Overfitting: When Good Models Go Bad

Gov 51 Section — Week 8

George
Harvard University

March 26, 2026

Can you predict house prices better than a straight line?

2,930 houses in Ames, Iowa. 80 features. One price tag.

You have square footage, bedrooms, year built, garage size, lot area, fireplaces... How many variables should you use?

More variables = better model? Let's find out.

The Overfitting Problem

Overfitting means memorizing noise instead of learning signal

Underfitting

Model is too simple
Misses real patterns
High bias, low variance

A flat line through a curved relationship

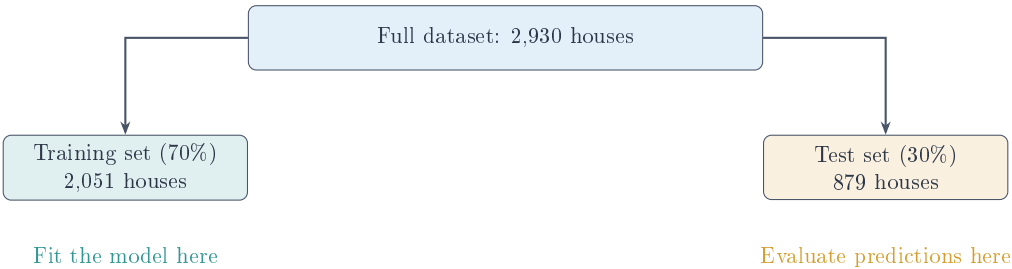
Overfitting

Model is too flexible
Memorizes quirks of this sample
Low bias, high variance

A wiggly curve that hits every training point

The goal is the sweet spot: complex enough to capture real patterns, simple enough to generalize

The fundamental tension: in-sample fit vs. out-of-sample prediction



The model never sees the test set — so RMSE there is an honest measure of prediction quality

RMSE measures how far off your predictions are, in real units

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- $Y_i - \hat{Y}_i$ = prediction error for observation i
- Square, average, take the root $\rightarrow$ back in original units
- **RMSE = \$40,000** means predictions are off by about \$40K on average

Lower RMSE = better predictions. Compute it on the **test set** for an honest answer.

Polynomial Fits: Seeing Overfitting

Same 60 houses, four polynomial fits

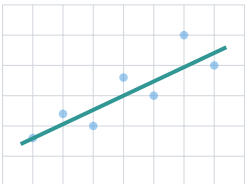
Predicting sale price from square footage alone

	Degree 1	**Degree 2**	**Degree 4**	**Degree 10**
Shape	straight line	gentle curve	wiggly	very wiggly
Train R^2	0.58	0.63	0.65	0.74

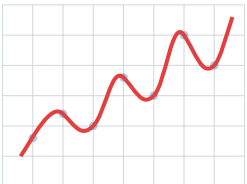
Train R^2 climbs from 0.58 to 0.74 —
but is Degree 10 really a better model?

Higher degree = the curve chases individual points

Degree 1



Degree 10



Degree 1 captures the trend. Degree 10 memorizes each data point's quirks.

If train R^2 always goes up with more flexibility, how do we know when we've gone too far?

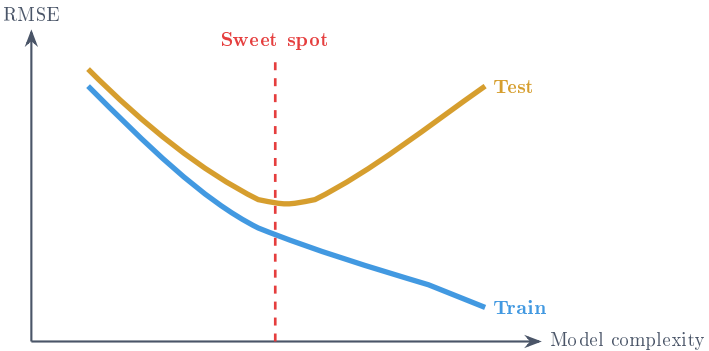
More Variables, More Problems?

Adding variables: train RMSE falls, but test RMSE eventually rises

	1 var	4 vars	20 vars	259 vars
Train RMSE	\$56,900	\$33,600	\$30,100	\$23,300
Test RMSE	\$55,700	\$36,000	\$33,200	\$50,000
Gap	small	\$2,400	\$3,100	\$26,700

259 variables: train RMSE keeps falling (\$23K), but test RMSE jumps back up (\$50K) — worse than using just 1 variable!

The overfitting curve has a characteristic shape



Train error always falls. Test error falls then rises. The gap between them is overfitting.

Your Turn: Diagnosing Overfitting

Which model is overfitting?

Scenario: predicting congressional vote share

	Model A	**Model B**
Variables	5	150
Train RMSE	6.2	1.8
Test RMSE	6.5	14.3
Train R^2	0.61	0.97
Test R^2	0.58	-0.42

Discuss with your neighbor:

1. Which model is overfitting and how can you tell?
2. What does negative R^2 mean in practice?
3. Which model would you use to predict next year's elections?

Model B is overfitting — the gap between train and test is the clue

Model A: healthy

Train RMSE: 6.2
Test RMSE: 6.5
Gap: 0.3

Small gap ⇒ model generalizes well

Model B: overfit

Train RMSE: 1.8
Test RMSE: 14.3
Gap: 12.5

Huge gap ⇒ memorized the training data

Negative R^2 means predictions are worse than just guessing the average — the model is actively harmful

Three warning signs of overfitting

1. **Train error much lower than test error**
   A small gap is normal. A large gap is trouble.

2. **Negative R^2 on test data**
   Your model is worse than just predicting $\bar{Y}$ for everyone.

3. **Coefficients that are enormous or change sign**
   When the model is desperate to fit noise, coefficients go wild.

The Fix: Penalize Complexity

Overfitting is the diagnosis; regularization is the treatment

OLS objective:

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

Penalized regression adds a cost for large coefficients:

$$\min_{\beta} \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{\text{fit the data}} + \underbrace{\lambda \cdot \text{Penalty}(\beta)}_{\text{stay simple}}$$

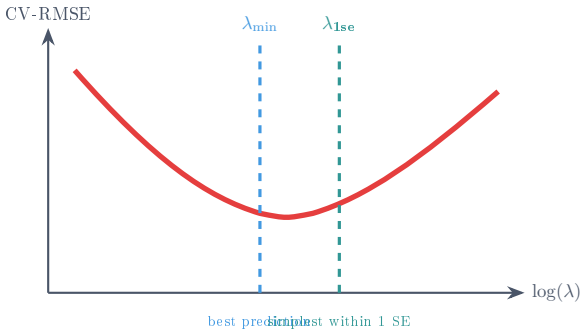
- $\lambda = 0$: no penalty $\rightarrow$ OLS (can overfit)
- λ large: heavy penalty $\rightarrow$ coefficients shrink toward zero

LASSO, Ridge, and Elastic Net differ in how they penalize

Method	Penalty	Effect
Ridge	$\lambda \sum \beta_j^2$	Shrinks all, keeps everything
LASSO	$\lambda \sum \lvert\beta_j\rvert$	Shrinks some to **exactly zero**
Elastic Net	Both combined	Shrinks and selects

Same regression you already know + a penalty term

Cross-validation picks the right amount of penalty



- λ_{min}: minimizes CV error — best prediction
- $\lambda_{\mathbf{1se}}$: simplest model within 1 SE of minimum — more parsimonious

Penalized methods beat OLS when there are many predictors

Ames Housing: predicting sale price with 34 numeric variables

Method	Test RMSE	Variables used
OLS	\$37,900	all 34
Ridge	\$37,100	all 34 (shrunk)
LASSO	\$37,100	28 of 34
Elastic Net	\$36,700	21 of 34

All three penalized methods outperform OLS — and LASSO/Elastic Net tell you which variables matter

Project Milestone 1
Due Thursday, March 26 by 11:59 pm

Prediction projects help real people make better decisions

Application	Why it matters
Bail prediction	Algorithms reduce crime 25% at same jailing rate (Kleinberg et al.)
Conflict forecasting	Better forecasts → earlier intervention
Election prediction	Campaigns allocate resources where races are closest
Recidivism (COMPAS)	Risk scores used in courts — but are they fair?

Prediction isn't about understanding *why* — it's about getting decisions right

Milestone 1: pick a question, find your data

Submit a 1–2 page proposal with four things:

1. **Research question** — one clear sentence
2. **Study type** — descriptive, predictive, or causal?
3. **Dataset** — source, observations, key variables
4. **Brief plan** — 2–3 sentences on your analysis approach

Still deciding between two questions? Submit both and ask for feedback.

Three study types — know which bucket you are in

Descriptive

What does the world look like?

"How has the racial wealth gap changed since 1990?"

Summary stats, visualizations, text analysis

Predictive

Can we forecast an outcome?

"Can we predict which defendants will be rearrested?"

Train/test splits, RMSE, LASSO

Causal

Does X cause Y?

"Does minimum wage reduce employment?"

Experiments, DiD, IV (coming soon)

Your study type determines your methods, your success metric, and how you interpret results

Your turn: find your question in 2 minutes

Think of one thing about politics, society, or economics that bugs you.

1. **Write it down** (30 seconds)
2. **Turn it into a one-sentence question** (30 seconds)
   Not "inequality" — instead: "Has the racial wealth gap widened since 2008?"
3. **Classify it:** descriptive, predictive, or causal? (30 seconds)
4. **Share with your neighbor** (30 seconds)

Eight sources cover most undergraduate projects

Source	Best for
IPUMS	Census, demographics, income, housing
ANES	Voting behavior, political attitudes
GSS	Social attitudes over time
ICPSR / openICPSR	16,000+ datasets + AEA replication packages
ProPublica	Criminal justice, investigations
FiveThirtyEight	Politics, sports, culture (clean, small)
Opportunity Insights	**Income mobility by neighborhood (free CSVs)**

Load your data into R **before** you submit. Make sure it works.

Replication data is fair game — but the question must be yours

- **Opportunity Insights** (opportunityinsights.org/data)
  Chetty et al.: tract-level mobility, earnings, patents — free CSVs, no registration
- **Abramitzky & Boustan** (openICPSR project 120490)
  Immigrant mobility — connects to Card et al. speeches we read in class
- **Any AEA replication package** (openicpsr.org)
  Thousands of datasets from published economics papers — all free

Rule: if you use replication data, your question must be new — you are not replicating, you are investigating something the authors did not

Overfitting is the core problem of prediction — next week we learn how LASSO and Ridge solve it systematically. Project proposals due Thursday, March 26 by 11:59 pm.