Today: The sample analog of the BLP—this is OLS.

Roadmap

1. The plug-in principle
2. OLS formula: sample analog of BLP
3. Why this is “least squares”
4. Fitted values and residuals
5. Example with real data

Part I: The Plug-In Principle

Recall: The Population BLP

The Best Linear Predictor minimizes:

$$\mathbb{E}[(Y - \alpha - \beta X)^2]$$

Solution:

$$\beta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \alpha^* = \mathbb{E}[Y] - \beta^* \mathbb{E}[X]$$

Problem: We don't know $\text{Cov}(X, Y)$, $\text{Var}(X)$, $\mathbb{E}[Y]$, or $\mathbb{E}[X]$.

We have data. What do we do?

The Plug-In Principle

A simple but powerful idea:

Replace population quantities with sample analogs.

Quantity	Population	Sample Analog
Mean of Y	$\mathbb{E}[Y]$	$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
Mean of X	$\mathbb{E}[X]$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variance of X	$\text{Var}(X)$	$\hat{S}_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
Covariance	$\text{Cov}(X, Y)$	$\hat{S}_{XY} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

Why Does Plug-In Work?

Law of Large Numbers:

Sample analogs converge to population quantities.

$$\begin{aligned}\bar{Y} &\xrightarrow{p} \mathbb{E}[Y], & \bar{X} &\xrightarrow{p} \mathbb{E}[X] \\ \hat{S}_{XY} &\xrightarrow{p} \text{Cov}(X, Y), & \hat{S}_X^2 &\xrightarrow{p} \text{Var}(X)\end{aligned}$$

Continuous mapping theorem:

If $g(\cdot)$ is continuous and $\hat{\theta} \xrightarrow{p} \theta$, then:

$$g(\hat{\theta}) \xrightarrow{p} g(\theta)$$

Part II: The OLS Formula

OLS: Sample Analog of BLP

Population BLP:

$$\beta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Sample analog (OLS):

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

The $\frac{1}{n}$ terms cancel—this simplifies the formula.

The Intercept

Population:

$$\alpha^* = \mathbb{E}[Y] - \beta^* \mathbb{E}[X]$$

Sample:

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

Key implication: The regression line passes through (\bar{X}, \bar{Y}) .

The “average person” sits on the regression line.

Alternative Formula for Slope

Expand the covariance formula:

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

You can also write:

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

Why? Because $\sum_i (X_i - \bar{X}) \bar{Y} = \bar{Y} \sum_i (X_i - \bar{X}) = 0$.

This form is useful for deriving properties of $\hat{\beta}$.

Part III: Why “Least Squares”?

The Least Squares Criterion

OLS minimizes the **sum of squared residuals**:

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

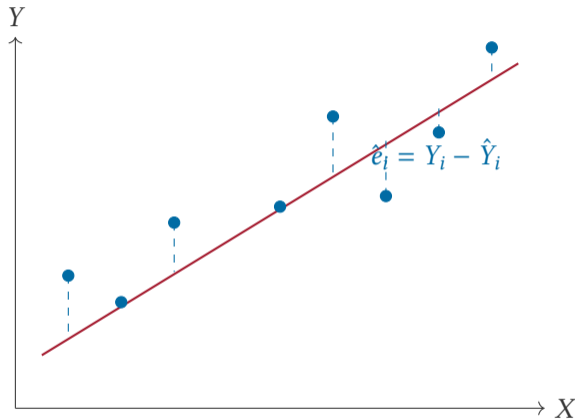
Compare to the population BLP:

$$\alpha^*, \beta^* = \arg \min_{a,b} \mathbb{E}[(Y - a - bX)^2]$$

Same logic: Minimize squared prediction errors.

Different: Population expectation vs. sample sum.

Visual Intuition



OLS chooses the line that makes $\sum \hat{e}_i^2$ as small as possible.

Why Squared Errors?

Why not minimize $\sum |Y_i - a - bX_i|$ (absolute errors)?

Mathematical reasons:

- ▶ Squared function is differentiable everywhere
- ▶ First-order conditions give closed-form solution
- ▶ Connects to variance and covariance

Statistical reasons:

- ▶ Sample analog of minimizing MSE in population
- ▶ Under normality, OLS = Maximum Likelihood

(Minimizing absolute deviations gives **Least Absolute Deviations**—different estimator, robust to outliers.)

Part IV: Fitted Values and Residuals

Fitted Values

The **fitted value** is our prediction for Y_i :

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

This is the point on the regression line at $X = X_i$.

Interpretation: Given someone's X value, what does the regression predict for their Y ?

Residuals

The **residual** is the prediction error:

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

- ▶ $\hat{e}_i > 0$: Observation above the regression line
- ▶ $\hat{e}_i < 0$: Observation below the regression line
- ▶ $\hat{e}_i = 0$: Observation exactly on the line

Residuals are sample analogs of errors:

$$\varepsilon_i = Y_i - \alpha^* - \beta^*X_i \quad (\text{population})$$

Key Properties of OLS Residuals

Property 1: Residuals sum to zero.

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

Property 2: Residuals are uncorrelated with X .

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

These are the **first-order conditions** from the minimization. They follow directly from solving $\frac{\partial SSR}{\partial a} = 0$ and $\frac{\partial SSR}{\partial b} = 0$.

Where Do These Come From?

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

First-order condition for a :

$$\begin{aligned}\frac{\partial SSR}{\partial a} &= -2 \sum_{i=1}^n (Y_i - a - bX_i) = 0 \\ \Rightarrow \sum_{i=1}^n \hat{e}_i &= 0\end{aligned}$$

First-order condition for b :

$$\begin{aligned}\frac{\partial SSR}{\partial b} &= -2 \sum_{i=1}^n X_i (Y_i - a - bX_i) = 0 \\ \Rightarrow \sum_{i=1}^n X_i \hat{e}_i &= 0\end{aligned}$$

Decomposing Variation

Every observation can be written as:

$$Y_i = \hat{Y}_i + \hat{e}_i$$

$$(\text{Actual}) = (\text{Fitted}) + (\text{Residual})$$

Sum of squares version:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SSE} + \underbrace{\sum_{i=1}^n \hat{e}_i^2}_{SSR}$$

$$\text{Total Variation} = \text{Explained Variation} + \text{Residual Variation}$$

R^2 : Goodness of Fit

The **coefficient of determination**:

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

Interpretation: Fraction of variance in Y “explained” by the regression.

- ▶ $R^2 = 0$: Regression explains nothing (horizontal line)
- ▶ $R^2 = 1$: All points on the line (perfect fit)
- ▶ $0 < R^2 < 1$: Typical case

R^2 Caution

Don't obsess over R^2 !

R^2 tells you:

- ▶ How much variation in Y is associated with X
- ▶ How well the line fits the data

R^2 does NOT tell you:

- ▶ Whether X causes Y
- ▶ Whether the relationship is practically important
- ▶ Whether you've included the right variables

A regression with low R^2 can still be useful and informative.

Part V: Consistency of OLS

OLS Is Consistent

Because OLS is a plug-in estimator, it inherits consistency from LLN.

$$\hat{\beta} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)} \xrightarrow{p} \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta^*$$

As $n \rightarrow \infty$:

- ▶ $\hat{\beta}$ converges in probability to β^*
- ▶ $\hat{\alpha}$ converges in probability to α^*

OLS recovers the population BLP with enough data.

What Exactly Are We Estimating?

Be precise:

$$\hat{\beta} \xrightarrow{p} \beta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

This is the **population BLP coefficient**—a well-defined quantity even if:

- ▶ The CEF is highly nonlinear
- ▶ There is no “true model”
- ▶ X doesn’t “cause” Y

The BLP always exists.

OLS consistently estimates it.

Part VI: Example

Example: Education and Wages

Suppose we have data on $n = 1000$ workers:

- ▶ Y_i : Log hourly wage
- ▶ X_i : Years of education

Summary statistics:

	Mean	Std Dev
Log wage (Y)	2.95	0.54
Education (X)	13.2	2.8

Correlation: $r_{XY} = 0.45$

Computing the OLS Estimates

Step 1: Compute sample covariance.

$$\widehat{\text{Cov}}(X, Y) = r_{XY} \cdot s_X \cdot s_Y = 0.45 \times 2.8 \times 0.54 = 0.68$$

Step 2: Compute sample variance.

$$\widehat{\text{Var}}(X) = s_X^2 = (2.8)^2 = 7.84$$

Step 3: Slope.

$$\hat{\beta} = \frac{0.68}{7.84} = 0.087$$

Step 4: Intercept.

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 2.95 - 0.087 \times 13.2 = 1.80$$

Interpreting the Results

The estimated regression:

$$\widehat{\log(\text{wage})} = 1.80 + 0.087 \times \text{Education}$$

Interpretation of $\hat{\beta} = 0.087$:

Each additional year of education is associated with an 8.7% higher wage, on average.

Important: This is a **descriptive** statement.

- ▶ It describes the linear relationship in the population
- ▶ It does NOT mean education *causes* 8.7% higher wages
- ▶ Causation requires additional assumptions

What Did We Learn?

OLS gave us an estimate of:

$$\beta^* = \frac{\text{Cov}(\text{Education}, \log \text{Wage})}{\text{Var}(\text{Education})}$$

This tells us:

- ▶ How education and wages move together in the population
- ▶ The best linear approximation to the CEF
- ▶ The slope of the line that minimizes mean squared prediction error

Whether this is a **causal effect** depends on:

- ▶ Selection: Who gets more education?
- ▶ Confounders: Ability, family background, motivation, etc.

Part VII: OLS in R

OLS in R: The `lm()` Function

R makes OLS simple with the `lm()` function.

```
# Read data (example: campaign spending data)
campaign_data <- read.csv("data/campaign_data.csv")

# View first few rows
head(campaign_data)
##      vote_share spending incumbent
## 1         48.2     125.3          0
## 2         52.7     203.4          1
## 3         44.1      87.6          0
```

Goal: Estimate the relationship between campaign spending and vote share.

Running OLS: Simple Regression

Basic syntax: `lm(Y ~ X, data = df)`

```
# Simple regression: vote_share on spending
model <- lm(vote_share ~ spending, data = campaign_data)

# View the results
summary(model)
```

The formula `vote_share ~ spending` means:

$$\text{vote_share}_i = \alpha + \beta \times \text{spending}_i + e_i$$

The tilde (~) reads as “is modeled as a function of.”

Extracting OLS Results

```
# Extract coefficients
coef(model)
## (Intercept)      spending
##      35.243912      0.082156

# Fitted values (predicted Y)
fitted(model)[1:5] # First 5 predictions

# Residuals
residuals(model)[1:5] # First 5 residuals

# R-squared
summary(model)$r.squared
## [1] 0.4532
```

Interpretation: Each additional \$1K in spending is associated with 0.08 percentage points higher vote share.

Multiple Regression

Add more predictors with +:

```
# Multiple regression: add incumbent status
model_full <- lm(vote_share ~ spending + incumbent,
                 data = campaign_data)

# View results
coef(model_full)
## (Intercept)      spending      incumbent
##    32.156234      0.067823      8.234521
```

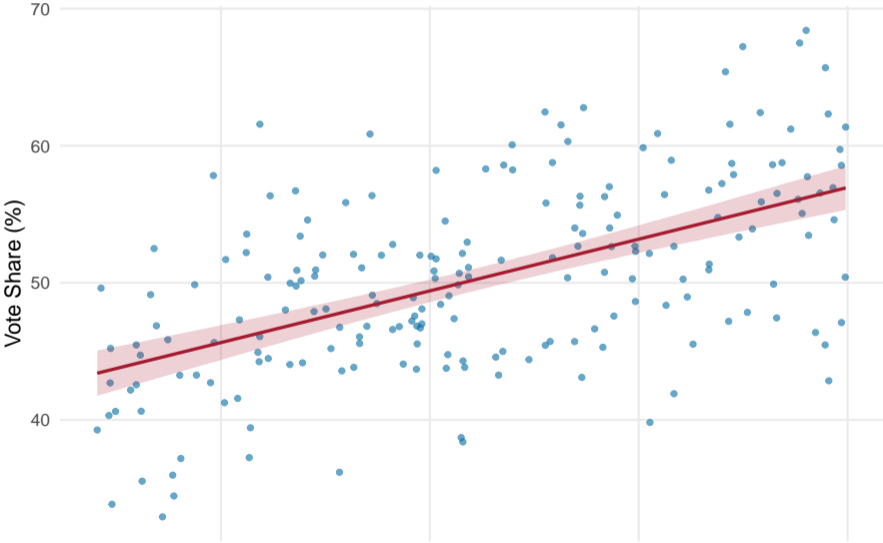
Interpretation:

- ▶ Spending: +\$1K \rightarrow +0.07 points (holding incumbency constant)
- ▶ Incumbent: +8.2 points (holding spending constant)

Visualizing the OLS Fit

Campaign Spending and Vote Share

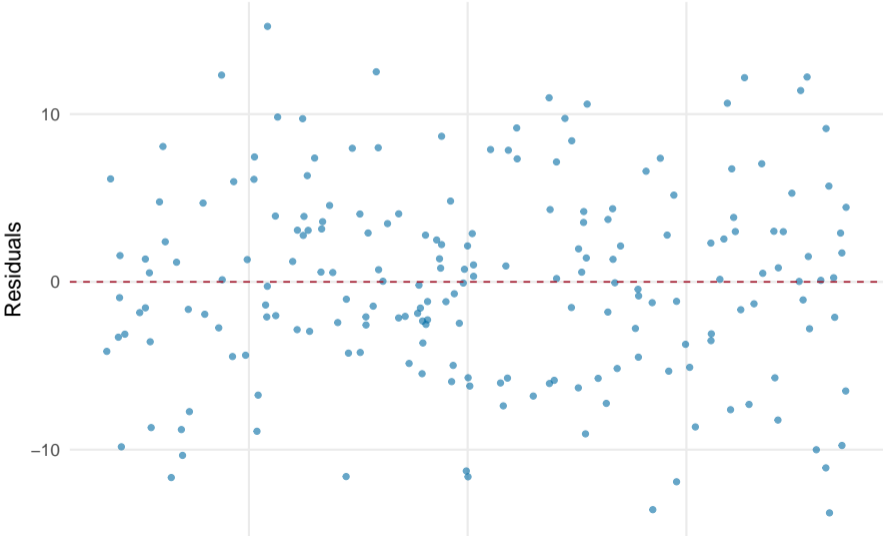
OLS regression line with 95% confidence band



Residual Diagnostics

Residual Plot

Checking for patterns in residuals

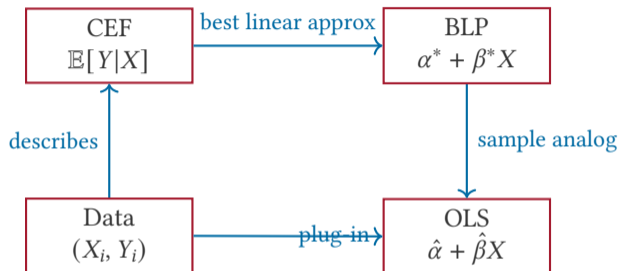


Summary

The big picture:

1. BLP is a **population** object: $\beta^* = \text{Cov}(X, Y) / \text{Var}(X)$
2. OLS is the **sample analog**: $\hat{\beta} = \widehat{\text{Cov}} / \widehat{\text{Var}}$
3. OLS minimizes $\sum \hat{e}_i^2$ —hence “least squares”
4. Fitted values: $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$
5. Residuals: $\hat{e}_i = Y_i - \hat{Y}_i$
6. Key properties: $\sum \hat{e}_i = 0$, $\sum X_i \hat{e}_i = 0$
7. OLS is **consistent**: $\hat{\beta} \xrightarrow{p} \beta^*$

Connecting the Ideas



CEF \rightarrow BLP: Restrict to linear predictors

BLP \rightarrow OLS: Replace population with sample

Looking Ahead

Next lecture: OLS mechanics in detail

- ▶ Full derivation of the OLS formula
- ▶ Algebraic properties
- ▶ Computation and implementation

Coming soon:

- ▶ Unbiasedness of OLS (when?)
- ▶ Sampling variance of $\hat{\beta}$
- ▶ Standard errors and hypothesis tests
- ▶ Multiple regression

OLS is simply the sample analog of the BLP.

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

It consistently estimates the best linear approximation to the conditional expectation function.

Whether the CEF is linear or not, causal or not—OLS gives you the best linear predictor. What you *interpret* it as is up to you.