Collinearity and Model Selection
with Boston Housing Data

DATA 200BP

January 14, 2026

Learning Goals

By the end of this module, you should be able to:

- Explain *why* collinearity causes unstable inference in multiple regression.
- Detect collinearity using correlations, condition indices, and the variance inflation factor (VIF).
- Understand how model selection (AIC/BIC/CV MSE; forward/backward/stepwise) can interact with collinearity.
- Motivate and apply regularization (ridge and lasso), including tuning via cross-validation.
- Use the Boston housing data (MEDV as response) as a running case study.

Roadmap

1. Why collinearity is an issue (variance, identifiability, geometry)
2. Detecting collinearity: correlations, eigenvalues/condition number, VIF
3. Boston Housing: full OLS model & VIF in practice
4. Model selection: AIC/BIC/CV MSE (forward/backward/stepwise)
5. Why selection does *not* guarantee low collinearity
6. Regularization: ridge & lasso (history, tuning, complexity)
7. Boston Housing: ridge vs lasso (CV, coefficient paths, sparsity)

Multiple Linear Regression Setup

Let $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ (columns are predictors). The multiple regression model is

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2 I_n).$$

The ordinary least squares (OLS) estimator is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y \quad \text{(when } X^\top X \text{ is invertible).}$$

What is Collinearity? (Intuition + Linear Algebra)

Definition (exact vs. approximate)

Exact collinearity: columns of X are linearly dependent, i.e. $\mathrm{rank}(X) < p$.
Approximate collinearity: columns are nearly dependent (some singular values are very small).

Invertibility of $X^\top X$

Recall $\hat{\beta} = (X^\top X)^{-1} X^\top y$ *only if* $X^\top X$ is invertible.

- If $\mathrm{rank}(X) < p$, then $X^\top X$ is **singular** (not invertible) and OLS is not unique.
- If $\mathrm{rank}(X) = p$ but columns are highly correlated, then $X^\top X$ is **ill-conditioned**: it has eigenvalues $\lambda_{\min} \approx 0$ and $\kappa(X^\top X) = \lambda_{\max}/\lambda_{\min}$ is large.

Why it matters

Small perturbations in y or in X can create large changes in $(X^\top X)^{-1}$, hence in $\hat{\beta}$.

Consequences for Inference

- **Variance inflation:** $(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$ can have very large diagonal entries.
- **Unstable signs/magnitudes:** coefficients may flip sign or change dramatically when adding/removing a correlated predictor.
- **Wide confidence intervals, weak t-tests:** interpretation becomes fragile.
- **Prediction vs interpretation:** prediction may remain strong (because $X\hat{\beta}$ can be stable), while individual coefficients are unstable.

Important nuance: Under correct specification, collinearity does *not* bias OLS; it increases uncertainty and instability (and model selection / shrinkage can introduce bias intentionally).

$p \gg n$: Another Motivation for Regularization

High-dimensional regression

When the number of predictors exceeds the sample size ($p > n$), we have $\mathrm{rank}(X) \leq n < p$, so $X^\top X$ is **singular**.

- OLS solutions are **non-unique**: there are infinitely many β with the same training RSS.
- Even when p is close to n, $(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$ (when defined) can explode.
- Regularization adds information/constraints to stabilize estimation:
  - **Ridge** yields a unique solution for any p, n because $X^\top X + \lambda I$ is invertible.
  - **Lasso** can set coefficients exactly to zero, achieving model selection.

Geometry (One Picture Worth Many Words)

OLS finds the point in the column space of X closest to y. If predictors are nearly collinear, the column space is "thin" in some directions, so many β values produce almost the same fitted values.

Practical takeaway

When collinearity is high, there are many nearly-equivalent explanations of the data, so coefficient-level inference becomes poorly identified.

Detection Methods (Quick Checklist)

- Pairwise correlation matrix / scatterplots (fast but incomplete).
- Variance Inflation Factor (VIF): quantifies how much $(\hat{\beta}_j)$ is inflated.
- Condition number / eigenvalues of $X^\top X$ (near-singularity diagnostics).
- Domain knowledge: known redundancies (e.g., overlapping measurements).

Variance Inflation Factor (VIF): Definition

Fix predictor x_j. Regress x_j on the remaining predictors X_{-j}:

$$x_j = X_{-j}\gamma + r_j, \qquad R_j^2 = 1 - \frac{\|r_j\|^2}{\|x_j - \bar{x}_j \mathbf{1}\|^2}.$$

Definition

$$\mathrm{VIF}_j := \frac{1}{1 - R_j^2}.$$

Rules of thumb: VIF > 5 suggests concern; VIF > 10 is often considered severe (context-dependent).

VIF: Where it Comes From (Key Derivation Sketch)

In multiple regression, the variance of $\hat{\beta}_j$ can be written as

$$(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \cdot \frac{1}{1 - R_j^2}.$$

- The first factor is the variance you would get in a simple regression on x_j alone.
- The second factor, $\frac{1}{1-R_j^2}$, is exactly VIF_j.

Interpretation

If R_j^2 is close to 1, then x_j is almost perfectly predicted by other covariates, so $(\hat{\beta}_j)$ explodes.

Boston Housing Data: Setup

- $n = 506$ census tracts, $p = 13$ predictors.
- Response: `MEDV` (median home value, in \$1000s).
- Model we start with:

$$\texttt{MEDV} = \beta_0 + \beta_1 \texttt{CRIM} + \cdots + \beta_{13} \texttt{LSTAT} + \varepsilon.$$

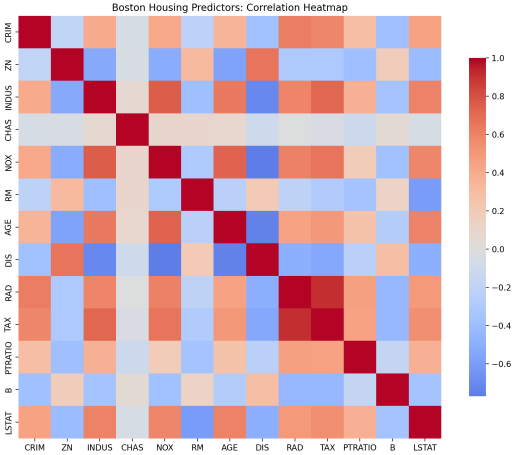
Full OLS model fit (all predictors)

For this dataset, the full model has $R^2 \approx 0.741$ and adjusted $R^2 \approx 0.734$.

Boston Housing Data: Variable Descriptions

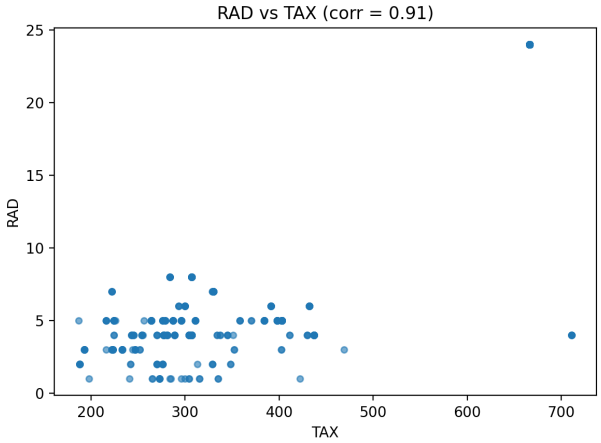
`CRIM`	per-capita crime rate by town
`ZN`	proportion of residential land zoned for lots over 25,000 sq.ft.
`INDUS`	proportion of non-retail business acres per town
`CHAS`	Charles River dummy (=1 if tract bounds river; 0 otherwise)
`NOX`	nitrogen oxides concentration (parts per 10 million)
`RM`	average number of rooms per dwelling
`AGE`	proportion of owner-occupied units built prior to 1940
`DIS`	weighted distances to five Boston employment centers
`RAD`	index of accessibility to radial highways
`TAX`	full-value property-tax rate per \$10,000
`PTRATIO`	pupil–teacher ratio by town
`B`	$1000(B_k - 0.63)^2$ where B_k is the proportion of Black residents (historical variable)
`LSTAT`	% lower status of the population
`MEDV`	median value of owner-occupied homes (in \$1000s)

Correlation Heatmap (First Pass)



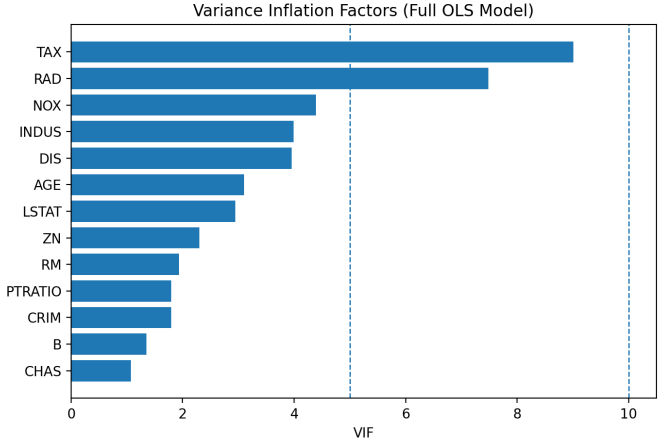
Pairwise correlations can reveal obvious relationships, but collinearity can involve *more than two* predictors.

A Classic Collinearity Pair: RAD vs TAX



In this dataset, $\mathrm{corr}(\texttt{RAD}, \texttt{TAX}) \approx 0.91$, suggesting potential redundancy.

VIF on the Full OLS Model



Observed pattern

The largest VIFs occur for `TAX` and `RAD`, indicating notable collinearity

How Collinearity Manifests: Coefficient Instability

A quick demonstration: compare estimated coefficients when dropping a correlated variable.

	Full model	Drop TAX	Drop RAD
$\hat{\beta}_{\texttt{RAD}}$	0.306	0.135	–
$\hat{\beta}_{\texttt{TAX}}$	-0.012	–	0.001

Interpretation

When two predictors carry overlapping information, their individual coefficients can change substantially across "nearby" models.

Addressing Collinearity (Preview)

Common strategies:

- **Drop** one of a redundant pair (guided by science and measurement quality).
- **Combine** variables (indices, PCA/PLS, domain-informed composites).
- **Regularize** (ridge/lasso): accept some bias to reduce variance and stabilize estimation.
- **Collect more data** (helps, but not always possible).

Next: model selection and why it does *not* automatically solve collinearity.

Why Do Model Selection?

We often want:

- **Interpretability:** a simpler model that highlights key drivers.
- **Prediction:** low out-of-sample error (not necessarily sparse).
- **Cost / measurement constraints:** fewer variables are cheaper to collect.

But:

selecting a model changes the inferential target and can interact with collinearity in nontrivial ways.

Common Selection Criteria

Let k be the number of free parameters (including intercept), and $\ell(\hat{\theta})$ the maximized log-likelihood.

- **AIC:** $\mathrm{AIC} = 2k - 2\ell(\hat{\theta})$
- **BIC:** $\mathrm{BIC} = k \log n - 2\ell(\hat{\theta})$
- **CV MSE:** $\mathrm{MSE}_{CV} = \frac{1}{K} \sum_{m=1}^{K} \frac{1}{|V_m|} \sum_{i \in V_m} (y_i - \hat{y}_i^{(-m)})^2$

Heuristic

AIC penalizes complexity lightly; BIC penalizes more strongly (often yields smaller models).

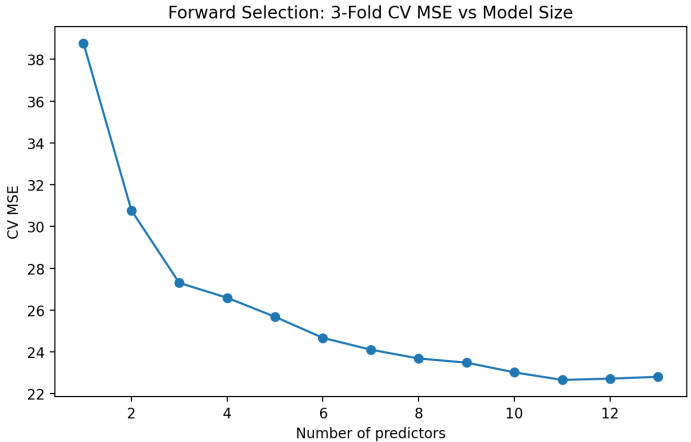
Selection Algorithms

- **Forward selection:** start empty, add the best variable each step.
- **Backward elimination:** start full, drop the least useful variable each step.
- **Stepwise (both):** alternate add/drop to improve a criterion (AIC/BIC).

Greedy nature

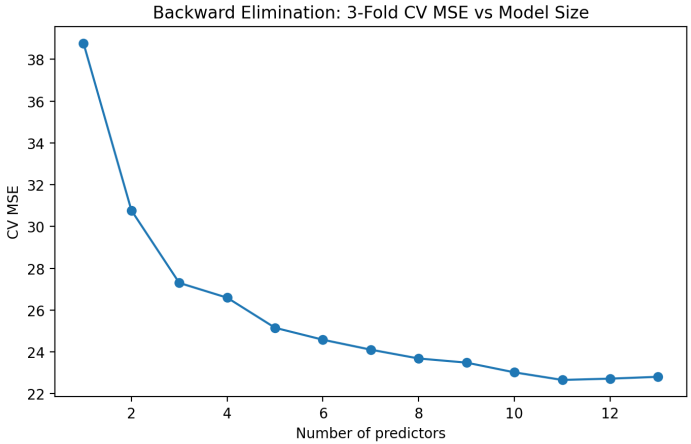
Stepwise methods do *not* guarantee the globally best subset (they search locally).

Forward Selection by CV MSE (Boston)



CV-based selection targets prediction error; the "best" model size is where CV MSE is minimized.

Backward Elimination by CV MSE (Boston)



Forward and backward can yield different paths, but often similar best model sizes in practice.

Stepwise AIC vs Stepwise BIC: Selected Predictors

Stepwise AIC selected (11 predictors)

`LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN, CRIM, RAD, TAX`

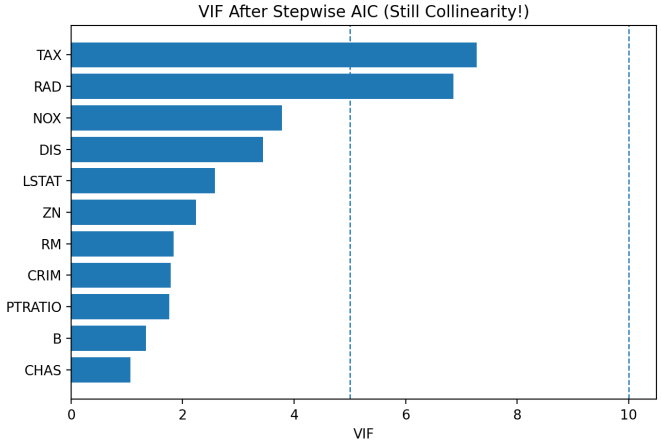
Stepwise BIC selected (8 predictors)

`LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN`

Key point

Model selection does not guarantee a collinearity-free model. You still need diagnostics after selection.

VIF After Stepwise AIC (Collinearity Can Remain)



Even after AIC-based selection, `RAD` and `TAX` remain strongly collinear (large VIFs).

Motivation for Regularization

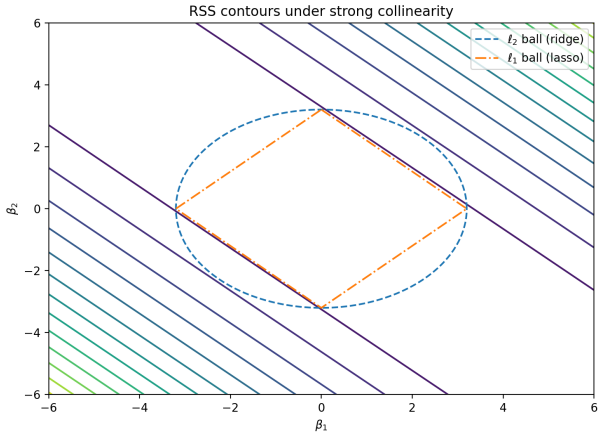
When predictors are correlated:

- OLS can have high variance and unstable coefficients.
- Stepwise selection can be unstable and may keep correlated predictors together.

Regularization idea

Allow a small amount of bias in exchange for a potentially large reduction in variance and improved stability/prediction.

Geometry: Why Collinearity Creates Instability



When predictors are highly correlated, the RSS contours become *elongated* in coefficient space, so many (β_1, β_2) pairs fit similarly. Ridge (circle constraint) shrinks toward the origin; lasso (diamond constraint) often hits an axis, producing sparsity.

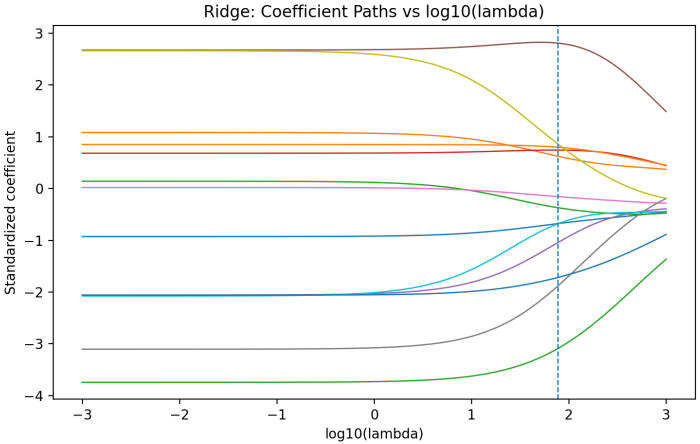
Ridge Regression (Hoerl & Kennard)

Ridge solves:

$$\hat{\beta}^{\text{ridge}}(\lambda) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

- $\lambda \geq 0$ is a tuning parameter.
- Larger λ shrinks coefficients toward 0 (but usually not exactly 0).
- **Standardize predictors** before fitting so the penalty is comparable across variables.

Ridge: Coefficient Paths (Boston)



Dashed line marks CV-chosen λ (here, $\lambda \approx 76.6$ on standardized predictors).

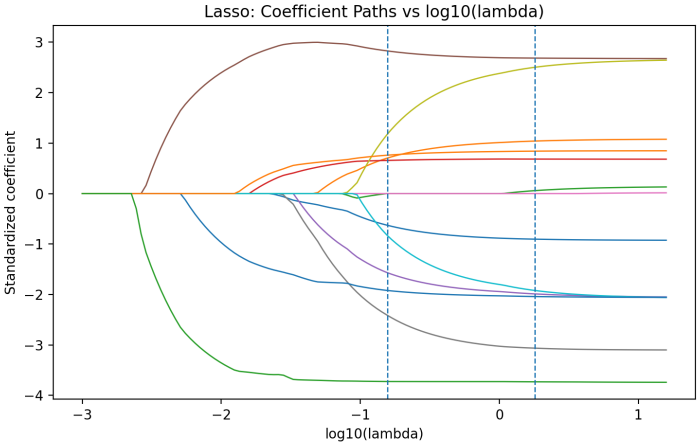
Lasso (Tibshirani): Sparsity via ℓ_1 Penalty

Lasso solves:

$$\hat{\beta}^{\text{lasso}}(\lambda) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

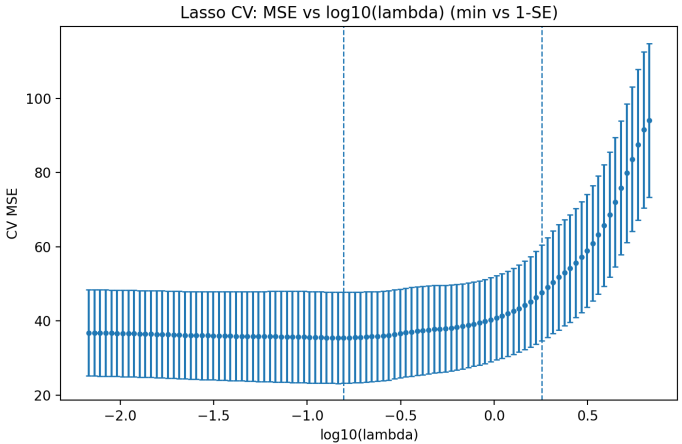
- The ℓ_1 penalty encourages **exact zeros** $\Rightarrow$ model selection.
- With correlated predictors, lasso often picks one and shrinks others to 0 (not guaranteed).
- We choose λ by cross-validation (often report λ_{min} and λ_{1SE}).

Lasso: Coefficient Paths (Boston)



Dashed lines mark $\lambda_{\min}$ and the more conservative λ_{1SE}.

Choosing λ by Cross-Validation (Boston)



- $\lambda_{\min}$: minimizes CV MSE.
- λ_{1SE}: largest λ within one standard error of the minimum (simpler model).

Boston Result: Lasso for Selection + (Often) Less Collinearity

For this dataset:

- $\lambda_{\text{min}} \approx 0.157$ (many variables retained).
- $\lambda_{1SE} \approx 1.80$ (much sparser).

Lasso at λ_{1SE} selects (3 predictors)

`RM, PTRATIO, LSTAT`

Check collinearity in the selected subset

VIFs for `RM, PTRATIO, LSTAT` are all close to 1–2 (low collinearity).

Key Takeaways

- Collinearity primarily hurts **interpretability and inference** by inflating variance and creating instability.
- VIF provides an interpretable diagnostic: how much variance is inflated due to dependence among predictors.
- Model selection (AIC/BIC/CV) targets parsimony/prediction but does *not* guarantee low collinearity.
- Ridge stabilizes estimates; lasso can perform variable selection (and can reduce collinearity by selecting fewer variables).
- Always do **post-selection diagnostics**: re-check VIF/conditioning and interpretability.

Next Steps / Practice

- Reproduce all analyses in the provided R and Python scripts.
- Try alternative responses (e.g., log(MEDV)) and compare selection stability.
- Explore elastic net (mixture of ridge + lasso) for correlated predictors.