Applied Regression Interpretation

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Thurs (March 5) and Tues (March 10), 2026

The Big Picture

Women earn 84 cents for every dollar men earn

Where does that number come from?

And what does it actually mean?

Today: one dataset, four skills, building complexity

526 workers from the 1976 CPS — every number today is a fact about these people

The Current Population Survey is the gold standard for labor data

- **What:** Monthly survey of ∼60,000 U.S. households
- **Who runs it:** Bureau of Labor Statistics + Census Bureau
- **How:** Interviewers visit or call sampled households
- **Why it exists:** Official source for unemployment rate, earnings, labor force participation
- Every jobs report you see in the news comes from the CPS

Wooldridge extracted 526 workers from the 1976 CPS

- 1976: pre-Equal Pay enforcement, large raw gender gaps
- Clean, real data — every number we compute today is a fact about these 526 people

$$n = 526 \cdot 274 \text{ men, } 252 \text{ women}$$

Six variables, one dataset, the whole lecture

Variable	Type	Description
`wage`	Continuous	Hourly wage ($/hr)
`educ`	Continuous	Years of education
`exper`	Continuous	Years of work experience
`tenure`	Continuous	Years at current job
`female`	Binary (0/1)	$= 1$ if woman
`married`	Binary (0/1)	$= 1$ if married

Model 1: The Raw Gap

Model 1 starts with the population regression

$$\text{wage}_i \quad = \quad \beta_0 \; + \; \beta_1 \; \cdot \; \text{female}_i \; + \; \epsilon_i$$

- wage_i — the **outcome** (*dependent variable*): hourly wage for person i
- female_i — the **predictor** (*independent variable*): equals 1 if woman, 0 if man
- β_0, β_1 — **unknown parameters** we want to learn
- ϵ_i — the **error term**: everything else that affects wages

OLS minimizes the sum of squared residuals

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{e}_i^2 \quad = \quad \sum_{i=1}^{n} \left(Y_i \; - \; \hat{\beta}_0 \; - \; \hat{\beta}_1 X_i\right)^2$$

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the residuals as small as possible

The solution gives us fitted values for every worker

$$\widehat{\text{wage}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{female}_i$$

Plug in female $= 0$ or 1 and you get a predicted wage

Women earn $2.51 less per hour than men on average

	Model 1
Intercept ($\hat{\beta}_0$)	7.10
	(0.21)
Female ($\hat{\beta}_1$)	**−2.51**
	(0.30)
R^2	0.116
n	526

Five numbers. By the end of today, you will know what every one of them means.

What do 7.10 and −2.51 mean?

$$\widehat{\text{wage}}_i = \underbrace{7.10}_{\hat{\beta}_0} + \underbrace{(-2.51)}_{\hat{\beta}_1} \cdot \text{female}_i$$

If female = 0 (men), what is $\hat{Y}$?

If female = 1 (women), what is $\hat{Y}$?

So what is $\hat{\beta}_1$ measuring?

$\hat{\beta}_0$ is the male mean; $\hat{\beta}_1$ is the gap

- Men (female $= 0$): $\hat{Y} = 7.10 \quad \Rightarrow \quad \hat{\beta}_0 = \overline{\text{wage}}_{\text{men}}$
- Women (female $= 1$): $\hat{Y} = 7.10 - 2.51 = 4.59 \quad \Rightarrow \quad \hat{\beta}_0 + \hat{\beta}_1 = \overline{\text{wage}}_{\text{women}}$
- Subtract: $\hat{\beta}_1 = 4.59 - 7.10 = -2.51$

With binary X: intercept = baseline mean, slope = difference in means

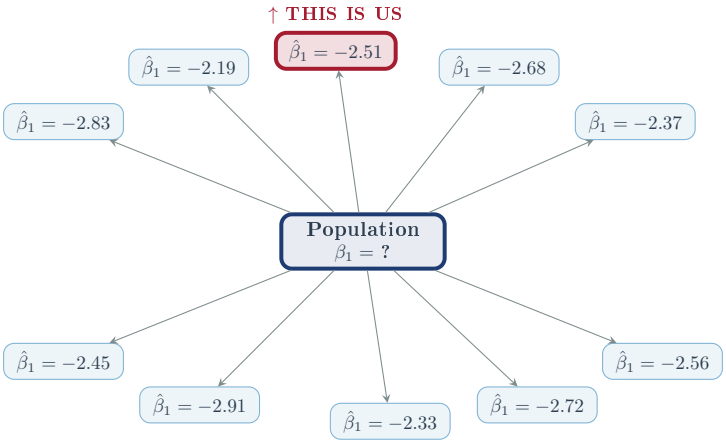
What is the (0.30) underneath the coefficient?

	Model 1
Female ($\hat{\beta}_1$)	-2.51
	(0.30) ← **what is this?**

We got -2.51 from 526 workers.

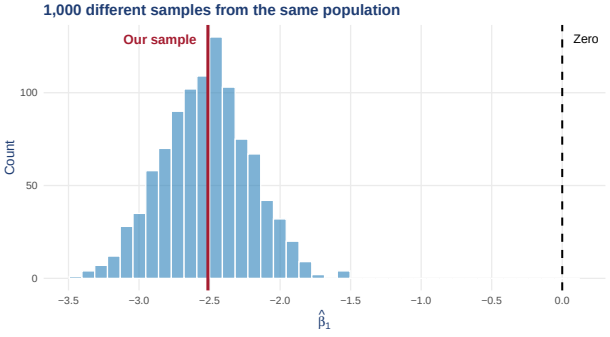
What if we had surveyed a *different* 526 workers?

Our sample is one of many we could have drawn



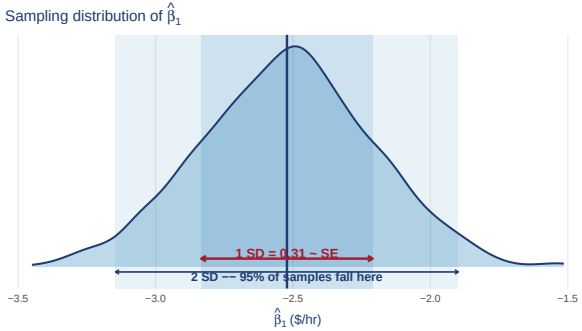
Every sample gives a different $\hat{\beta}_1$ — all are estimates of the same β_1

Here are 1,000 different samples from the same population



The **standard error** measures the width of this distribution: SE = 0.30

The SE is the standard deviation of this distribution



$$\text{SE}(\hat{\beta}_1) \quad \approx \quad \text{SD of all possible } \hat{\beta}_1\text{'s} \quad \approx \quad 0.30$$

β_1 is the truth; $\hat{\beta}_1$ is our guess from data

The population (all U.S. workers)

$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{female}_i + \epsilon_i$

- β_1 = true average wage gap
- This is what we **want to know**
- Not a causal claim — just a difference

Our sample (526 workers)

$\hat{\beta}_1 = -2.51$

- One draw from that distribution
- From *this particular* sample
- The SE tells us how much it bounces

Hats mean "estimated from this sample" — not the true population value

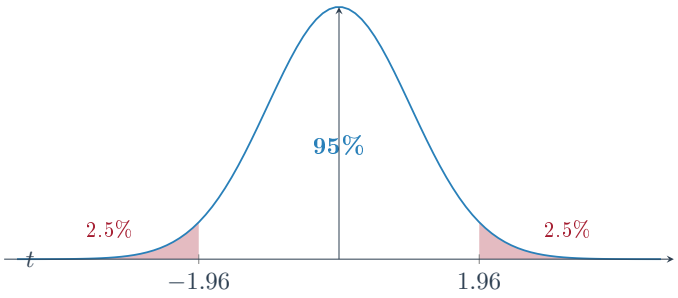
The t-statistic asks: could the true gap be zero?

- **Thought experiment:** Suppose $\beta_1 = 0$ (no real wage gap in the population)
- We got $\hat{\beta}_1 = -2.51$ — how surprising is that?
- Measure surprise in units of SE:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{-2.51}{0.30} = -8.28$$

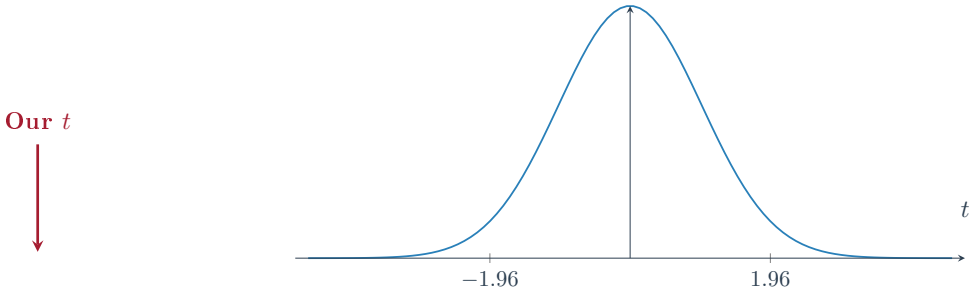
Our estimate is **8.28 standard errors** away from zero

The t-distribution shows which values are plausible under H_0



- Blue: 95% of t-statistics fall between ± 1.96 if $\beta_1 = 0$
- Red tails: only 5% land here — these are "unlikely under H_0"

Our $t = -8.28$ is not even on the chart



The p-value is the chance of landing this far out if $\beta_1 = 0$ — here it is ≈ 0

$p < 0.001$ means: if the true gap were zero, you'd essentially never see $t = -8.28$

From hypothesis test to confidence interval

- We rejected $\beta_1 = 0$ — zero is implausible
- But which values of β_1 **are** plausible?
- **Answer:** values we *cannot* reject — within 1.96 SEs of $\hat{\beta}_1$

$$95\%\ \text{CI} \;=\; \hat{\beta}_1 \pm 1.96 \times \text{SE} \;=\; -2.51 \pm 1.96(0.30)$$

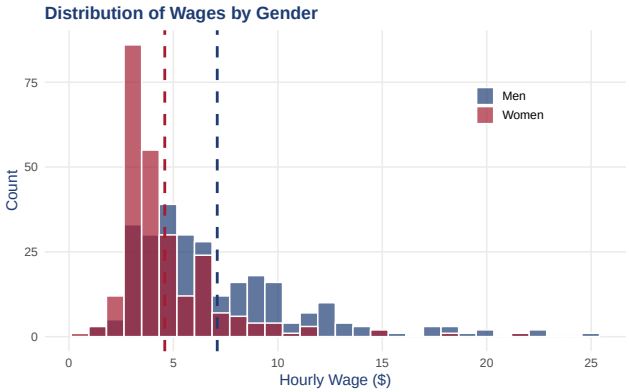
$$95\%\ \text{CI} \quad = \quad (-3.10,\ -1.92)$$

We are 95% confident the true gap is between $-\$3.10$ and $-\$1.92$

- **Correct:** If we repeated this sampling 100 times, $\sim$95 of those intervals would contain β_1
- **Wrong:** "There is a 95% chance β_1 is in this interval"
  - β_1 is fixed — it's either in there or it isn't
- Zero is **not** in $(-3.10,\ -1.92)$ — consistent with rejecting H_0

You will need **1.96** for confidence intervals on the exam

Is knowing someone's gender enough to predict their wage?



Look at the overlap. How much wage variation does gender explain?

$R^2 = 0.116$ means gender alone explains 12% of wage variation

- 88% of variation is *within* gender groups
- Some men earn very little; some women earn a lot
- Gender matters, but it is far from the whole story

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum \hat{e}_i^2}{\sum (Y_i - \bar{Y})^2} = 0.116$$

Turn to your neighbor

"The wage gap is $2.51."

What does that number **not** tell you?

Model 2: Adding Covariates

Model 2 adds education, experience, and tenure

$$\widehat{\text{wage}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{female}_i + \hat{\beta}_2 \text{educ}_i + \hat{\beta}_3 \text{exper}_i + \hat{\beta}_4 \text{tenure}_i$$

- Same outcome, same treatment variable
- Now we **hold constant** education, experience, and tenure
- This asks a *different question* than Model 1

Before we look: will the female coefficient get bigger or smaller?

Model 1: $\hat{\beta}_{\text{female}} = -2.51$

We are about to add education, experience, and tenure.

	Men	**Women**
Mean education	12.8 yrs	12.3 yrs
Mean experience	17.6 yrs	16.4 yrs
Mean tenure	6.5 yrs	3.6 yrs

Will the magnitude of $\hat{\beta}_{\text{female}}$ increase or decrease? Why?

The gender gap shrinks from $2.51 to $1.81 after adding controls

	Model 1	Model 2
Intercept	7.10	−1.57
	(0.21)	(0.72)
Female	−2.51	**−1.81**
	(0.30)	(0.26)
Education		0.57
		(0.05)
Experience		0.03
		(0.01)
Tenure		0.14
		(0.02)
R^2	0.116	**0.364**
n	526	526

Every coefficient has a "holding constant" interpretation

- **Female:** $-\$1.81$/hr, holding education, experience, and tenure constant
- **Education:** $+\$0.57$/hr per additional year, holding gender, experience, and tenure constant
- **Experience:** $+\$0.03$/hr per additional year, holding gender, education, and tenure constant
- **Tenure:** $+\$0.14$/hr per additional year, holding gender, education, and experience constant

“Holding constant” changes the question we are asking

Model 1

$\hat{\beta}_{\text{female}} = -2.51$

Gap between the *average* man and the *average* woman

Model 2

$\hat{\beta}_{\text{female}} = -1.81$

Gap between a man and woman with the *same* education, experience, and tenure

Adding covariates changes what $\hat{\beta}_1$ means

R^2 tripled — the model explains 3× more variation

- Model 1: $R^2 = 0.116$
- Model 2: $R^2 = 0.364$
- Education, experience, and tenure explain a lot of wage variation
- The SE on Female *shrank*: $0.30 \rightarrow 0.26$
  - More precise estimate when the model fits better

Turn to your neighbor: plug-in prediction

Using Model 2:

$$\hat{Y} = -1.57 + (-1.81)\text{female} + 0.57 \cdot \text{educ} + 0.03 \cdot \text{exper} + 0.14 \cdot \text{tenure}$$

Predict hourly wage for:

Profile: educ $= 16$, exper $= 10$, tenure $= 5$

1. A woman with this profile: $\hat{Y} = ?$

2. A man with this profile: $\hat{Y} = ?$

3. What is the gap?

Plug-in prediction: the gap is always $\hat{\beta}_{\text{female}}$

- **Man** (female $= 0$):
  $\hat{Y} = -1.57 + 0(-1.81) + 0.57(16) + 0.03(10) + 0.14(5) = \$8.54$
- **Woman** (female $= 1$):
  $\hat{Y} = -1.57 + 1(-1.81) + 0.57(16) + 0.03(10) + 0.14(5) = \$6.72$
- **Gap**: $6.72 - 8.54 = -\$1.81 = \hat{\beta}_{\text{female}}$

With no interaction, the gap is the same for every profile

Model 3: Interactions

Does marriage affect everyone's wages the same way?

Think about this:

- In 1976, who was more likely to be the primary earner in a married household?
- If employers perceive married men as "more stable," what happens to their wages?
- If employers perceive married women as "likely to leave," what happens?

*How would you **model** this — one slope for marriage, or two?*

Model 3 asks: does marriage affect wages differently for men and women?

$$\widehat{\text{wage}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{female}_i + \hat{\beta}_2 \text{married}_i + \hat{\beta}_3 (\text{female} \times \text{married})_i + \cdots$$

- $\hat{\beta}_2$ = effect of marriage **for men** (when female $= 0$)
- $\hat{\beta}_2 + \hat{\beta}_3$ = effect of marriage **for women**
- $\hat{\beta}_3$ = *difference* in marriage effect across genders

Marriage boosts men's wages by \$1.82 but *reduces* women's

	Model 1	Model 2	Model 3
Intercept	7.10	−1.57	−2.39
	(0.21)	(0.72)	(0.73)
Female	−2.51	−1.81	−0.31
	(0.30)	(0.26)	(0.41)
Married			1.82
			(0.40)
Female × Married			**−2.40**
			(0.53)
Education		0.57	0.55
		(0.05)	(0.05)
Experience		0.03	0.02
		(0.01)	(0.01)
Tenure		0.14	0.13
		(0.02)	(0.02)
R^2	0.116	0.364	**0.392**
n	526	526	526

The marriage premium: \$1.82 for men, −\$0.57 for women

Men (female = 0)

Marriage effect = $\hat{\beta}_2$
= +\$1.82/hr

Married men earn significantly more

Women (female = 1)

Marriage effect = $\hat{\beta}_2 + \hat{\beta}_3$
= $1.82 + (-2.40) = -\$0.57$/hr

Married women earn slightly less

$\hat{\beta}_3 = -2.40$ captures the *difference* in marriage premiums

Turn to your neighbor: 2×2 predictions

Using Model 3 with educ = 12, exper = 10, tenure = 5:

$$\hat{Y} = -2.39 + (-0.31)\text{female} + 1.82 \cdot \text{married} + (-2.40)\text{female} \times \text{married}$$
$$+0.55 \cdot \text{educ} + 0.02 \cdot \text{exper} + 0.13 \cdot \text{tenure}$$

Fill in the table:

	Single	**Married**
Men	?	?
Women	?	?

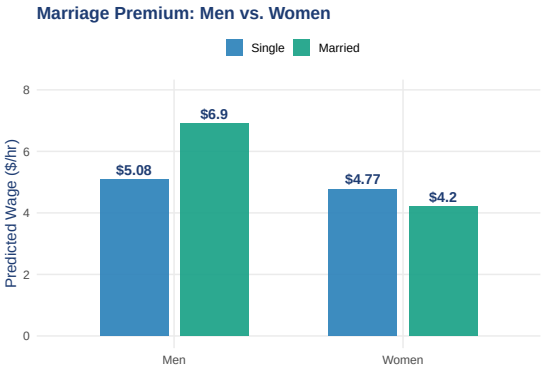
Which group earns the most? The least?

Married men earn the most; married women earn the least

	Single	Married
Men	\$5.08	\$6.90
Women	\$4.77	\$4.20

Marriage premium:

Men: +\$1.82 Women: −\$0.57



Putting It All Together

The gap shrinks as you add controls:

$-\$2.51 \rightarrow -\$1.81 \rightarrow -\$0.31$

Adding controls changes how large the gap is — and what it *means*

Each model answers a different question

Model	What $\hat{\beta}_{\text{female}}$ measures
1	Raw gap: average man vs. average woman
2	Adjusted gap: same education, experience, tenure
3	Gap for single workers with same human capital

The "wage gap" depends on which question you ask

Polynomials

The 30th year of experience matters less than the 5th

$$\widehat{\text{wage}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{female}_i + \hat{\beta}_2 \text{educ}_i + \hat{\beta}_3 \text{exper}_i + \hat{\beta}_4 \text{exper}_i^2 + \hat{\beta}_5 \text{tenure}_i$$

- $\hat{\beta}_3 = 0.205, \quad \hat{\beta}_4 = -0.004$
- Negative $\hat{\beta}_4$: each additional year adds *less* than the last

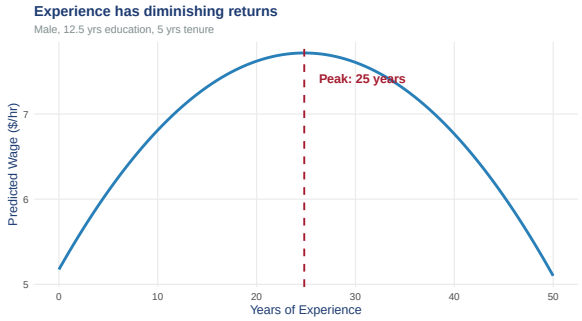
R: `lm(wage ~ female + educ + exper + I(exper^2) + tenure)`

The marginal effect depends on where you start

$$\frac{\partial \widehat{\text{wage}}}{\partial \text{exper}} = \hat{\beta}_3 + 2\hat{\beta}_4 \cdot \text{exper} = 0.205 - 0.008 \cdot \text{exper}$$

Experience	Marginal effect (\$/hr)
5 years	+\$0.16
15 years	+\$0.08
25 years	$\approx$ \$0.00
35 years	−\$0.08

Wages peak at about 25 years of experience



$$\text{exper}^* = \frac{-\hat{\beta}_3}{2\hat{\beta}_4} = \frac{-0.205}{2(-0.004)} \approx 25 \text{ years}$$

Quadratics are still OLS — exper^2 is just another column

	X_1	X_2	X_3	$\cdots$
Worker 1	5	25	12	
Worker 2	20	400	16	
Worker 3	8	64	10	
	exper	exper^2	educ	

exper^2 is a **mechanical transformation** — OLS doesn't know it's a square

Five Ways to Read a Coefficient

The interpretation depends on what X and Y are

Same $\hat{\beta}_1$ formula. Five different English sentences.

Y **can be:**	continuous (wage) or binary (highwage) or logged
X **can be:**	continuous (educ) or binary (female, college)

The units of $\hat{\beta}_1$ change with the units of X and Y

Continuous → Continuous: \$0.54/hr per year of education

$$\widehat{\text{wage}}_i = -0.90 + 0.54 \cdot \text{educ}_i$$

$$\hat{\beta}_1 = 0.54$$

One more year of education → +\$0.54/hr

Units of $\hat{\beta}_1$: dollars per year

Continuous → Log: +8.3% per year of education

$$\widehat{\log(\text{wage})}_i = 0.58 + 0.083 \cdot \text{educ}_i$$

$$\hat{\beta}_1 = 0.083$$

One more year of education → +8.3% higher wages

- ▷ **Rule:** $\hat{\beta}_1 \times 100 =$ percent change in Y
- ▷ Works well when $\hat{\beta}_1$ is small

Binary → Continuous: women earn $2.51 less per hour

$$\widehat{\text{wage}}_i = 7.10 + (-2.51) \cdot \text{female}_i$$

$$\hat{\beta}_1 = -2.51$$

$$\hat{\beta}_1 = \bar{Y}_{\text{women}} - \bar{Y}_{\text{men}} = \text{difference in group means}$$

You already know this one from Model 1

Binary → Log: women earn ∼40% less

$$\widehat{\log(\text{wage})}_i = 1.81 + (-0.40) \cdot \text{female}_i$$

$$\hat{\beta}_1 = -0.40$$

Approximate: $\hat{\beta}_1 \times 100 = -40\%$

Exact: $(e^{\hat{\beta}_1} - 1) \times 100 = -32.8\%$

When $|\hat{\beta}_1|$ is large, the approximation overshoots

Binary → Binary: college grads are 42pp more likely to earn above median

- **Define:** `highwage` $= 1$ if wage $>$ median; `college` $= 1$ if educ ≥ 16

$$\widehat{\text{highwage}}_i = 0.35 + 0.42 \cdot \text{college}_i$$

$$\hat{\beta}_1 = 0.42$$

College graduates are **42 percentage points** more likely to earn above median

This is the **Linear Probability Model** (LPM): OLS with a binary Y

Five specifications, five interpretations

Y	**X**	$\hat{\beta}_1$	**Interpretation**
wage	educ (cont.)	0.54	+\$0.54/hr per year
log(wage)	educ (cont.)	0.083	+8.3% per year
wage	female (binary)	−2.51	Women earn \$2.51 less
log(wage)	female (binary)	−0.40	Women earn ∼40% less
highwage	college (binary)	0.42	+42 pp more likely

Same formula. Same estimator. The **units** change with X and Y.

Turn to your neighbor: interpret every coefficient

Full model:

$$\widehat{\log(\text{wage})}_i = 0.42 - 0.30\,\text{female}_i + 0.088\,\text{educ}_i + 0.005\,\text{exper}_i + 0.017\,\text{tenure}_i$$

1. Female: -0.30 means ...
2. Education: 0.088 means ...
3. Experience: 0.005 means ...
4. Tenure: 0.017 means ...

Hint: Y is logged, so $\hat{\beta} \times 100 =$ percent change

Prediction and Overfitting

Today's plan

We are not doing causal inference right now

No treatments, no counterfactuals — just: can we predict Y well?

1. **OLS prediction basics** — covariates, polynomials, R^2
2. **Recognizing overfitting** — when "better fit" makes predictions worse
3. **Train/test splits** — honest evaluation of prediction accuracy (review from earlier)
4. **After spring break:** machine learning methods designed for prediction

People predict things all the time

High stakes

- Will this patient respond to treatment?
- Which counties will swing in the next election?
- Where will the next violent crime occur?
- How many ER visits will flu season bring?
- What will GDP growth be next quarter?

Everyday stakes

- Who should I start in fantasy football?
- Will this movie bomb at the box office?
- How much should I list my house for?
- Which team covers the spread Saturday?
- How long will my commute take tomorrow?

The common thread: predicting what has not yet happened

All of them need to predict **outcomes they have not yet observed**

Not "what happened?" but "what will happen?"

You already know a prediction machine: OLS

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

- The $\hat{\beta}$'s describe patterns *in the sample*
- But OLS draws a **line** — and lines extend beyond the data
- Plug in *any* values of X and you get a $\hat{Y}$, even for combinations nobody in the data has

That is prediction: estimating Y for observations the model has never seen

The intercept is already a prediction

$$\hat{\beta}_0 = \text{predicted } Y \text{ when } \textbf{every } X = 0$$

Model	Intercept meaning
Wages $\sim$ Female	Mean wages for men (the $X = 0$ group) — sensible
Wages $\sim$ Female + Exper	Mean wages for men with zero experience — plausible
Wages $\sim$ Female + Exper + Age	Mean wages for a zero-year-old man with zero experience — impossible

The intercept predicts "leftward" to the Y-axis
— even when no one in the data lives there

Description vs. prediction

Description

Goal: patterns *in* the data

X**'s:** observed in the sample

$\hat{Y}$**:** summarizes what we have

Prediction

Goal: Y for *new* observations

X**'s:** may be unseen combinations

$\hat{Y}$**:** a bet on the unseen

Same model, same $\hat{\beta}$'s — the difference is whether the X's are in-sample or out-of-sample

Ames, Iowa: 2,930 houses, 74 variables, one goal

Can we predict a house's sale price?

Outcome:	Sale price (\$12K–\$755K)
Predictors:	Square footage, year built, garage size, lot area, bedrooms, bathrooms, pool, ...
Question:	How many variables should we use?

R^2 looks great when you add more variables

	1 var	4 vars	20 vars	259 vars
Train R^2	0.52	0.83	0.86	**0.92**

R^2 **always** goes up (or stays the same) when you add variables

So why not add *everything*?

With enough variables, $R^2 = 1.00$ is guaranteed

Variables (k)	Observations (n)	Train R^2
1	2,051	0.52
4	2,051	0.83
20	2,051	0.86
259	2,051	0.92
$k \geq n$	2,051	**1.00**

If $k \geq n$, OLS can **perfectly** fit every data point.
$R^2 = 1.00$ is not learning — it is memorization.

Think about this

You build a model with 2,051 variables on 2,051 houses.

$R^2 = 1.00$. Every prediction is perfect.

Your friend lists a new house for sale.

How confident are you in your prediction of its price?

R^2 only measures how well you fit THIS dataset

- R^2 rewards **memorization**
- A model with 259 terms can memorize patterns in the noise
- **The real question:** can the model predict *new* houses it hasn't seen?

A perfect fit on your data might be a terrible fit on new data

On new data, R^2 can go *negative*

Imagine: you fit a model on your dataset, then a *new* batch of houses arrives

- **On the original data:** $R^2 \geq 0$ always — OLS guarantees SSR $\leq$ SST
- **On new data:** if predictions are worse than guessing $\bar{Y}$, then SSR $>$ SST

$$R^2_{\text{new}} = 1 - \frac{\text{SSR}_{\text{new}}}{\text{SST}_{\text{new}}} < 0 \quad \text{when SSR} > \text{SST}$$

But is $R^2_{new} < 0$ possible? Both SSR and SST are sums of squares — both positive. What would make the ratio exceed 1?

Yes — here it is, with our Ames data

630-coefficient model (all pairwise interactions) trained on 2,051 houses, then applied to 879 *different* houses

$$\begin{aligned}
\text{SSR}_{\text{new}} &= \textstyle\sum_{i\in\text{new}}(Y_i - \hat{Y}_i)^2 \;\approx\; 32{,}648 \text{ billion} \\
\text{SST}_{\text{new}} &= \textstyle\sum_{i\in\text{new}}(Y_i - \bar{Y})^2 \;\approx\; 4{,}973 \text{ billion} \\
\frac{\text{SSR}}{\text{SST}} &= \frac{32{,}648}{4{,}973} = 6.57 \\
R^2_{\text{new}} &= 1 - 6.57 = \mathbf{-5.57}
\end{aligned}$$

The model's squared errors are **6.5× larger** than just guessing \$181,515 for every house

The fix: train on some data, test on the rest

- **Step 1:** Split data — **training set** (70%) and **test set** (30%)
- **Step 2:** Fit the model on training data only
- **Step 3:** Predict on test data the model has *never seen*
- **Step 4:** Measure prediction error

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \left(Y_i - \hat{Y}_i\right)^2}$$

RMSE = Root Mean Squared Error — same units as Y (dollars)

Adding good predictors helps everywhere

	1 variable	4 variables
Train RMSE	\$57,000	\$34,000
Test RMSE	\$56,000	\$36,000

Both train *and* test error improve
— the model learns real patterns

R^2 rising = better fit. RMSE falling = better fit. Two ways to say the same thing.

Throwing in everything: test predictions go bananas

	1 var	4 vars	20 vars	259 vars
Train R^2	0.52	0.83	0.86	0.92
Test R^2	0.45	0.77	**0.81**	0.56
Train RMSE	\$57K	\$34K	\$30K	\$23K
Test RMSE	\$56K	\$36K	\$33K	**\$50K**

Train RMSE keeps falling. Test RMSE **rises** for the 259-variable model.

What does $50K of prediction error mean in practice?

- **The 259-variable model** promised $23K average error (training RMSE)
- **On new houses,** the actual error is $50K — more than double
- A buyer relying on this model could overpay by $50,000
- A seller could underprice by $50,000

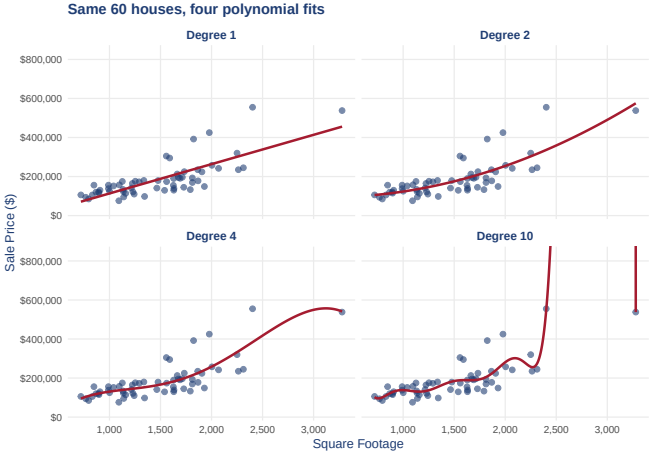
The model looked great on paper. It failed when it mattered.

Why does this happen? The model memorized noise.

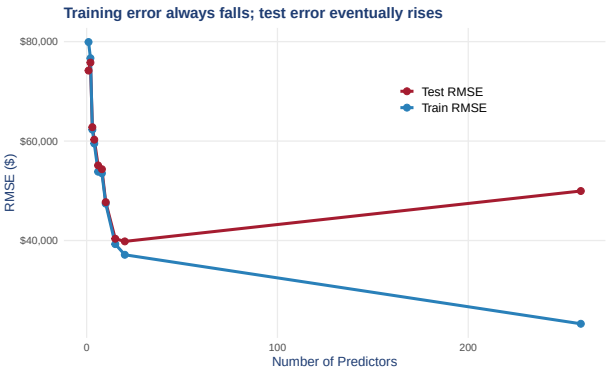
- With 259 variables and 2,051 houses, the model has **one parameter for every 8 observations**
- It learns real patterns (square footage matters) *and* fake patterns (noise in the training data)
- On new data, the real patterns generalize — the fake ones do not

Overfitting = fitting signal *and* noise, instead of just signal

Degree 10 chases individual houses instead of learning the pattern



Training error always falls; test error eventually rises



A model that memorizes training data **fails** on new data

More variables $\neq$ better model

- **Prediction:** kitchen-sink models memorize noise $\rightarrow$ terrible forecasts
- **Causal inference:** throwing in every control variable can amplify bias
- **Lesson:** choose predictors with purpose, not volume

More variables $\neq$ better model — this is true for prediction *and* causal inference

Can we find the *optimal* prediction model?

What we know so far:

- Too few variables → underfitting (misses real patterns)
- Too many variables → overfitting (memorizes noise)
- Test RMSE reveals the sweet spot — but only for models we tried

Open question: is there a principled way to *search* for the best model, instead of guessing?

Why $k \geq n$ breaks OLS: the algebra

OLS requires inverting $(X'X)$:

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

- X is $n \times k$ — when $k \geq n$, the columns are linearly dependent
- $(X'X)$ becomes **singular** — no unique inverse exists
- Infinitely many $\hat{\boldsymbol{\beta}}$ solve the normal equations

With $k \geq n$, OLS has **no unique solution**. The model can perfectly fit the data in infinitely many ways.

Coming soon: penalized regression & cross-validation

The machine learning solution:

1. **Penalize complexity** — add a cost for large coefficients (Ridge, LASSO)
   Forces the model to keep only variables that earn their place
2. **Cross-validation** — systematically rotate which data is "train" vs. "test"
   Finds the penalty strength that minimizes out-of-sample error
3. **Result:** a principled, automated search for the best-predicting model

Not today — but this is where we're headed next

Exam Preparation

Five skills to master before the exam

1. **Interpret coefficients in context**
2. **Compute $\hat{Y}$ by plug-in**
3. **Compare models side-by-side**
4. **Read interactions and build 2×2 tables**
5. **Recognize overfitting**

Study the review guide — the exam comes from it

- Not everything on the review will appear
- But the exam is drawn *from* the review

Practice plug-in predictions, interpret every number in a regression table, build 2×2 tables

- **Cheat sheet:** 2 pages, front and back, handwritten or typed
- **Exam 1:** Thursday, March 12 — 75 minutes

Three papers you are responsible for

1. **Card et al.** — anti-immigration speeches and media framing
2. **Broockman, Kalla & Aronow (2015)** — irregularities in the LaCour study
3. **Broockman & Kalla (2016)** — transgender rights canvassing experiment

I reserve the right to ask you **anything** about these three papers

We spent two full lectures on these — use your lecture notes and slides to review

Every number in a regression table has a precise English meaning — your job is to find it