Econometrics

Ridge and Lasso Regression

Jose Angel Garcia Sanchez

Université Paris 1 Panthéon-Sorbonne
jagarsanc@gmail.com

November 2025

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Outline

Motivation

- OLS minimizes the sum of squared residuals: $\min_\beta \sum_{i=1}^{n}(y_i - x_i'\beta)^2$
- OLS works well when $n >> p$ (many observations, few variables)
- But OLS can fail when:
  - p is large relative to n (high-dimensional data)
  - There is multicollinearity among regressors
  - We want to improve out-of-sample prediction
- Solution: add a *penalty term* to the objective function
- This is called *regularization* or *shrinkage*

The Bias-Variance Trade-off

- Recall that the Mean Squared Error can be decomposed: $MSE = Bias^2 + Variance$
- OLS is unbiased but can have high variance, especially when:
  - Predictors are highly correlated (multicollinearity)
  - The number of predictors is large
- Regularization introduces some bias but reduces variance
- If the reduction in variance exceeds the increase in squared bias, we get lower MSE
- This improves prediction accuracy, especially out-of-sample

The General Penalized Regression Framework

- General form: $\min_\beta \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \cdot P(\beta) \right\}$
- $\lambda \geq 0$ is the *tuning parameter* (or regularization parameter)
- $P(\beta)$ is the *penalty function*
- When $\lambda = 0$: we get OLS
- When $\lambda \to \infty$: coefficients shrink toward zero
- Different choices of $P(\beta)$ give different estimators:
  - Ridge: $P(\beta) = \sum_{j=1}^{p} \beta_j^2$ (L2 penalty)
  - Lasso: $P(\beta) = \sum_{j=1}^{p} |\beta_j|$ (L1 penalty)
  - Elastic Net: combination of both

Important Preliminary: Standardization

- Regularization penalizes the size of coefficients
- But coefficient size depends on the scale of variables
- A variable measured in millions will have a tiny coefficient compared to one measured in units
- Solution: *standardize* all variables before estimation
- For each variable x_j: $\tilde{x}_j = \frac{x_j - \bar{x}_j}{s_{x_j}}$
- After standardization, all variables have mean 0 and standard deviation 1
- The intercept is typically not penalized
- After estimation, coefficients can be transformed back to original scale if needed

Ridge Regression: Definition

- Ridge regression adds an L2 penalty to the OLS objective:

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Equivalently, in matrix notation:

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ (Y - X\beta)'(Y - X\beta) + \lambda \beta'\beta \right\}$$

- The penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks coefficients toward zero
- Large coefficients are penalized more heavily (quadratic penalty)

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Ridge Regression: Closed-Form Solution

- Taking the first-order condition and setting it to zero:

$$\frac{\partial}{\partial \beta}\left[(Y - X\beta)'(Y - X\beta) + \lambda \beta'\beta\right] = 0$$

- We get: $-2X'Y + 2X'X\beta + 2\lambda\beta = 0$
- Solving for β:

$$\hat{\beta}^{Ridge} = (X'X + \lambda I)^{-1}X'Y$$

- Compare with OLS: $\hat{\beta}^{OLS} = (X'X)^{-1}X'Y$
- The term λI is added to $X'X$ before inversion
- This ensures $(X'X + \lambda I)$ is always invertible, even when $X'X$ is singular

Ridge Regression: Properties (1)

- Ridge estimator is biased: $E(\hat{\beta}^{Ridge}) \neq \beta$
- The bias is: $Bias(\hat{\beta}^{Ridge}) = -\lambda(X'X + \lambda I)^{-1}\beta$
- The variance is: $Var(\hat{\beta}^{Ridge}) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$
- As λ increases:
  - Bias increases (coefficients shrink more toward zero)
  - Variance decreases (estimates become more stable)
- There exists an optimal λ^* that minimizes MSE

Ridge Regression: Properties (2)

- Ridge does NOT set coefficients exactly to zero
- All predictors remain in the model (no variable selection)
- Ridge is particularly useful when:
  - Predictors are highly correlated (multicollinearity)
  - We believe all predictors are relevant
  - We want to stabilize estimates
- The ridge penalty can be viewed as a Bayesian prior: $\beta \sim N(0, \sigma^2/\lambda)$

Ridge Regression: Geometric Interpretation

- Ridge regression can be written as a constrained optimization:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t$$

- The constraint region is a sphere (or hypersphere in higher dimensions)
- The OLS solution is at the center of the elliptical contours of the RSS
- The ridge solution is where the elliptical contours first touch the sphere
- Because the sphere has no corners, the solution typically has all coefficients non-zero
- Smaller t (equivalently, larger λ) means more shrinkage

Ridge and Multicollinearity

- Recall: with perfect multicollinearity, $X'X$ is singular and OLS fails
- With near-perfect multicollinearity, $(X'X)^{-1}$ has very large elements
- This leads to inflated variances and unstable OLS estimates
- Ridge regression solves this by adding λI to $X'X$
- Even if $X'X$ is singular, $(X'X + \lambda I)$ is always invertible for $\lambda > 0$
- The eigenvalues of $(X'X + \lambda I)$ are $d_j + \lambda$, where d_j are eigenvalues of $X'X$
- Small eigenvalues (source of instability) become λ instead of nearly zero

Lasso Regression: Definition

- Lasso = Least Absolute Shrinkage and Selection Operator
- Lasso adds an L1 penalty to the OLS objective:

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

- The key difference from Ridge: absolute values instead of squares
- This seemingly small change has profound implications
- Lasso performs both shrinkage AND variable selection

Lasso: No Closed-Form Solution

- Unlike Ridge, Lasso has no closed-form solution
- The absolute value function $|\beta_j|$ is not differentiable at $\beta_j = 0$
- Optimization requires iterative algorithms:
  - Coordinate descent (most common)
  - LARS (Least Angle Regression)
  - Proximal gradient methods
- For a single coefficient with orthonormal design:

$$\hat{\beta}_j^{Lasso} = \text{sign}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda)_+$$

- Where $(x)_+ = \max(0, x)$ is the soft-thresholding operator

Lasso: Variable Selection Property

- The key feature of Lasso: it sets some coefficients exactly to zero
- This performs automatic variable selection
- As λ increases, more coefficients become exactly zero
- The model becomes more sparse (fewer non-zero coefficients)
- This is very useful when:
  - We have many potential predictors
  - We believe only a subset are truly relevant
  - We want an interpretable, parsimonious model
- Lasso identifies which variables matter and estimates their effects simultaneously

Lasso: Geometric Interpretation

- Lasso can be written as a constrained optimization:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

- The constraint region is a diamond (or cross-polytope in higher dimensions)
- The diamond has corners on the axes
- The elliptical RSS contours often touch the diamond at a corner
- At a corner, one or more coefficients are exactly zero
- This is why Lasso produces sparse solutions

Lasso: Properties

- Lasso estimator is biased (like Ridge)
- Lasso is consistent for variable selection under certain conditions
- However, Lasso has some limitations:
  - With highly correlated predictors, Lasso tends to select one and ignore the others
  - When $p > n$, Lasso selects at most n variables
  - Lasso estimates for selected variables are biased toward zero
- These limitations motivate extensions like Elastic Net and Adaptive Lasso

Comparing Ridge and Lasso

Property	Ridge	Lasso
Penalty	$\sum \beta_j^2$ (L2)	$\sum \lvert\beta_j\rvert$ (L1)
Closed-form solution	Yes	No
Variable selection	No	Yes
Handles multicollinearity	Yes	Partially
Correlated predictors	Shrinks together	Selects one
When $p > n$	Works	Selects $\leq n$
Bayesian interpretation	Normal prior	Laplace prior

Elastic Net: Combining Ridge and Lasso

- Elastic Net combines L1 and L2 penalties:

$$\hat{\beta}^{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Alternative parameterization with mixing parameter $\alpha \in [0, 1]$:

$$\hat{\beta}^{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

- $\alpha = 1$: Lasso; $\alpha = 0$: Ridge

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Elastic Net: Properties

- Combines the best of both worlds:
  - Variable selection (from Lasso)
  - Grouping effect (from Ridge)
- With correlated predictors, Elastic Net tends to select groups of correlated variables together
- Can select more than n variables when $p > n$
- Particularly useful when:
  - Predictors are highly correlated
  - We want both sparsity and stability
  - The number of predictors exceeds sample size

The Role of λ

- The tuning parameter λ controls the amount of regularization
- $\lambda = 0$: no penalty, we get OLS
- $\lambda \to \infty$: maximum penalty, all coefficients go to zero
- Intermediate λ: trade-off between bias and variance
- How to choose λ? Several approaches:
  - Cross-validation (most common)
  - Information criteria (AIC, BIC)
  - Theoretical considerations

Cross-Validation (1)

- K-fold cross-validation is the standard approach
- Procedure:
  1. Divide data into K roughly equal parts (folds)
  2. For each fold $k = 1, \ldots, K$:
     - Fit model on all data except fold k
     - Compute prediction error on fold k
  3. Average the K prediction errors
- Repeat for a grid of λ values
- Choose λ that minimizes cross-validation error
- Common choices: $K = 5$ or $K = 10$

Cross-Validation (2)

- The CV error as a function of λ typically has a U-shape
- Small λ: high variance, low bias (overfitting)
- Large λ: low variance, high bias (underfitting)
- Two common choices:
  - λ_{min}: minimizes CV error
  - λ_{1se}: largest λ within 1 standard error of minimum
- λ_{1se} gives a more parsimonious model (more regularization)
- This is the "one-standard-error rule"

Information Criteria

- Alternative to cross-validation: use information criteria
- For Lasso, we can use:
  - AIC: $AIC = n \log(RSS/n) + 2df$
  - BIC: $BIC = n \log(RSS/n) + \log(n) \cdot df$
- The degrees of freedom for Lasso is approximately the number of non-zero coefficients
- For Ridge: $df = \sum_{j=1}^{p} \frac{d_j}{d_j + \lambda}$ where d_j are eigenvalues of $X'X$
- BIC tends to select sparser models than AIC
- Cross-validation is generally preferred in practice

Inference After Selection

- Standard inference (t-tests, confidence intervals) is invalid after Lasso selection
- Why? The selection process introduces additional randomness
- Naive standard errors are too small (underestimate uncertainty)
- Confidence intervals have incorrect coverage
- Several approaches have been developed:
  - Sample splitting
  - Selective inference
  - Debiased/Desparsified Lasso
  - Bootstrap methods

Sample Splitting

- Simple approach to valid inference:
  1. Split sample randomly into two parts
  2. Use first part for variable selection (Lasso)
  3. Use second part for estimation and inference (OLS on selected variables)
- Advantages: simple, valid inference
- Disadvantages:
  - Loses statistical power (uses only half the data for each step)
  - Results depend on the random split
- Can average over multiple random splits

Post-Lasso OLS

- Also called "Post-Selection OLS" or "OLS post Lasso"
- Procedure:
  1. Run Lasso to select variables (non-zero coefficients)
  2. Run OLS using only the selected variables
- The OLS estimates are less biased than Lasso estimates
- Removes the shrinkage bias for selected variables
- Caution: standard errors from this OLS are still not valid for inference
- Valid inference requires additional corrections

Adaptive Lasso

- Standard Lasso penalizes all coefficients equally
- This can lead to too much bias for large coefficients
- Adaptive Lasso uses weighted penalties:

$$\hat{\beta}^{ALasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}$$

- Weights: $w_j = 1/|\hat{\beta}_j^{init}|^\gamma$ for some initial estimator $\hat{\beta}^{init}$
- Typically $\gamma = 1$ or $\gamma = 2$
- Variables with small initial estimates are penalized more
- Adaptive Lasso has the "oracle property": asymptotically selects the correct model and estimates are efficient

Implementation in Software

- Ridge and Lasso are implemented in all major statistical software:
  - R: `glmnet` package (gold standard)
  - Python: `scikit-learn` (Ridge, Lasso, ElasticNet)
  - Stata: `lasso`, `elasticnet` commands
- The `glmnet` package uses coordinate descent and is very fast
- Computes entire solution path (all values of λ) efficiently
- Built-in cross-validation functions

Practical Recommendations

- Always standardize predictors before regularization
- Use cross-validation to choose λ
- Consider the one-standard-error rule for more parsimony
- If predictors are highly correlated, consider Elastic Net
- For inference, use appropriate methods (not naive standard errors)
- Check model performance on held-out test data
- Compare with OLS on a subset of variables (sanity check)
- Remember: regularization is primarily for prediction, not causal inference

When to Use What?

- **OLS**: $n >> p$, no multicollinearity, focus on inference
- **Ridge**: Multicollinearity, all predictors believed relevant, prediction focus
- **Lasso**: Many predictors, only some believed relevant, want variable selection
- **Elastic Net**: Many correlated predictors, want both selection and grouping
- **Adaptive Lasso**: Want consistent variable selection with oracle properties
- In practice: try several methods and compare via cross-validation

Summary

- Regularization adds a penalty to the OLS objective function
- Ridge (L2): shrinks coefficients, handles multicollinearity, no selection
- Lasso (L1): shrinks and selects variables, produces sparse models
- Elastic Net: combines Ridge and Lasso benefits
- The tuning parameter λ is chosen via cross-validation
- Standard inference is invalid after Lasso selection
- These methods are powerful tools for prediction in high-dimensional settings
- But remember: regularization introduces bias, which may not be desirable for causal inference