OLS Mechanics I

Derivation, Fit, and Optimality

Gov 2001 · Scott Cunningham · April 15, 2026

Every four years, economists ask whether the economy elects presidents

- **1948:** Economy growing → Truman wins
- **1980:** Recession → Carter loses to Reagan
- **1992:** "It's the economy, stupid" → Clinton wins
- **2008:** Financial crisis → Obama wins

Fair (1978)

"The Effect of Economic Events on Votes for President"

Review of Economics and Statistics

Fair quantified this intuition: 24 elections, 1916–2016. Today we work through exactly how.

Ray Fair spent five decades studying how economic conditions shape presidential elections

- **Ray Fair** — John M. Musser Professor of Economics, Yale University
- **Sharon Oster** (Yale School of Management) — his wife, pioneering industrial economist; died June 10, 2022
- **Emily Oster** (Harvard PhD, 2005) — their daughter; Professor at Brown; author of *Expecting Better*, *Cribsheet*
- **Jesse Shapiro** (Harvard PhD, 2005) — son-in-law; Harvard Professor; MacArthur Fellow, 2021

An economics dynasty

Two Harvard PhDs (2005)
One MacArthur Fellow
One bestselling author
All connected to Yale

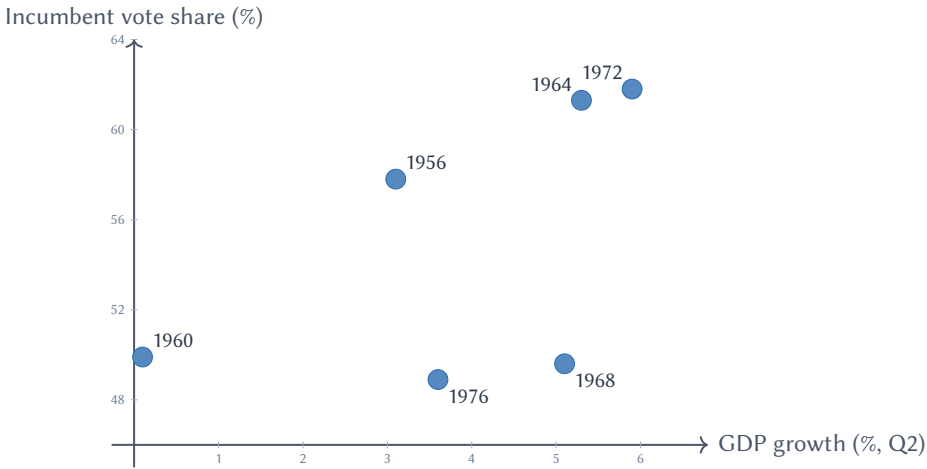
As of 2024, Fair's model still forecasts presidential elections on his Yale website.

Fair's data: six elections, three predictors, one outcome

Year	GDP growth (%)	Inflation (%)	Vote share (%)	Incumbent?
1956	3.1	2.1	57.8	Yes
1960	0.1	2.5	49.9	No
1964	5.3	1.1	61.3	No
1968	5.1	3.5	49.6	No
1972	5.9	4.8	61.8	Yes
1976	3.6	6.6	48.9	Yes

Y = incumbent party presidential vote share (%); GDP growth = Q2 of election year. Full dataset: $n = 24$ elections, 1916–2016.

The pattern: good economies re-elect incumbents



OLS estimates the BLP coefficient β — here is what it finds

Variable	$\hat{\beta}$	SE	Interpretation
Intercept	46.3	1.8	baseline vote (%)
GDP growth	+0.70	0.23	+1% growth $\Rightarrow$ +0.7 pp
Inflation	−0.68	0.22	+1% inflation $\Rightarrow$ −0.7 pp
Incumbent	+4.0	1.4	incumbency bonus: +4 pp

$n = 24$ elections (1916–2016), $R^2 = 0.73$

Estimand: β (BLP in the population) **Estimator:** OLS **Estimate:** $\hat{\beta}$ above

Today: how does OLS produce $\hat{\beta}$ from data?

Running Fair's model in R takes two lines

```r
# Load Fair election data (n=24, 1916-2016)
fair <- read.csv("fair_elections.csv")

# Fit the OLS model
fit <- lm(vote ~ gdpg + inflation + incumbent,
          data = fair)

summary(fit)
```

`lm()` minimizes the sum of squared residuals. Everything else in today's lecture explains *why* that works.

OLS is the BLP's sample analog: replace $\mathbb{E}[\,\cdot\,]$ with $\frac{1}{n}\sum_i$

Population BLP

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

uses $\mathbb{E}[\,\cdot\,]$

DATA →

Sample OLS

$$\hat{\beta}_1 = \frac{\hat{S}_{XY}}{\hat{S}_{XX}}$$

uses $\frac{1}{n}\sum_i$

Plug-in principle: same minimization problem, sample moments replace population moments

Today: four building blocks

1. **Derivation** — bivariate FOCs $\longrightarrow \hat{\beta}_1 = \hat{S}_{XY}/\hat{S}_{XX}$
2. **Matrix form** — $\hat{\beta} = (X'X)^{-1}X'Y$
3. **Model fit** — R^2, adjusted R^2, overfitting
4. **Gauss–Markov** — when OLS is BLUE

Part I: Deriving $\hat{\beta}$

Minimize squared residuals; solve two equations

Notation: (α, β_1) are the true parameters; (a, b) are the candidate values we search over

The model:

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

Residual for any candidate line (a, b):

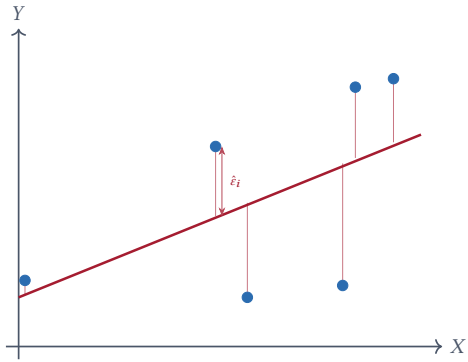
$$e_i(a, b) = Y_i - a - bX_i$$

OLS finds the minimizing pair:

$$(\hat{\alpha},\ \hat{\beta}_1) = \arg\min_{a,\,b} \sum_{i=1}^{n} e_i^2$$

The OLS residual: $\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}_1 X_i$ computed from data, unlike ε_i

The OLS problem: find the line that minimizes squared vertical gaps



Choose (a, b) to minimize:

$$\text{SSR}(a, b) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

where $\hat{\varepsilon}_i = Y_i - a - bX_i$

$$(\hat{\alpha}, \hat{\beta}_1) = \arg\min_{a,\, b} \text{SSR}(a, b)$$

FOC 1 (w.r.t. a): residuals sum to zero

Differentiate SSR with respect to a, set to zero:

$$\frac{\partial \text{SSR}}{\partial a} = -2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i) = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

Divide by n:

$$\bar{Y} - \hat{\alpha} - \hat{\beta}_1 \bar{X} = 0$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X} \quad \textbf{The OLS line passes through } (\bar{X}, \bar{Y})$$

Fair: $\bar{X} = 3.85$, $\bar{Y} = 54.88$. Given $\hat{\beta}_1$, $\hat{\alpha}$ is immediate.

FOC 2 (w.r.t. b): residuals are uncorrelated with X

Differentiate SSR with respect to b, set to zero:

$$\frac{\partial \text{SSR}}{\partial b} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i) = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$$

Substitute $\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}$ and rearrange:

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \frac{\hat{S}_{XY}}{\hat{S}_{XX}} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}$$

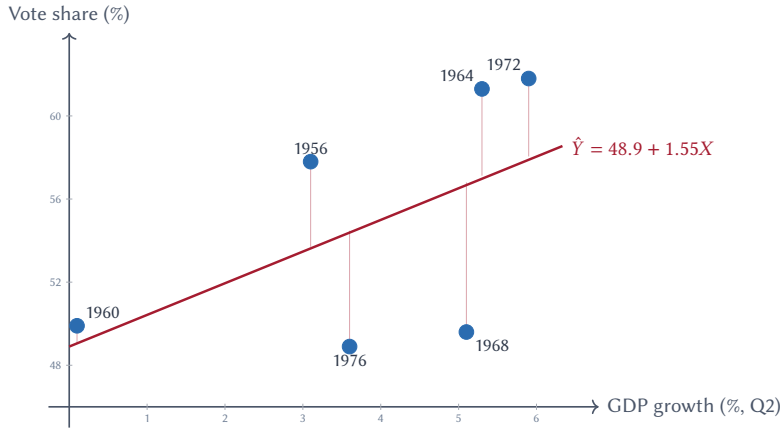
Two orthogonality conditions: $\sum \hat{\varepsilon}_i = 0$ and $\sum X_i \hat{\varepsilon}_i = 0$. Both reappear in matrix form.

Hand-computing $\hat{\beta}_1$ for the six Fair elections

Year	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1956	3.1	57.8	−0.75	+2.92	−2.19	0.56
1960	0.1	49.9	−3.75	−4.98	+18.68	14.06
1964	5.3	61.3	+1.45	+6.42	+9.31	2.10
1968	5.1	49.6	+1.25	−5.28	−6.60	1.56
1972	5.9	61.8	+2.05	+6.92	+14.18	4.20
1976	3.6	48.9	−0.25	−5.98	+1.50	0.06
Sum	23.1	329.3	0	0	**34.88**	**22.55**
Mean	3.85	54.88				

$$\hat{\beta}_1 = \frac{34.88}{22.55} \approx 1.55 \qquad \hat{\alpha} = 54.88 - 1.55 \times 3.85 \approx 48.9$$

The OLS line through Fair's six elections



1968: GDP growth 5.1% but vote share only 49.6%. Model predicts ≈57%. The residual is Vietnam and the Democratic convention.

Part II: Matrix Form

$$\hat{\beta} = (X'X)^{-1}X'Y$$

OLS imposes two orthogonality conditions on the residuals

Condition 1

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

Residuals average to zero

Line passes through $(\bar{X}, \bar{Y})$

Condition 2

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

Residuals uncorrelated with X

X cannot predict the leftovers

OLS residuals are orthogonal to every variable in the model

X_i appears because the chain rule forces it out when we minimize SSR

Differentiate SSR with respect to b, set to zero:

$$\frac{\partial}{\partial b}\sum_i e_i^2 = \sum_i 2\left(Y_i - a - bX_i\right)\cdot \underbrace{\left(-X_i\right)}_{\text{chain rule}} = 0$$

Divide by -2:

$$\sum_i X_i\hat{\varepsilon}_i = 0 \qquad \Longleftrightarrow \qquad \widehat{\text{Cov}}(X,\hat{\varepsilon}) = 0$$

$\hat{\beta}_1$ absorbs all the covariance between X and Y. Whatever remains in $\hat{\varepsilon}$ is orthogonal to X.

Both conditions hold in Fair's elections: verified with actual numbers

Year	X_i	Y_i	$\hat{\varepsilon}_i$	$X_i\hat{\varepsilon}_i$
1956	3.1	57.8	+4.1	+12.7
1960	0.1	49.9	+0.8	+0.1
1964	5.3	61.3	+4.2	+22.3
1968	5.1	49.6	−7.2	−36.7
1972	5.9	61.8	+3.7	+21.8
1976	3.6	48.9	−5.6	−20.2
Sum			**0.0**	**0.0**

Condition 1: $(+4.1) + (+0.8) + (+4.2) + (-7.2) + (+3.7) + (-5.6) = \mathbf{0.0}$

Condition 2: $3.1(4.1) + 0.1(0.8) + 5.3(4.2) + 5.1(-7.2) + 5.9(3.7) + 3.6(-5.6) = \mathbf{0.0}$

Six scalar equations are one matrix equation: $Y = X\beta + \varepsilon$

Six separate equations:

$$
\begin{aligned}
57.8 &= \alpha \cdot 1 + \beta_1 \cdot 3.1 + \varepsilon_1 \\
49.9 &= \alpha \cdot 1 + \beta_1 \cdot 0.1 + \varepsilon_2 \\
61.3 &= \alpha \cdot 1 + \beta_1 \cdot 5.3 + \varepsilon_3 \\
49.6 &= \alpha \cdot 1 + \beta_1 \cdot 5.1 + \varepsilon_4 \\
61.8 &= \alpha \cdot 1 + \beta_1 \cdot 5.9 + \varepsilon_5 \\
48.9 &= \alpha \cdot 1 + \beta_1 \cdot 3.6 + \varepsilon_6
\end{aligned}
$$

One matrix equation:

$$
\underbrace{\begin{bmatrix} 57.8 \\ 49.9 \\ 61.3 \\ 49.6 \\ 61.8 \\ 48.9 \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & 3.1 \\ 1 & 0.1 \\ 1 & 5.3 \\ 1 & 5.1 \\ 1 & 5.9 \\ 1 & 3.6 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \alpha \\ \beta_1 \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}}_{\varepsilon}
$$

The 1s column exists because α multiplies 1 in every equation. One column per parameter.

Each observation is a row in X; transposing makes each variable a row in X'

Design matrix X (6×2)**:**

$$X = \begin{bmatrix} 1 & 3.1 \\ 1 & 0.1 \\ 1 & 5.3 \\ 1 & 5.1 \\ 1 & 5.9 \\ 1 & 3.6 \end{bmatrix} \begin{matrix} \leftarrow 1956 \\ \leftarrow 1960 \\ \leftarrow 1964 \\ \leftarrow 1968 \\ \leftarrow 1972 \\ \leftarrow 1976 \end{matrix}$$

Transpose X' (2×6)**:**

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3.1 & 0.1 & 5.3 & 5.1 & 5.9 & 3.6 \end{bmatrix}$$

Row 1 $= [1, 1, \ldots, 1]$ intercept column of X

Row 2 $= [X_1, \ldots, X_6]$ GDP column of X

Row j of X' = column j of X = the multipliers from FOC j

$X'\hat{\varepsilon} = 0$: Fair's numbers confirm both conditions in one matrix product

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3.1 & 0.1 & 5.3 & 5.1 & 5.9 & 3.6 \end{bmatrix}}_{X',\ 2\times 6} \underbrace{\begin{bmatrix} +4.1 \\ +0.8 \\ +4.2 \\ -7.2 \\ +3.7 \\ -5.6 \end{bmatrix}}_{\hat{\varepsilon},\ 6\times 1} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

k regressors $\Rightarrow$ k rows in X' $\Rightarrow$ k orthogonality conditions at once. **That is why we need matrix notation.**

Preview: $(X'X)\hat{\beta} = X'Y$ — variation times slope equals co-movement.

The full Fair model: GDP growth, inflation, and incumbency predict vote share

$$\text{Vote}_i = \beta_0 + \beta_1 \text{ GDP growth}_i + \beta_2 \text{ Inflation}_i + \beta_3 \text{ Incumbent}_i + \varepsilon_i$$

	Variable	1956	1960
β_0	Intercept	—	—
β_1	GDP growth (%)	3.1	0.1
β_2	Inflation (%)	2.1	2.5
β_3	Incumbent party	1	0
Y	Vote share (%)	57.8	49.9

Four parameters. Six elections. The numbers 3.1, 2.1, 1 on the next slide are GDP growth, inflation, and incumbent.

With three regressors, scalar notation becomes unmanageable

Scalar: 6 equations, 4 unknowns

$$
\begin{aligned}
57.8 &= \beta_0 + \beta_1(3.1) + \beta_2(2.1) + \beta_3(1) + \varepsilon_1 \\
49.9 &= \beta_0 + \beta_1(0.1) + \beta_2(2.5) + \beta_3(0) + \varepsilon_2 \\
&\vdots
\end{aligned}
$$

Solve for $(\beta_0, \beta_1, \beta_2, \beta_3)$ simultaneously? Painful.

Matrix: one equation

$Y = X\beta + \varepsilon$, where

$$
Y = \begin{pmatrix} 57.8 \\ 49.9 \\ \vdots \\ 48.9 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 3.1 & 2.1 & 1 \\ 1 & 0.1 & 2.5 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}
$$

$$
Y : n \times 1 \quad X : n \times k \quad \beta : k \times 1
$$

Each entry of $X'X$ is a dot product — the intercept column makes $(1,1) = n$

For $n = 3$ symbolic observations:

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \end{pmatrix}, \qquad X' = \begin{pmatrix} 1 & 1 & 1 \\ X_1 & X_2 & X_3 \end{pmatrix}$$

Entry of $X'X$	Dot product
$(1,1)$	$[1,1,1] \cdot [1,1,1] = \mathbf{n}$
$(1,2)$	$[1,1,1] \cdot [X_1, X_2, X_3] = \sum X_i$
$(2,2)$	$[X_1, X_2, X_3] \cdot [X_1, X_2, X_3] = \sum X_i^2$

$$(1,1) = n\text{: intercept column is all 1s} \implies \underbrace{[1,\ldots,1]}_{n} \cdot \underbrace{[1,\ldots,1]}_{n} = n$$

$X'X$ and $X'Y$ encode all the variation we need

Bivariate Fair model (intercept + GDP growth only, 6 elections):

$$X'X = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} = \begin{pmatrix} 6 & 23.1 \\ 23.1 & 111.5 \end{pmatrix}$$

$(1,1) = n; \quad (1,2) = \sum X_i = 23.1$

$(2,2) = \sum X_i^2 = 111.5$

$$X'Y = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix} = \begin{pmatrix} 329.3 \\ 1302.7 \end{pmatrix}$$

$(1) = \sum Y_i = 329.3$

$(2) = \sum X_i Y_i = 1302.7$

$X'X$: variation in X $\qquad$ $X'Y$: covariation of X and Y

Matrix form rewrites the same objective: $\min_\beta \sum e_i^2 = \min_\beta \; e'e$

The residual vector $(n \times 1)$:

$$e(\beta) \;=\; Y - X\beta, \qquad e_i(\beta) \;=\; Y_i - X_i'\beta \quad \text{(vertical gap, observation } i\text{)}$$

Why $e'e = \sum e_i^2$**:**

$$e'e \;=\; \underbrace{[\, e_1,\; e_2,\; \ldots,\; e_n \,]}_{1\times n} \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}}_{n\times 1} \;=\; e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\hat{\beta} = \arg\min_\beta \; (Y - X\beta)'(Y - X\beta)$$

Each e_i is one vertical gap from the scatter plot. $e'e$ sums all the squares — same objective as Part I.

The normal equations lead directly to $\hat{\beta} = (X'X)^{-1}X'Y$

Minimize $(Y - X\beta)'(Y - X\beta)$. Differentiate, set to zero:

$$-2X'Y + 2X'X\hat{\beta} = 0 \quad \Longrightarrow \quad \underbrace{X'X\hat{\beta} = X'Y}_{\textbf{normal equations}}$$

Pre-multiply by $(X'X)^{-1}$**:**

$$\hat{\beta} = (X'X)^{-1}X'Y$$

The scalar FOCs from Part I are these two equations written out entry by entry. Matrix form is the same arithmetic, organized.

A matrix inverse *undoes* multiplication — not the same as a transpose, which rearranges

Transpose A': rows $\leftrightarrow$ columns

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}' = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

A rearrangement. Doesn't undo anything.

Inverse A^{-1}: $A^{-1}A = I$

$$A^{-1}A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ (identity)}$$

Like $3^{-1} \cdot 3 = 1$: cancels A out.

A^{-1} is the unique matrix such that $A^{-1}A = I$ — it is the matrix that cancels A out completely

$(X'X)^{-1}$ isolates $\hat{\beta}$ the same way 3^{-1} isolates x in $3x = 12$

Scalar:

$$3x = 12$$

$$3^{-1} \cdot 3x = 3^{-1} \cdot 12$$

$$x = \tfrac{1}{3} \cdot 12 = 4$$

Matrix:

$$X'X\hat{\beta} = X'Y$$

$$(X'X)^{-1}X'X\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

No matrix "division" — only inverses. But the logic is identical to the scalar case.

Applied to Fair's $X'X$: formula first, then $\det = 135.4$, then the inverse

General 2×2 formula:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Swap $a \leftrightarrow d$, negate b and c, divide every entry by $\det = ad - bc$.

For $X'X$:

$$X'X = \begin{pmatrix} a = 6 & b = 23.1 \\ c = 23.1 & d = 111.5 \end{pmatrix}, \qquad \det = 6(111.5) - 23.1^2 = 669.0 - 533.6 = \mathbf{135.4}$$

$$(X'X)^{-1} = \frac{1}{135.4} \begin{pmatrix} 111.5 & -23.1 \\ -23.1 & 6 \end{pmatrix} \approx \begin{pmatrix} 0.823 & -0.171 \\ -0.171 & 0.044 \end{pmatrix}$$

$\det(X'X) = 0$: no inverse, $\hat{\beta}$ undefined — perfect multicollinearity.

Multiplying $(X'X)^{-1}X'Y$ recovers the same $\hat{\beta}$ as the scalar formula

Recall: $(X'X)^{-1} \approx \begin{pmatrix} 0.823 & -0.171 \\ -0.171 & 0.044 \end{pmatrix}$, $X'Y = \begin{pmatrix} 329.3 \\ 1302.7 \end{pmatrix}$

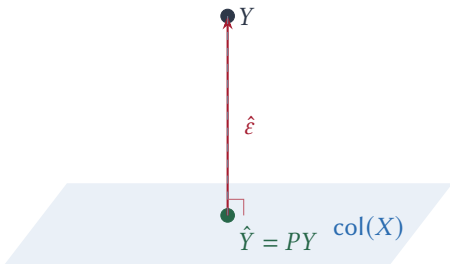
Multiply row by column:

$$\hat{\beta}_0 = 0.823(329.3) - 0.171(1302.7) \approx 48.9$$

$$\hat{\beta}_1 = -0.171(329.3) + 0.044(1302.7) \approx 1.55$$

$(X'X)^{-1}X'Y$ gives the same answer as the scalar formula: $\hat{\beta}_1 \approx 1.55$, $\hat{\beta}_0 \approx 48.9$

OLS is a projection: $\hat{Y}$ is the shadow of Y onto col(X)



$$P = X(X'X)^{-1}X' \quad \text{(hat matrix)}$$

$$\hat{Y} = PY \qquad \hat{\varepsilon} = (I - P)Y$$

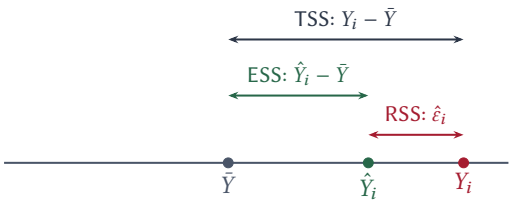
$X'\hat{\varepsilon} = 0$
Orthogonality is the projection condition

Kaixiao will develop the full geometry of projection on Monday — hat matrix, leverage, and FWL in matrix form. Today: the picture.

Part III: Model Fit

How well does the model explain the data?

Variance decomposes: total = explained + residual



$$\underbrace{(Y_i - \bar{Y})}_{\text{TSS}_i} = \underbrace{(\hat{Y}_i - \bar{Y})}_{\text{ESS}_i} + \underbrace{\hat{\varepsilon}_i}_{\text{RSS}_i}$$

$$\text{TSS} = \text{ESS} + \text{RSS}$$

Holds because $\hat{\varepsilon} \perp \hat{Y}$
(consequence of $X'\hat{\varepsilon} = 0$)

R^2 measures the fraction of variance the model explains

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Fair model	Regressors	R^2	
GDP growth only	1	0.44	economy alone
GDP + inflation + incumbent	3	0.73	+29 pp from two regressors

$R^2 \in [0, 1]$. Zero: model explains nothing. One: perfect fit.

R^2 always rises when you add regressors — adjusted R^2 penalizes

```r
# Full model: R^2 = 0.730
fit1 <- lm(vote ~ gdpg + inf + inc,
           data = fair)

# Add random noise
fair$noise <- rnorm(24)
fit2 <- lm(vote ~ gdpg + inf + inc
             + noise, data = fair)

summary(fit2)$r.squared      # 0.731
summary(fit2)$adj.r.squared  # 0.660
```

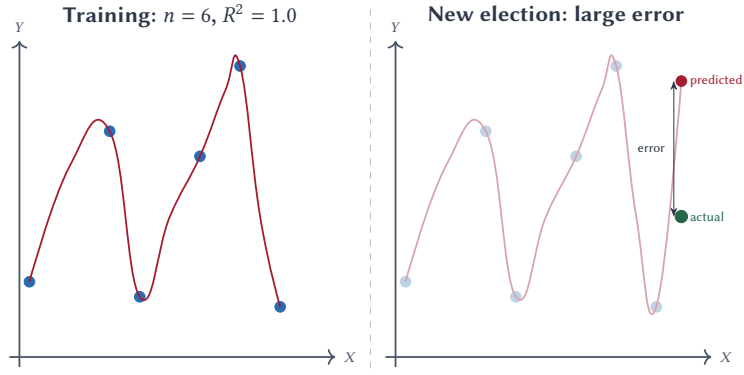
$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$$

	R^2	$\bar{R}^2$
3 regressors	0.730	0.680
+ noise	0.731	0.660

Adding noise raises R^2; $\bar{R}^2$ falls

Why: each regressor costs one degree of freedom ($n-k$ shrinks). Noise barely cuts RSS but burns that df, so $\text{RSS}/(n-k)$ rises.

Perfect R^2 on training data can predict new data badly



R^2 measures fit on the data you used. It says nothing about new data.

$R^2 = 1$ on the data you used tells you nothing about how the model performs on new data.

Part IV: Gauss–Markov

When is OLS the Best Linear Unbiased Estimator?

Four assumptions make OLS optimal

1. **Linearity:** $Y = X\beta + \varepsilon$ — linear in parameters
2. **Full rank:** $\text{rank}(X) = k$ — no perfect multicollinearity
3. **Strict exogeneity:** $\mathbb{E}[\varepsilon \mid X] = 0$ — mean-zero errors given X
4. **Spherical errors:** $\text{Var}(\varepsilon \mid X) = \sigma^2 I_n$ — homoskedastic, no autocorrelation

Under (1)–(4): OLS is **BLUE** — Best Linear Unbiased Estimator

OLS estimates the BLP regardless of A3 — strict exogeneity makes that BLP coefficient causal

Without A3 (always true):

OLS FOC: $X'\hat{\varepsilon} = 0$ algebraic

By LLN: $\hat{\beta} \xrightarrow{p} \beta_{\text{BLP}}$

$\beta_{\text{BLP}} = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$

Valid population estimand.
$\mathbb{E}[Xe] = 0$ holds by construction.
May have selection bias.

With A3: $\mathbb{E}[\varepsilon \mid X] = 0$

Makes $\beta_{\text{BLP}} = \beta_{\text{causal}}$

Also gives: $\mathbb{E}[\hat{\beta} \mid X] = \beta$
(finite-sample unbiasedness)

$\mathbb{E}[\varepsilon \mid X] = 0$ is strictly stronger than $\mathbb{E}[Xe] = 0$. A3 licenses a causal reading of $\hat{\beta}$.

Without A3: $\hat{\beta}$ estimates the BLP — descriptive, valid, possibly biased for causal effect.
With A3: $\hat{\beta}$ estimates the causal β.

Assumption 3 delivers unbiasedness: $\mathbb{E}[\hat{\beta} \mid X] = \beta$

$$\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$$

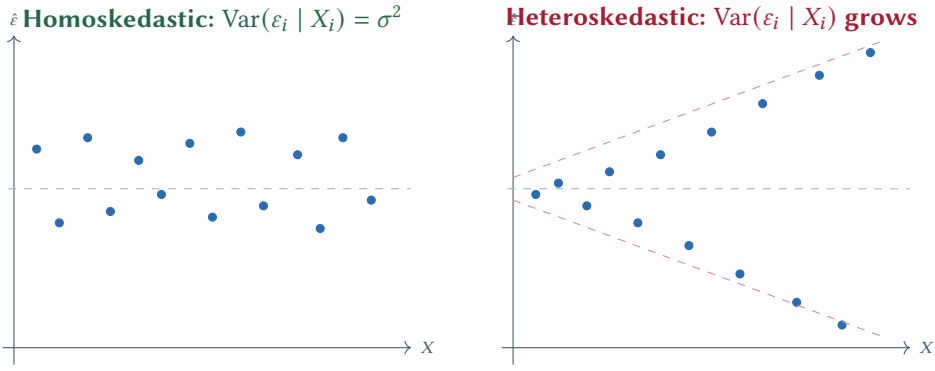
Take $\mathbb{E}[\,\cdot \mid X]$:

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X'X)^{-1}X' \underbrace{\mathbb{E}[\varepsilon \mid X]}_{=0} = \beta$$

$\mathbb{E}[\hat{\beta} \mid X] = \beta$ — OLS is **unbiased** under strict exogeneity

Strict exogeneity rules out OVB, measurement error in X, and lagged-dependent-variable bias. Contemporaneous exogeneity ($\mathbb{E}[\varepsilon_i \mid X_i] = 0$) suffices for consistency.

Assumption 4: homoskedasticity vs. the fan



$\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$: (a) homoskedasticity **and** (b) no autocorrelation

When A4 fails, OLS is still unbiased — but no longer efficient. Robust SEs fix inference. That is next lecture.

$\hat{\beta}$ varies across samples — it has a sampling distribution

Decompose:

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

Rewrite as sample means:

$$\hat{\beta} - \beta = \underbrace{\left(\frac{X'X}{n}\right)^{-1}}_{\to Q^{-1}\text{ (LLN)}} \cdot \underbrace{\frac{1}{n}\sum_i X_i\varepsilon_i}_{\to N(0,\Sigma)\text{ (CLT)}}$$

$$\hat{\beta} \approx N\left(\beta,\ Q^{-1}\Sigma Q^{-1}/n\right) \quad \text{— normal because } \hat{\beta} - \beta \text{ is linear in a sample mean}$$

Not normal because Y is normal — normal because $\hat{\beta} - \beta$ is linear in $\frac{1}{n}\sum_i X_i\varepsilon_i$. Inference next lecture.

OLS is the sample BLP — and it is optimal when the four conditions hold

$$\hat{\beta} = (X'X)^{-1}X'Y \quad \textbf{(plug-in estimator)}$$

Unbiased and BLUE under Gauss–Markov

Approximately normal because $\hat{\beta} - \beta$ is linear in $\frac{1}{n}\sum_i X_i\varepsilon_i$

Monday (12b):

Hat matrix, leverage, FWL in matrix form

Wednesday (13a):

SEs, t-tests, F-tests, robust SEs