The Best Line Through the Data

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 5

February 26 & March 3, 2026

Least Squares Was Invented to Find a Lost Asteroid

January 1801 — Piazzi discovers the asteroid Ceres. Tracks it for 40 days. Then it vanishes behind the Sun.

The problem: Where will it reappear, months later?

Gauss (age 24) fits a curve through Piazzi's 40 noisy observations by minimizing $\sum(\text{error})^2$.

December 31, 1801 — Ceres is found exactly where Gauss predicted.

The "best" curve through noisy data, chosen to *predict* where something would be. That is least squares. That is regression.

Who Deserves the Credit?

Legendre published the method first (1805).

Gauss published later (1809) — but claimed he'd been using it since 1795. His only evidence: a cryptic diary entry.

> *"There is no discovery that one cannot claim for oneself by saying that one had found the same thing some years previously."*
>
> — *Legendre, to Gauss*

Same Method, Many Purposes

Gauss (1801)	Fit a curve to predict where Ceres would appear
Us (today)	Fit a line for *description*, *prediction*, or *causal inference*

Same tool — different questions. Today we learn the mechanics.

Broockman and Kalla Ran the Real Experiment

Broockman and Kalla (2016) — the real experiment after LaCour's retraction:

- Voters randomly assigned to a conversation with a canvasser (`treated` $= 1$) or no contact (`treated` $= 0$)
- Feeling thermometer (0–100) measured *before* (`therm1`) and *after* (`therm2`)
- Higher scores = warmer feelings toward transgender people

```r
bk <- read_csv("broockman_kalla_2016.csv")
```

The Treated Group Scored 5.94 Points Higher

	Mean `therm2`	n
Control (no contact)	**60.80**	401
Treated (conversation)	**66.74**	284
Difference	**5.94**	

Remember these two numbers: **60.80** and **66.74**.

Can regression recover those exact numbers?

Write Down the Regression Equation

$$therm2_i = \beta_0 + \beta_1 \cdot treated_i + \varepsilon_i$$

Plug in the two groups:

	Equation	$\hat{Y}$
Control ($treated = 0$)	$\hat{Y} = \hat{\beta}_0$	$= 60.80$
Treated ($treated = 1$)	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$	$= 60.80 + 5.94 = 66.74$

The Coefficients Are the Means in Disguise

$$\widehat{therm2}_i = \underbrace{60.80}_{\hat{\beta}_0} + \underbrace{5.94}_{\hat{\beta}_1} \cdot treated_i$$

With a binary X, regression IS the comparison of two group means.

$$\begin{aligned} \hat{\beta}_0 &= 60.80 \text{ is the } \textbf{control group mean}. \\ \hat{\beta}_1 &= 5.94 \text{ is their } \textbf{difference}. \\ \hat{\beta}_0 + \hat{\beta}_1 &= 66.74 \text{ is the } \textbf{treated group mean}. \end{aligned}$$

R Confirms: One Line of Code, Same Numbers

```r
fit_a <- lm(therm2 ~ treated, data = bk)
summary(fit_a)
```

	Coefficient	Std. Error	t-statistic	p-value
Intercept	60.80	1.42	42.82	< 0.001
treated	5.94	2.01	2.96	0.003
$R^2 = 0.013$	$n = 685$			

Every Number in That Table Tells You Something

Coefficient	$\hat{\beta}_1 = 5.94$ — the treated group scored 5.94 points higher
Std. Error	$\text{SE} = 2.01$ — how much $\hat{\beta}_1$ would vary across samples
***t*-statistic**	$t = 5.94/2.01 = 2.96$ — $\lvert t\rvert > 1.96$, so reject H_0 at $\alpha = 0.05$
95% CI	$5.94 \pm 1.96 \times 2.01 = [2.00,\ 9.88]$ — does not contain 0
***p*-value**	$p = 0.003 < 0.05$ — probability of $\lvert t\rvert \geq 2.96$ under H_0
R^2	0.013 — treatment explains 1.3% of the variation in `therm2`

Three ways to say "statistically significant":
$\lvert t\rvert > 1.96$, CI excludes 0, $p < 0.05$.

SDM: A Unitless Ruler for Comparing Groups

One more statistic you'll use in Problem Set 2: the **standardized difference in means** (SDM).

$$\text{SDM} \quad = \quad \frac{\bar{x}_{\text{treated}} - \bar{x}_{\text{control}}}{\sqrt{(s^2_{\text{treated}} + s^2_{\text{control}})\,/\,2}}$$

Dividing by the pooled SD cancels the units — thermometers (0–100), dollars, years, all on one scale.

Randomization Worked: $|\text{SDM}| = 0.01$

	Treated	Control	SDM
Baseline `therm1`	60.22	60.43	-0.01

Rule of thumb: $|\text{SDM}| < 0.25 =$ good balance.

Where Are We?

Done:

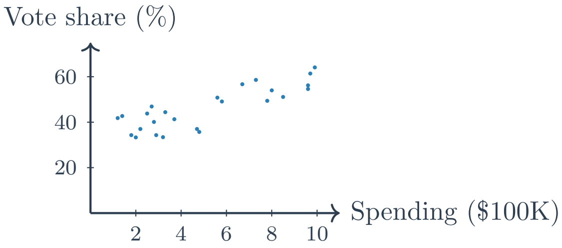
- Binary $X \Rightarrow$ regression = difference in means
- $\hat{\beta}_0$ = control mean, $\quad \hat{\beta}_1$ = difference in means
- Read every number in the `lm()` output

Next: What if X is continuous? What makes one line "better" than another?

What Makes OLS “Best”?

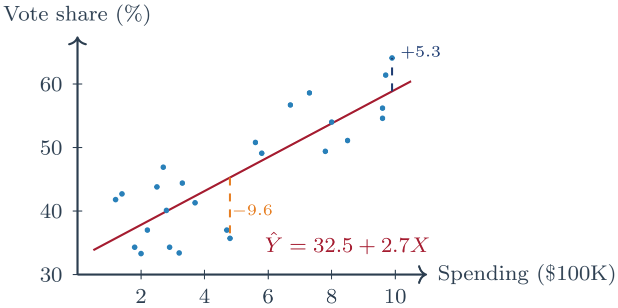
Campaign Spending and Vote Share: A Political Science Example

```r
set.seed(51); n <- 25
spending <- round(runif(n, 1, 10), 1)
vote_share <- round(30 + 3*spending + rnorm(n, 0, 5), 1)
```



What line should we draw?

The Residual Is How Far Each Point Misses the Line



Square each, sum them: $(+5.3)^2 + (-9.6)^2 + \cdots$

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 625$$

Why Square the Residuals? Because They Always Sum to Zero

For any OLS regression: $$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

WRONG: "The residuals sum to zero, so the line fits perfectly."
They *always* sum to zero. That tells you nothing about fit.

Fix: Square first, then sum. Large misses get penalized more. That gives us **SSR**.

Residuals Sum to Zero for the Same Reason Deviations from the Mean Do

Deviations from the mean:

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

OLS residuals:

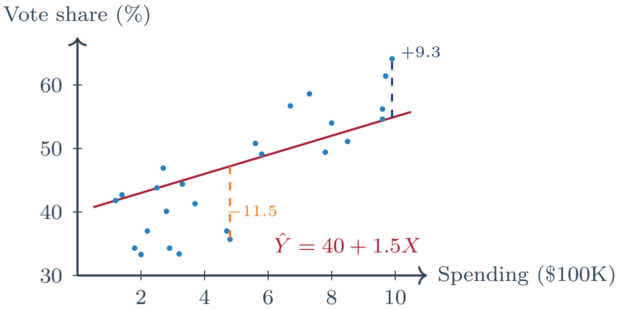
$$\boxed{\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0}$$

Why? The calculus that finds the best line forces it

Only OLS? Yes — any other line would not have this property

Connection: The mean is regression without an X: $Y_i = \beta_0 + \varepsilon_i$

A Different Line Creates Larger Residuals



$$\text{SSR} \;=\; (+9.3)^2 + (-11.5)^2 + \cdots \;=\; 969$$

$625 < 969$. Every line creates residuals — OLS finds the smallest SSR.

The Formulas Connect to What You Already Know

Slope:

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

Intercept:

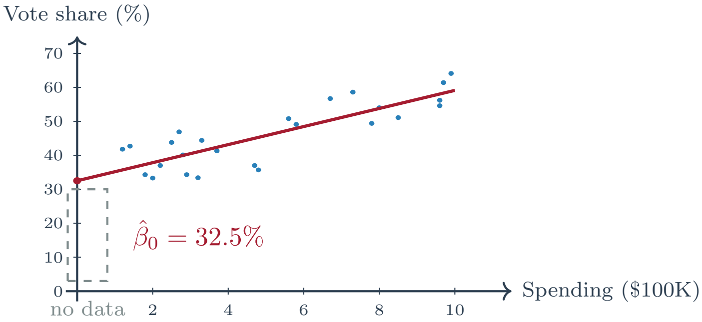
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

The regression line always passes through $(\bar{X}, \bar{Y})$.

You already know Cov and Var from Week 3 — the slope is just their ratio.

$\hat{\beta}_0$ Predicts Y at $X = 0$ — but Is That Real?

Binary X	$X = 0$ is control — $\hat{\beta}_0$ is the control mean
Spending	$X = 0$ means \$0 spent — maybe plausible
Earnings $\sim$ Age	Age $= 0$? *Extrapolation*



Thought Experiment: What If We Recentered X?

Define: $X^* = X - \bar{X}$

Run: $Y_i = \alpha + \beta X_i^* + \varepsilon_i$

Question 1: What does the intercept $\hat{\alpha}$ mean now?

Question 2: Has $\hat{\beta}$ changed?

Where Are We Now?

Done:

- Binary X: regression = difference in means (B&K)
- Continuous X: OLS minimizes SSR (campaign spending)
- Slope $= \text{Cov}(X, Y)/\text{Var}(X)$; line passes through $(\bar{X}, \bar{Y})$

Next: We got $\hat{\beta}_1 = 5.94$ with SE $= 2.01$. Can we get a more precise estimate?

Adding a Covariate

Baseline Attitudes Explain Most of the Noise

Problem: $\hat{\beta}_1 = 5.94, \quad \text{SE} = 2.01$ — but `therm2` varies hugely

Idea: Account for where each person *started*

$$\boxed{\textit{therm2}_i \;=\; \beta_0 + \beta_1 \cdot \textit{treated}_i + \beta_2 \cdot \textit{therm1}_i + \varepsilon_i}$$

The Treatment Effect Barely Changes, But Precision Jumps

	Coef.	SE	t
Reg. A: therm2 $\sim$ treated			
Intercept	60.80	1.42	42.82
treated	5.94	2.01	2.96
$R^2 = 0.013$			
Reg. B: therm2 $\sim$ treated + therm1			
Intercept	14.21	2.18	6.52
treated	5.58	1.34	4.16
therm1	0.77	0.03	24.10
$R^2 = 0.576$			

$\hat{\beta}_1$: $5.94 \rightarrow 5.58$ (barely moved). SE: $2.01 \rightarrow 1.34$ (sharper).

Total Variation Splits into Explained + Residual

$$\boxed{\text{TSS} \quad = \quad (\text{TSS} \; - \; \text{SSR}) \quad + \quad \text{SSR}}$$

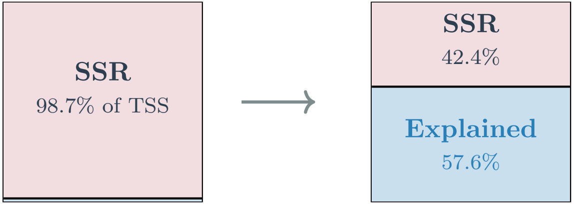
$$\textbf{SSR} = \sum (Y_i - \hat{Y}_i)^2$$

$$\textbf{Explained} = \sum (\hat{Y}_i - \bar{Y})^2$$

$$\textbf{TSS} = \sum (Y_i - \bar{Y})^2$$

R^2 = blue share of the box: $\dfrac{\text{TSS} - \text{SSR}}{\text{TSS}}$

Covariates Absorb Noise, Leaving a Cleaner Signal



Reg. A: $R^2 = 0.013$

Reg. B: $R^2 = 0.576$

Less unexplained variation $\Rightarrow$ smaller SE $\Rightarrow$ more precise $\hat{\beta}_1$

R^2 Measures How Much Variation the Model Explains

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

- **TSS** = total variation in Y (deviations from the mean)
- **SSR** = residual variation (deviations from the fitted line)
- R^2 = fraction of variation "explained" by the model

	R^2	Interpretation
Reg. A (treated only)	0.013	1.3% explained
Reg. B (treated + therm1)	0.576	57.6% explained

Most variation comes from where people *started*, not from the treatment.

A Note on Naming: Every Textbook Is Different

Source	**Residual SS**	**Total SS**
Imai (QSS)	SSR	TSS
Blackwell	SSR	TSS
Some econometrics texts	SSR	SST
Some statistics texts	SSE	SST

We use **SSR** (sum of squared residuals) and **TSS** (total sum of squares) — matching both course textbooks.

See the Decomposition in R: SSR, TSS, and R^2

```r
fit_a <- lm(therm2 ~ treated, data = bk)
fit_b <- lm(therm2 ~ treated + therm1, data = bk)

# Compute by hand
SSR_a <- sum(residuals(fit_a)^2)
SSR_b <- sum(residuals(fit_b)^2)
y     <- bk[["therm2"]]
TSS   <- sum((y - mean(y))^2)

1 - SSR_a / TSS  # 0.013 -- matches summary()
1 - SSR_b / TSS  # 0.576 -- matches summary()
```

R^2 = fraction of variation *not* left in the residuals.

Same Regression, Two Different Questions

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

Goal	**You care about**	**Why**
Prediction	$\hat{Y}_i$	Forecast new cases
Causality	$\hat{\beta}_1$	Estimate the effect

What do covariates buy you in each case?

For Causality: Covariates Buy Precision, Not a Different Answer

	Without `therm1`	With `therm1`
$\hat{\beta}_1$	5.94	5.58
SE	2.01	1.34
t	2.96	4.16

In a randomized experiment, covariates shrink the SE without changing $\hat{\beta}_1$.

For Prediction: Covariates Bring $\hat{Y}$ Closer to Y

	Without `therm1`	With `therm1`
R^2	0.013	0.576
SSR	98.7% of TSS	42.4% of TSS

Adding `therm1` explains 57.6% of the variation in `therm2`. Predictions improve.

The Full Picture

1. **Binary X:** regression = difference in means
2. **Continuous X:** OLS minimizes SSR; slope = Cov/Var
3. **Adding covariates** — same mechanism (lower SSR), two purposes:
   - **Causality:** same $\hat{\beta}_1$, smaller SE, more precision
   - **Prediction:** more explained variation, better $\hat{Y}$

Coming up:

- Prediction — overfitting and underfitting (more covariates don't always help)
- Causal inference — covariates can reduce *bias*, not just variance

Regression is a line.
OLS finds the best one.
With a binary X, it is
the difference in means.
Adding covariates sharp-
ens the estimate without
changing the answer.

Questions?