Regularization

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 9
Starting March 24, 2026

Exams are graded; today we start prediction and ML

- Exams graded and being uploaded as we speak
- Today: **prediction and machine learning**
- **Project Proposal (Milestone 1) due Thursday, March 26 by 11:59 pm**

Your Project:
Finding a Question, Finding Data

The project is where you use everything we've built

- Every problem set has been practice
- The project is the game
- Pick something you're genuinely curious about
- It doesn't need to be groundbreaking — it needs to be **careful, honest, and interesting to you**

Milestone 1 is a 1–2 page proposal: research question, study type, dataset, and brief plan

Every project falls into one of three buckets

Descriptive

What does the world look like?

Summary stats, visualizations, text analysis, mapping

Card et al. immigration speeches — that was a descriptive project

Predictive

Can we forecast an outcome?

Train/test splits, RMSE, LASSO — exactly what we learn this week

COMPAS recidivism prediction is this type

Causal

Does X cause Y?

Experiments, DiD, IV — we'll cover these in Weeks 11–12

Broockman & Kalla was a causal study

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

Good descriptive projects

- How has the racial wealth gap in the U.S. changed since 1990?
- What topics dominate congressional floor speeches about immigration?
- How do commute times vary across metropolitan areas?
- What share of defendants in Broward County are rearrested within two years?

Descriptive work is underrated. Good description is hard, and it's the foundation for everything else.

Good predictive projects

- Can we predict which defendants will be rearrested?
- Can we forecast which countries will experience civil war?
- Can we predict election outcomes from polling data?
- Can we identify fraudulent insurance claims from claim characteristics?

The key metric is **out-of-sample performance** — an overfit model that looks great on training data but fails on new data is worthless

Good causal projects

- Does increasing the minimum wage reduce employment?
- Does body camera adoption reduce police use of force?
- Does access to early childhood education improve long-term earnings?
- Did a specific policy change affect voter turnout?

You need a credible identification strategy — why is your comparison valid? We'll cover the methods in Weeks 11–12.

Good questions start with what bugs you

1. **Start with what bugs you**
   Housing costs frustrate you → "How have rents changed relative to wages?"
2. **Read the news with a social scientist's eye**
   "Crime is surging" — compared to what? That "compared to what?" is where questions live
3. **Narrow it down**
   "Inequality" is not a question. "Has the Gini coefficient changed differently in Medicaid-expansion vs. non-expansion states?" is.
4. **Start with the data**
   Find an interesting dataset, then ask what questions it can answer. No shame in that.

Eight sources cover most undergraduate projects

Source	Best for
IPUMS (ipums.org)	Census, ACS, demographics, income, housing
ANES (electionstudies.org)	Voting behavior, political attitudes
GSS (gss.norc.org)	Social attitudes since 1972
ICPSR / openICPSR	16,000+ datasets + AEA replication packages
Harvard Dataverse	Replication data from published papers
ProPublica	Criminal justice, investigations
FiveThirtyEight GitHub	Politics, sports (clean, small)
Opportunity Insights	**Income mobility by neighborhood (free CSVs)**

You can also collect your own: web scraping, text data, surveys, or LLM classification

Published studies post their data — use it, but ask a new question

- **Opportunity Insights** (opportunityinsights.org/data)
  Chetty et al.: tract-level mobility, earnings, incarceration, patents by race/income. Free CSV download, no registration.
- **Abramitzky & Boustan** (openICPSR project 120490)
  Immigrant mobility across two centuries — connects to Card et al. speeches we read
- **AEA replication packages** (openicpsr.org)
  Thousands of datasets from published economics papers — all free

Rule: if you use replication data, your question must be new — you're not replicating, you're investigating something the authors didn't

Load your data before you submit

- **Load your data into R before you submit**
  Make sure the file reads, the key variables exist, and there's enough variation
- **State your question in one sentence**
  If you can't, it's too big
- **Deciding between two questions? Submit both**
  That's what Milestone 1 is for — we'll give you feedback
- **It doesn't need to be ambitious**
  A well-executed descriptive analysis of something you care about is a great project

Due Wednesday, March 26 — 1–2 pages

Now: Prediction and Machine Learning

An algorithm could reduce crime by 25% — using math you already know

24.7%

reduction in crime with no change in jailing rates

Kleinberg, Lakkaraju, Lestre, Ludwig & Mullainathan (2018, *QJE*)

Judges make predictions every day — and the math can help

At every bail hearing, the judge predicts:

- Will this defendant show up for court?
- Will they commit a new crime before trial?

Kleinberg et al. results:

- Reduce crime by 24.7% at same jailing rate
- **Or** reduce jailing by 42% at same crime rate
- **Tools:** OLS, LASSO, Ridge — what we learn this week

Two experts, same data, opposite conclusions

Halliburton Co. v. Erica P. John Fund — \$5 billion securities case.

	Defense Expert	**Plaintiff Expert**
Peer selection method	S&P Energy + custom index	Analyst-report peers
Excess return	-2.9%	-3.7%
p-value	0.20	0.02
Conclusion	Not significant	Significant

Key problem: different peer choices $\rightarrow$ opposite answers

How do we build models where the data picks the variables, not the analyst?

The Bias-Variance Tradeoff

Adding predictors improves fit but worsens prediction

Before spring break:

- More predictors $\rightarrow$ better in-sample fit
- More predictors $\rightarrow$ worse out-of-sample predictions
- R^2 always goes up when you add variables — even useless ones

How do we build models that predict *new* data well, not just the data we already have?

In Week 6 we split the data once — today we generalize

Week 6: Single split

- 80% train, 20% test
- Fit on train, evaluate on test
- One RMSE estimate

Problem: result depends on which 20% you held out

Today: Cross-validation

- Split into k folds
- Each fold takes a turn as the test set
- Average k RMSE estimates

More stable estimate of out-of-sample performance

Same idea as Week 6 — but repeated k times for a more reliable answer

What makes a prediction "good"?

MSE measures total prediction error

For a single prediction:

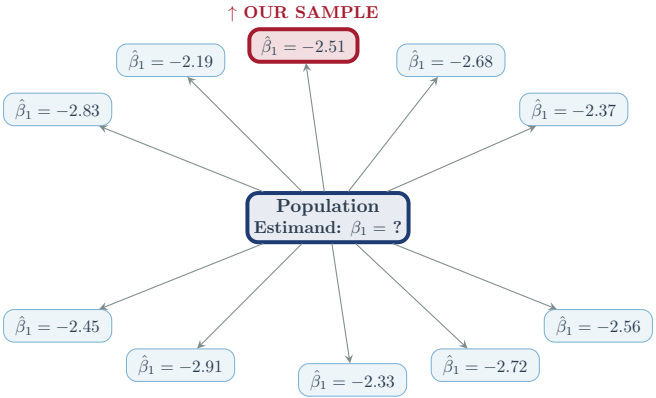
$$\text{Squared Error}_i = (\hat{Y}_i - Y_i)^2$$

Average over all predictions:

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

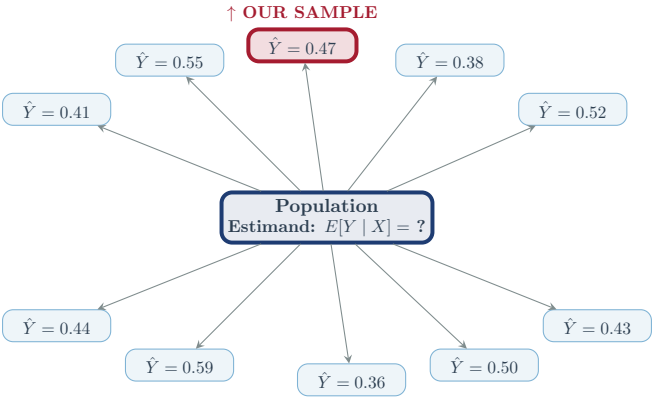
RMSE $= \sqrt{\text{MSE}}$ (back in original units)

Remember this? $\hat{\beta}_1$ is an estimator, β_1 is the estimand



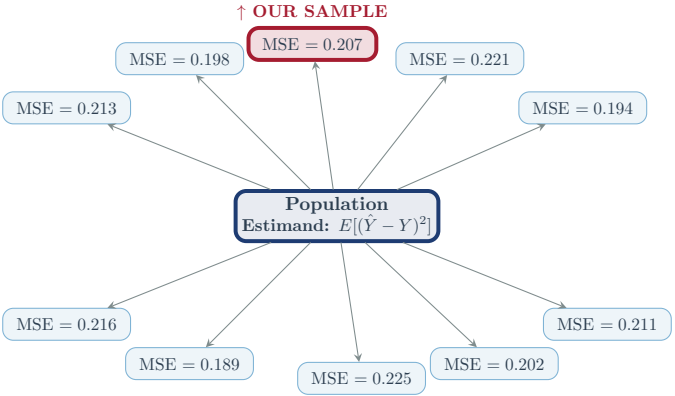
Each spoke is an **estimator**. The hub is the **estimand**.

Each sample also gives a different prediction $\hat{Y}$



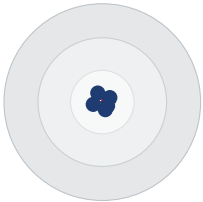
Variance = how spread out the $\hat{Y}$'s are across samples

Our sample MSE is an estimator of the population MSE

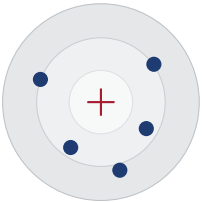


Same logic as $\bar{X}$ estimating $E[X]$. The population MSE is what we decompose.

Bias means consistently wrong; variance means scattered



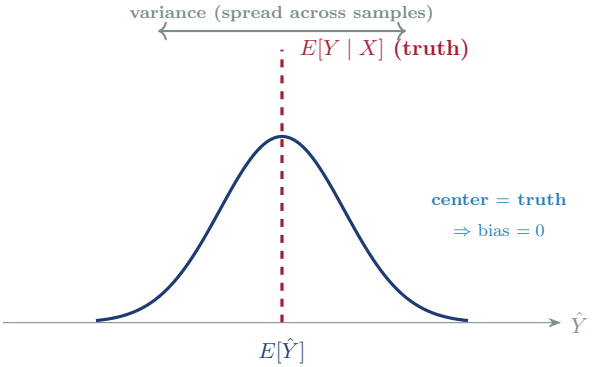
Low Bias, Low Variance
The goal



Low Bias, High Variance
OLS with many predictors

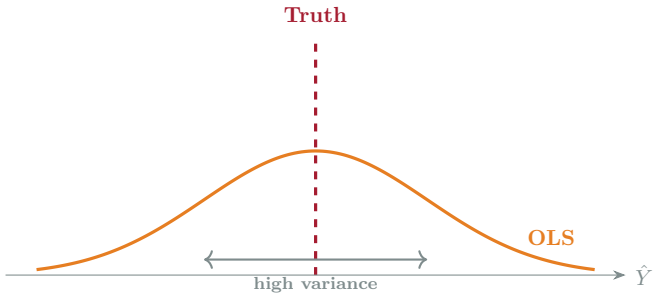
High Bias, Low Variance
Regularized regression

Now think of the same three scenarios as sampling distributions



Each random sample $\rightarrow$ different $\hat{Y}$. The distribution has a **center** (bias) and a **spread** (variance)

OLS with many predictors: centered on truth, but the spread is wide



Unbiased: $E[\hat{Y}]$ hits the truth on average.
But any single sample's $\hat{Y}$ could be far off

Overfitting is the reason OLS has high variance

- Each sample is different — different people, different quirks
- An overfit model memorizes the quirks of *this* sample
- Draw a new sample $\Rightarrow$ different quirks $\Rightarrow$ very different $\hat{Y}$
- That sample-to-sample instability is what "high variance" means

On average across all possible samples, OLS is right (unbiased).
But for *your* sample, it might be badly off (high variance)

Gauss-Markov: OLS is the best you can do *without* bias

- **BLUE:** Best Linear Unbiased Estimator
- Among all linear estimators with zero bias, OLS has the lowest variance
- But "best among unbiased" $\neq$ best overall
- What if you *allowed* a little bias — and got much lower variance?

That is exactly what Ridge, LASSO, and Elastic Net do

Now let's name what we saw

- The **center** of the sampling distribution = **Bias**
- The **spread** of the sampling distribution = **Variance**
- The randomness in the outcome itself = **Noise**

You saw all three in the pictures.
Now let's write them as equations

Bias means consistently wrong

Definition:

$$\text{Bias}(\hat{Y}) = E[\hat{Y}] - E[Y \mid X]$$

- **Question:** over repeated samples, are predictions centered on the truth?
- **OLS:** unbiased — centered correctly on average
- **Key insight:** small bias can be useful if it buys stability

Variance means unpredictable

Definition:

$$\mathrm{Var}(\hat{Y}) = E\big[(\hat{Y} - E[\hat{Y}])^2\big]$$

- **Question:** over repeated samples, how much do predictions change?
- **Many parameters** $\rightarrow$ high variance
- **Few parameters** $\rightarrow$ low variance

Noise is what no model can predict

Definition:

$$\epsilon_i = Y_i - E[Y_i \mid X_i]$$

- **Irreducible:** even the perfect model misses
- $E[\epsilon^2]$ = inherent randomness in outcomes
- No data or methods can reduce this term

The three sources of error are independent, so they add

In English:

1. Your prediction misses because you aimed at the wrong spot (bias)
2. Your prediction misses because your aim wobbles (variance)
3. Your prediction misses because the world is noisy (irreducible)

Independence: noise $\perp$ model choice $\Rightarrow$ the three terms add

Total miss = systematic miss + wobble + noise

Every prediction error comes from bias, variance, or noise

$$E\left[(\hat{Y} - Y)^2\right] = \underbrace{\text{Bias}^2}_{\text{systematic error}} + \underbrace{\text{Variance}}_{\text{instability}} + \underbrace{E[\epsilon^2]}_{\text{irreducible noise}}$$

We control bias and variance. We cannot control $E[\epsilon^2]$.

Where does the decomposition come from?

The "add and subtract the mean" trick:

$$
\begin{aligned}
E\big[(\hat{Y} - Y)^2\big] &= E\big[\big(\hat{Y} - E[\hat{Y}] + E[\hat{Y}] - Y\big)^2\big] \\
&= E\big[\big(\hat{Y} - E[\hat{Y}] + E[\hat{Y}] - E[Y \mid X] + E[Y \mid X] - Y\big)^2\big] \\
&= E\big[\big(\underbrace{\hat{Y} - E[\hat{Y}]}_{\text{estimation error}} + \underbrace{E[\hat{Y}] - E[Y \mid X]}_{\text{bias}} + \underbrace{E[Y \mid X] - Y}_{\epsilon} \big)^2\big] \\
&= E\big[(\hat{Y} - E[\hat{Y}])^2\big] + \big(E[\hat{Y}] - E[Y \mid X]\big)^2 + E[\epsilon^2] \\
&= \text{Variance} + \text{Bias}^2 + E[\epsilon^2]
\end{aligned}
$$

Key step: cross terms vanish because $\epsilon \perp \hat{Y}$ and $E[\hat{Y} - E[\hat{Y}]] = 0$.

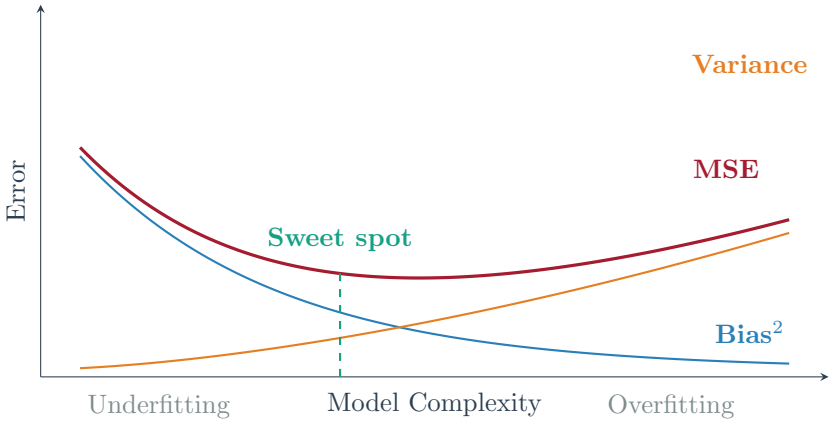
A little bias can buy a lot of stability

$$\text{MSE} \quad = \quad \text{Bias}^2 \; + \; \text{Variance}$$

Allow a *small* increase in Bias^2 $\rightarrow$
get a *large* decrease in Variance

Net effect: lower MSE, better predictions

The best model lives between underfitting and overfitting



Why not just use fewer variables?

If overfitting comes from too many predictors, just use fewer:

- 1 variable: high bias, low variance — misses real patterns
- 4 variables: better, but which 4?
- 20 variables: which 20? There are $\binom{35}{20} = 3.2$ billion subsets

The real problem:

- **You** would have to decide which variables to include
- Different researchers pick different variables → different answers
- No principled way to choose — until now

What if we kept all the variables but forced the model to be disciplined about how much weight each one gets?

The Big Idea: Penalize Complexity

What if we forced our regression coefficients to be smaller?

OLS has no constraints on coefficient size

Recall: OLS chooses $\hat{\beta}$ to minimize

$$\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{j=1}^{p} \hat{\beta}_j X_{ij} \right)^2$$

- No limit on how large $\hat{\beta}_j$ can be
- With many predictors, OLS chases noise

Penalized regression forces coefficients to earn their place

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

- $\lambda = 0$: no penalty $\rightarrow$ OLS
- $\lambda \rightarrow \infty$: all coefficients shrink toward zero
- Cross-validation picks the right λ

Ridge Regression (1970)

Hoerl & Kennard invented Ridge at DuPont in 1970

- **Problem:** correlated predictors $\rightarrow$ wildly unstable OLS estimates
- **Solution:** add a squared penalty to shrink coefficients
- **Objection:** "Ridge is biased! Gauss-Markov says OLS is best!"
- **Response:** BLUE $\neq$ lowest MSE

Ridge shrinks all coefficients but never eliminates any

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

- Penalty on the **sum of squared coefficients**
- Every variable stays in the model, just with smaller weight
- Never sets any coefficient *exactly* to zero

LASSO (1996)

Tibshirani invented LASSO at Toronto in 1996

Least **A**bsolute **S**hrinkage and **S**election **O**perator

- **Author:** Robert Tibshirani, Toronto $\rightarrow$ Stanford
- **Key move:** swap β_j^2 penalty for $|\beta_j|$ penalty
- **Result:** some coefficients shrink to **exactly zero**
- **LASSO selects**, not just shrinks

LASSO shrinks coefficients and sets some exactly to zero

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

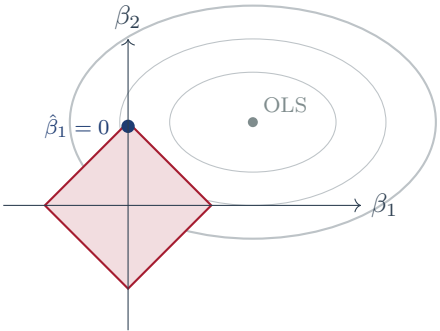
- Penalty on the **sum of absolute values**
- As λ increases, coefficients shrink
- Some hit exactly zero and drop out
- You get a sparse model: few variables, each one meaningful

LASSO tells you which variables the data actually needs

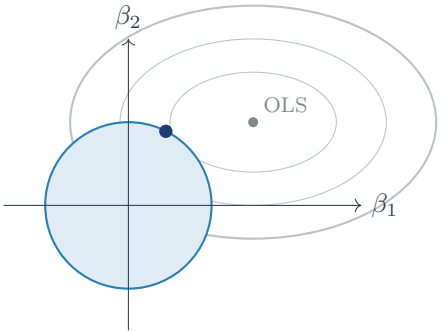
- **Prediction:** fewer noisy variables → less overfitting
- **Interpretation:** a short list of what matters
- **Honesty:** the data picks the model, not the analyst

The penalty selects variables → the analyst doesn't have to

The geometry explains why LASSO zeros out coefficients



LASSO (L1)

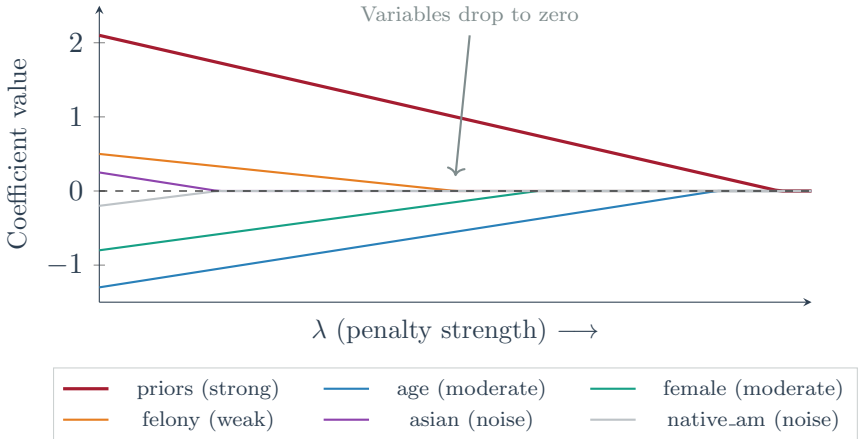


Ridge (L2)

Diamond corners → solutions land on axes → coefficients hit zero.

Circle has no corners → solutions never land exactly on an axis.

Prior record survives longest; noise variables drop out first



Elastic Net and the Full Picture

Elastic Net blends the LASSO and Ridge penalties

Zou & Hastie (2005)

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \Big[\alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \Big]$$

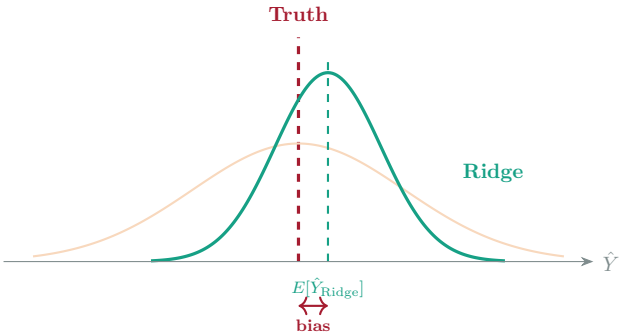
- $\alpha = 1$: pure LASSO
- $\alpha = 0$: pure Ridge
- $0 < \alpha < 1$: blend of both

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

	Ridge	LASSO	Elastic Net
Penalty	$\sum \beta_j^2$	$\sum \lvert\beta_j\rvert$	Both
Shrinks coefficients?	Yes	Yes	Yes
Sets coefficients to zero?	No	Yes	Yes
Variable selection?	No	Yes	Yes
Handles correlation?	Well	Poorly	Well

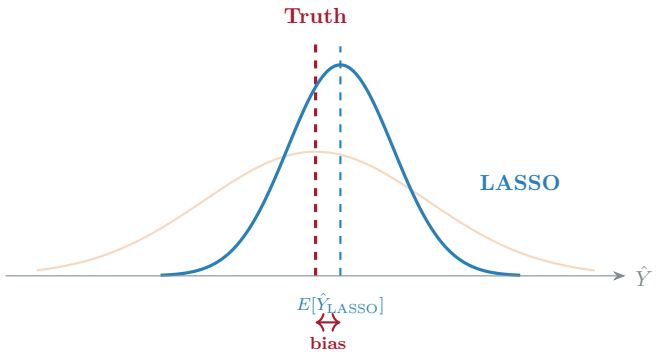
What Do These Methods Do to the Sampling Distribution?

Remember the wide OLS distribution? Here's what Ridge does to it



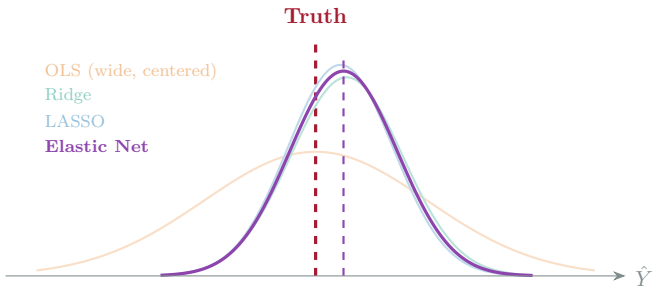
Narrower = lower variance. Small shift = small bias. Total MSE is *lower* than OLS

LASSO does the same thing — and kills noise variables



Same bias-variance tradeoff as Ridge — but LASSO also **selects variables**, setting noise predictors to exactly zero

All three penalized methods: tighter, slightly off-center



Every penalized method: *tighter and slightly off-center.*
The small bias is the price. The low variance is the payoff

How Do We Choose λ?

Back to crime prediction: which penalty strength works best?

LASSO for rearrest prediction: 30 candidate variables from admin data (age, prior arrests, charge type, time served, . . .)

- $\lambda = 0$: keep all 30 variables $\rightarrow$ OLS, overfits
- λ small: keep 22 variables, drop 8 weak ones
- λ medium: keep 9 variables, drop 21
- λ large: keep 1 variable (prior record), drop 29 $\rightarrow$ underfits

Which λ gives the most honest prediction of who gets rearrested?

The λ dial controls how much the model trusts the data

- λ **too small** — model trusts the data too much

  It memorizes quirks of the training sample. On new defendants, predictions are noisy.

- λ **too large** — model doesn't trust the data enough

  It ignores real patterns. Predictions are stable but systematically off.

- λ **just right** — the sweet spot

We need a method to **test** each candidate λ on data the model hasn't seen

How would you test λ by hand?

Imagine you have 500 defendants and want to predict rearrest:

1. Set aside the first 100 defendants — don't touch them yet
2. Fit your LASSO on the other 400, using some λ
3. Predict rearrest for the 100 you held out
4. Compute RMSE on those 100 predictions
5. Now put the first 100 back and hold out the next 100
6. Repeat: fit on 400, predict on 100, compute RMSE

After 5 rounds, every defendant has been in the "held-out" group exactly once

Each round's held-out group is called a **fold**. Five rounds = five folds = "5-fold cross-validation."

Try every λ, pick the one with the lowest average RMSE

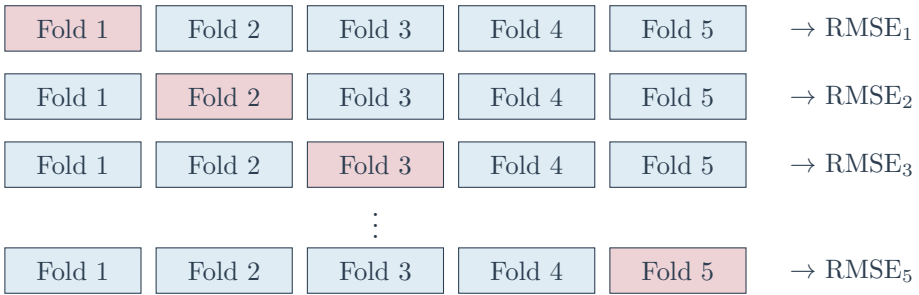
For each candidate λ:

1. Split data into 5 groups (folds)
2. Rotate through: each fold takes a turn as the test set
3. Average the 5 RMSE values $\rightarrow$ CV-RMSE for this λ

Then compare across λ values:

λ	Variables kept	CV-RMSE
0 (OLS)	30	0.48
0.01	22	0.44
0.05	**9**	**0.39 $\leftarrow$ best**
0.50	3	0.43
5.00	1	0.51

k-fold cross-validation: each observation gets a turn



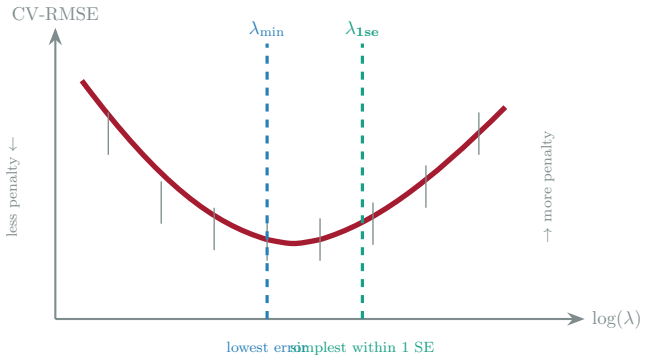
$$\text{CV-RMSE} = \frac{1}{5}(\text{RMSE}_1 + \cdots + \text{RMSE}_5)$$

Minimizing training error always picks $\lambda = 0$

- $\lambda = 0$ means no penalty $\rightarrow$ OLS
- OLS gives the best in-sample fit by definition
- But we already know OLS overfits!

Cross-validation estimates **out-of-sample** error $\Rightarrow$ optimal $\lambda > 0$

The CV error curve tells you which λ to use



- λ_{min}**:** lowest CV error — best raw prediction
- $\lambda_{\textbf{1se}}$**:** simplest model within 1 SE of minimum — fewer variables, nearly as good
- **Default:** use $\lambda_{1\text{se}}$ — simpler and almost as accurate

What's Next

Thursday: we fix overfitting with penalized regression

- **Review** the bias-variance tradeoff
- **Learn** Ridge, LASSO, and Elastic Net — the tools that fix overfitting
- **See** how cross-validation picks the right penalty
- **Apply** it: Baker (2025) and a $5 billion securities case
- **Quantify uncertainty** with prediction intervals

Project Milestone 1 due Thursday by 11:59pm
PS3 (COMPAS prediction) due Thursday, April 2

The data should pick the model, not the analyst.

A little bias buys a lot of stability.