Ridge Regression

Predictive Modeling & Statistical Learning

Gaston Sanchez

Ridge Regression

Introduction

- Ridge Regression introduced by Hoerl (1962)
- Refined by Hoerl and Kennard (1970)
- Motivated by multicollinearity problems in regression analysis
- Developed to address the instability of estimated regression coefficients (when predictors are highly correlated)

Motivation

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Motivation

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Predicted model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{b}$$

Motivation

Assuming that $\mathbf{X}$ is of full column-rank, the OLS solution for

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{b}$$

is given by:

$$\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Multicollinearity Issues

Potential instability—due to multicollinearity—in the OLS solution affecting

$$(\mathbf{X}^\top \mathbf{X})^{-1}$$

Ridge Idea

Ridge Idea

Hoerl and Kennard proposed to modify $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ by adding a small constant k to the diagonal entries of $\mathbf{X}^\mathsf{T}\mathbf{X}$ before taking the inverse:

$$\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I}$$

Why/how did they do that?

EVD

Before moving on, let's recall that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is a symmetric matrix, and that we can use its eigendecomposition to rewrite it as:

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$$

where $\mathbf{V}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix containing the eigenvalues in decreasing order.

EVD

We can use the eigendecomposition to find $(\mathbf{X}^\top\mathbf{X})^{-1}$:

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{V}\mathbf{\Lambda}_*^{-1}\mathbf{V}^\top$$

where $\mathbf{\Lambda}_*$ is a diagonal matrix with inverses of non-null eigenvalues.

Ridge Idea

There is a close relationship between:

$$\mathbf{X}^\mathsf{T}\mathbf{X} \qquad \text{and} \qquad \mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I}$$

Can you guess which one it is?

Ridge Ideas

The matrices $\mathbf{X}^\mathsf{T}\mathbf{X}$ and $\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I}$

- have the same eigenvectors
- but different eigenvalues: $\{\lambda_j\}_{j=1}^p$ and $\{\lambda_j + k\}_{j=1}^p$

Ridge Ideas

In the OLS solution:

$$\mathbf{b} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

we could replace $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ by $(\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I})^{-1}$

$$\mathbf{b} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{kI})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

which would allow us to increase the magnitude of the eigenvalues, especially the ones that are close to zero.

Ridge Ideas

The question is how to find an adequate value for the constant k to be added

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I})^{-1}$$

Ridge Idea

Ridge Reression Estimator

The result is the ridge regression estimator (or ridge rule)

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{rr}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + k\mathbf{I}_{\mathbf{p}})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

- This gives us a class of estimators, indexed by a parameter k
- When $k > 0$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{rr}}$ is a biased estimator of $\boldsymbol{\beta}$
- When $k = 0$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{rr}}$ reduces to the OLS estimator

Properties

The ridge regression estimator can be characterized in three different ways

- as a **constrained** estimator with restricted length that minimizes the RSS
- as a **shrinkage** estimator that shrinks the LS estimator toward the origin
- as a **Bayes** estimator (given suitable priors)

Penalized Least Squares Problem

A ridge regression estimator is the solution of a penalized least squares problem

$$ESS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

subject to

$$\|\boldsymbol{\beta}\|^2 \leq c$$

where c is an arbitrary constant

Estimating the Ridge Parameter

Controlling (or *regularizing*) the parameter estimates can be accomplished by adding a penalty to the ESS if the estimates become large.

The effect of the penalty is that the parameter estimates are only allowed to become large if there is a proportional reduction in ESS.

By adding the penalty, we are making a trade-off between the model variance and bias. By sacrificing some bias, we can often reduce the variance enough to make the overall ESS lower than unbiased models.

Shrinkage Estimator

A ridge regression estimator is a **shrinkage** estimator that shrinks the OLS estimator toward zero.

Recall the SVD of $\mathbf{X} = \mathbf{UDV}^\mathsf{T}$, and let $\mathbf{P} = \mathbf{XV}$

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{rr}} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I_n})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \\
&= (\mathbf{VD^2V}^\mathsf{T} + k\mathbf{VV}^\mathsf{T})^{-1}\mathbf{VDU}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}(\mathbf{D^2} + k\mathbf{I_p})^{-1}\mathbf{DU}^\mathsf{T}\mathbf{y} \\
&= \mathbf{V}(\mathbf{D^2} + k\mathbf{I_p})^{-1}\mathbf{P}^\mathsf{T}\mathbf{y}
\end{aligned}
$$

Estimating the Ridge Parameter

We can use very small values of k to study how the OLS estimates would behave if the input data were mildly perturbed.

If we observe large fluctuations in ridge estimates for very small k, such instability would reflect the presence of collinearity in the input variables.

The main challenge in ridge regression is to decide upon the best value of k. To do this, we use cross-validation.

Penalized Least Squares Problem

Consider the general form of the penalized least squares criterion, which can be written as:

$$ESS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda p(\boldsymbol{\beta})$$

where:

- $p(.)$ is a given *penalty function* and
- λ is a *regularization parameter*

Penalized Least Squares Problem

A family of penalty functions $p(.)$, (indexed by $q > 0$), are given by:

$$p_q(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^q$$

This general penalty function bounds the L_q-norm of the parameters in the model as

$$\sum_{j} |\beta_j|^q \leq c$$

Penalized Least Squares Problem

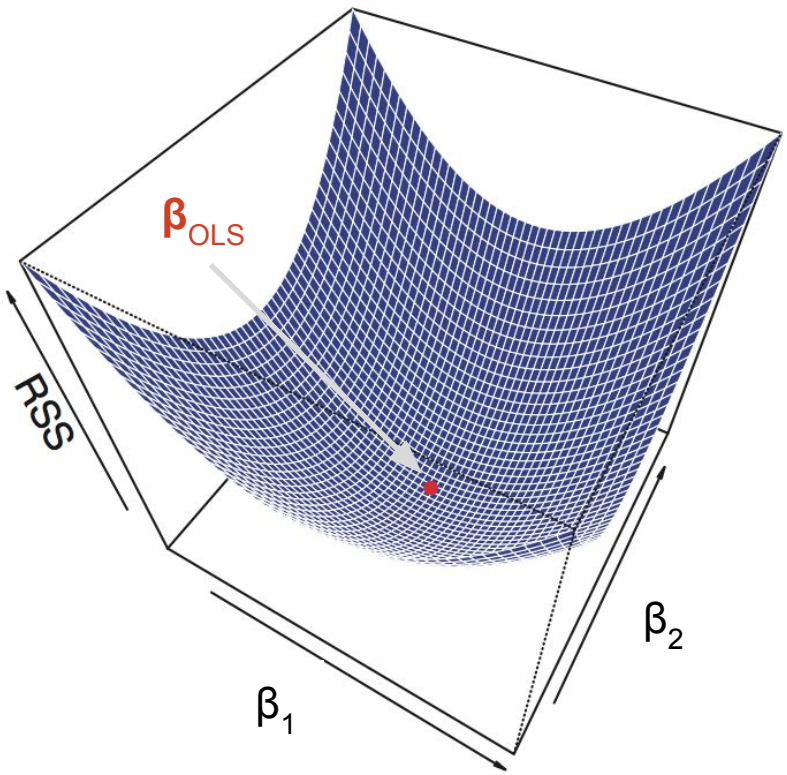
Ridge regression corresponds to $q = 2$, and its corresponding penalty function is a circular disk ($p = 2$) or sphere ($p = 3$), or for a general p, a rotationally invariant hypersphere centered at the origin.

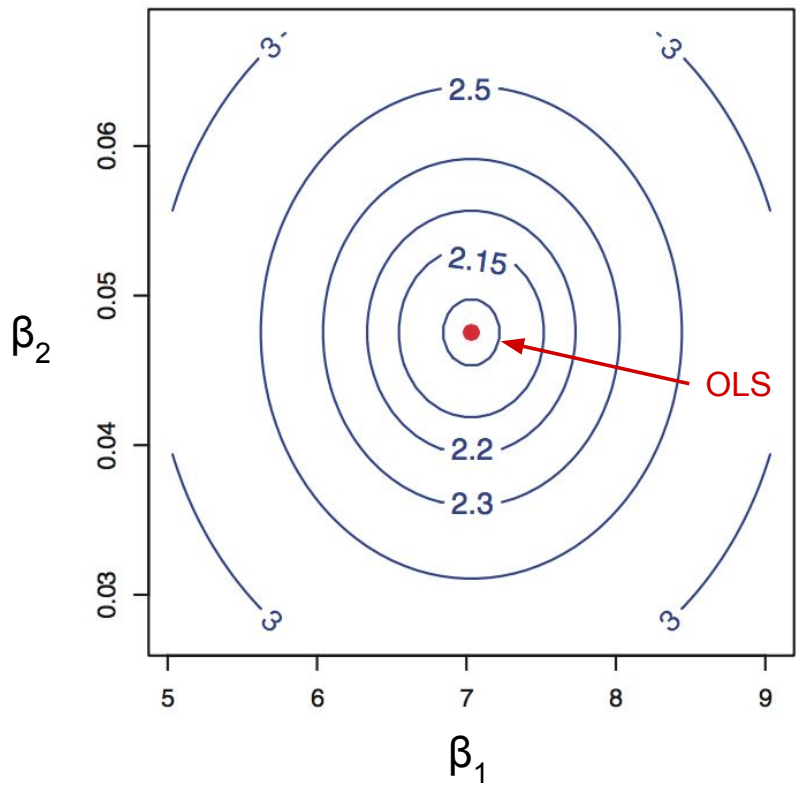
$$p_2(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^2$$

which bounds the L_2-norm of the parameters in the model as

$$\sum_{j} |\beta_j|^2 \leq c$$

RSS Surface with OLS solution





RSS Surface with OLS solution

From the image in the previous slide:

- Three-dimensional plot of the RSS on two variables
- Contour plot—with ellipses—of the RSS
- The red dots correspond to the least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$

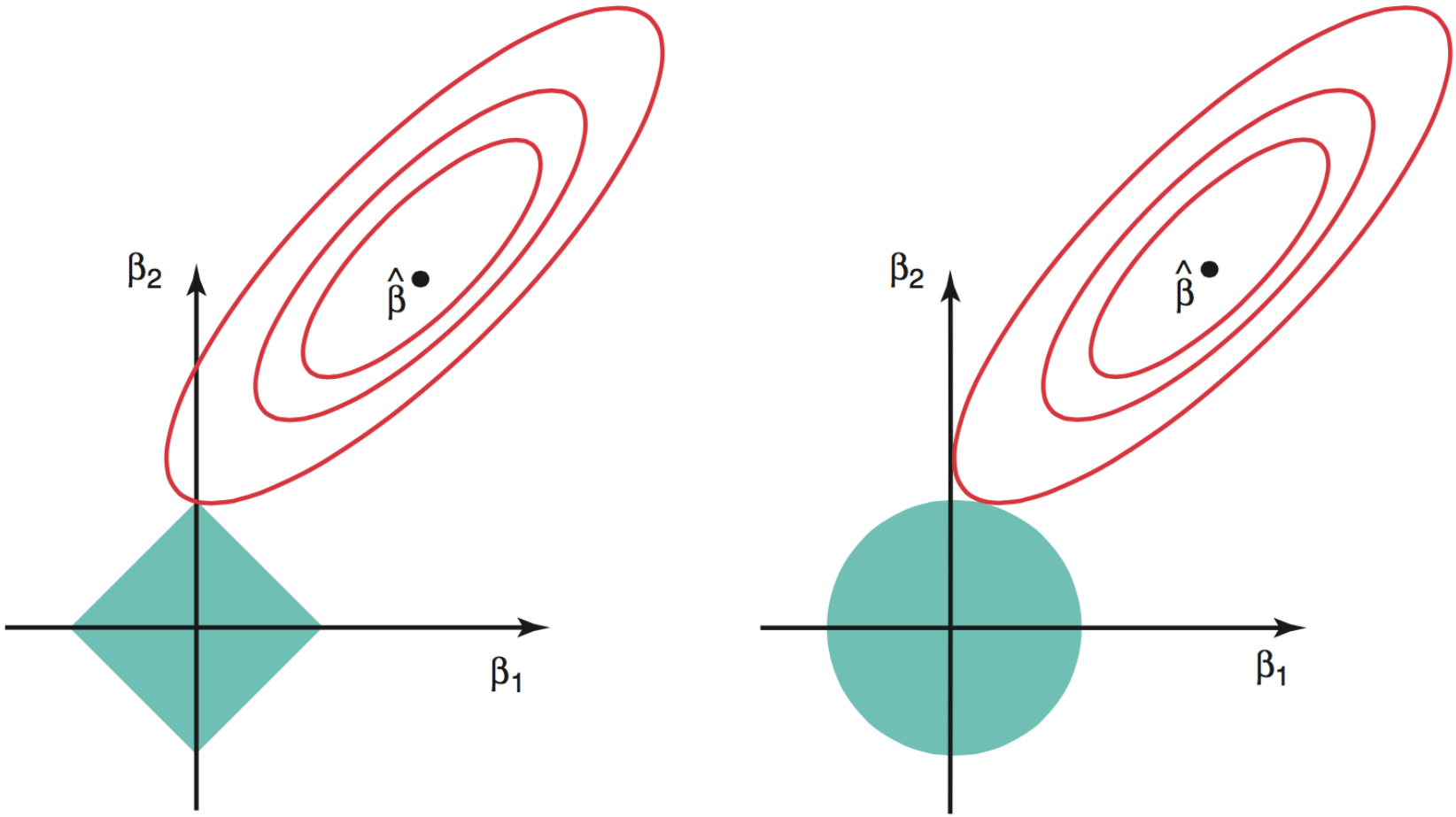


FIGURE 6.7. *Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

Ridge Regression in Practice

Standardized Predictors

In general, the value of β_j depends on the scale in which X_j is measured.
$\hat{\beta}_j$ will be different if the variable is measured in grams or in kilograms.
So, when we calculate the norm of the estimates, it is convenient that each coefficient is treated in the same manner.
One way to treat all coefficients in a "fair" way is to scale them (usually we standardize the predictors)

Standardized Predictors

In general, the value of β_j depends on the scale in which X_j is measured.
$\hat{\beta}_j$ will be different if the variable is measured in grams or in kilograms.
So, when we calculate the norm of the estimates, it is convenient that each coefficient is treated in the same manner.
One way to treat all coefficients in a "fair" way is to scale them (usually we standardize the predictors)

References

- **Statistical Learning from a Regression Perspective** by Richard Berk (2008). *Chapter 2, section 2.3.1: Shrinkage*. Springer.
- **Modern Regression Methods** by Thomas Ryan (1997) *Chapter 12: Ridge Regression*. Wiley.
- **Linear Models with R** by Julian Faraway (2015). *Chapter 11, section 11.3: Ridge Regression*. CRC Press.
- **Modern Multivariate Statistical Techniques** by A.J. Izenman (2008). *Chapter 5, section 5.6.4: Ridge Regression*. Springer.

References (French Literature)

- **Regression avec R** by Cornillon and Matzner-Lober (2011). *Chapter 8: Ridge et Lasso*. Springer.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 17, section 17.5.2: La regression ridge*. Editions Technip, Paris.