MONASH University

Business Analytics

Week 8

Advanced regression

11 September 2017

Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3,7	Tas & David
4	Classification	4	Souhaib
5	Classification	4, 9	Souhaib
6	Model selection and resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	**Advanced regression**	6	Souhaib
	Ridge regression	6	Souhaib
	Lasso regression	6	Souhaib
9	Advanced learning methods	8	Souhaib
	Semester break		
10	Clustering	10	Souhaib
11	Visualization and data wrangling		Souhaib
12	Project presentation		Souhaib

Regression

$$Y = f(X) + \varepsilon$$

where $X = (X_1, \ldots, X_p)$, $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2 | X] = \sigma^2$.

$$m^* = \underset{m \in \mathcal{M}}{\text{argmin}}\ \mathbb{E}[(Y - m(X))^2]$$

Linear regression

$$m(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \to \beta^* = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}[(Y - m(X))^2]$$

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 := (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{ls}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

For a truly linear underlying model, the linear regression has expected test error

$$\sigma^2 + 0 + \frac{\sigma^2 p}{n}.$$

Shortcomings in high-dimension

- The shortcomings don’t even have to do with the linearity assumption!

- It might happen that the columns of $\boldsymbol{X}$ are not linearly independent, so that $\boldsymbol{X}$ is not of full rank. Then $\boldsymbol{X}'\boldsymbol{X}$ is singular and the least squares coefficients are not uniquely defined.

- **Predictive ability**: tradeoff between bias and variance.

- **Interpretative ability**: When the number of variables p is large, we may sometimes seek, for the sake of interpretation, a smaller set of *important variables*

Alternatives

- Dimension Reduction: We project the p predictors into a M-dimensional subspace, where $M < p$. Then these M projections are used as predictors to fit a linear regression model by least squares.

- Subset Selection: We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

- **Shrinkage**: We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.

Best subset selection

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \boldsymbol{I}(\beta_j \neq 0) \leq s, \quad s \geq 0.$$

where $s \geq 0$ is a tuning parameter.

- Need to consider $\binom{p}{s}$ models containing s predictors $\rightarrow$ Computationally infeasible when p and s are large + larger the search space, the higher the chance of overfitting

- Stepwise procedures: forward, backward, etc.

Best subset selection

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \boldsymbol{I}(\beta_j \neq 0) \leq s, \quad s \geq 0.$$

where $s \geq 0$ is a tuning parameter.

- Need to consider $\binom{p}{s}$ models containing s predictors $\rightarrow$ Computationally infeasible when p and s are large + larger the search space, the higher the chance of overfitting

- Stepwise procedures: forward, backward, etc.

Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq s$$

where $s \geq 0$ is a tuning parameter.

- $s = 0$? $\rightarrow$ $\hat{\beta}^{\mathrm{R}} = (0, \ldots, 0)$
- $s = \infty$? $\rightarrow$ $\hat{\beta}^{\mathrm{R}} = \hat{\beta}^{\mathrm{ls}}$ (least squares)
- $s \in (0, \infty)$ $\rightarrow$ tradeoff

Ridge regression

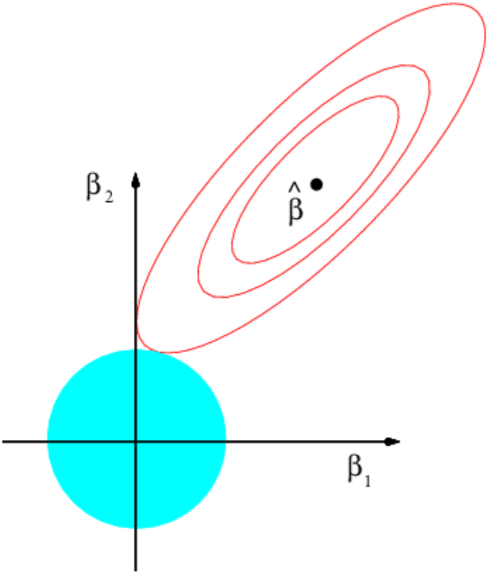
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq s$$

where $s \geq 0$ is a tuning parameter.

- $s = 0$? → $\hat{\beta}^{\mathrm{R}} = (0, \ldots, 0)$
- $s = \infty$? → $\hat{\beta}^{\mathrm{R}} = \hat{\beta}^{\mathrm{ls}}$ (least squares)
- $s \in (0, \infty)$ → tradeoff

Ridge regression: geometry



Ridge regression: another formulation

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **tuning parameter**.

- $\lambda = 0$? $\rightarrow$ $\hat{\beta}^{\mathrm{R}} = \hat{\beta}^{\mathrm{ls}}$ (least squares)
- $\lambda = \infty$? $\rightarrow$ $\hat{\beta}^{\mathrm{R}} = (0, \ldots, 0)$
- $\lambda \in (0, \infty)$ $\rightarrow$ tradeoff

- We can solve it by *data augmentation*
- $\hat{\beta}^{\mathrm{R}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}$

Ridge regression: another formulation

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **tuning parameter**.

- $\lambda = 0$? → $\hat{\beta}^{\text{R}} = \hat{\beta}^{\text{ls}}$ (least squares)
- $\lambda = \infty$? → $\hat{\beta}^{\text{R}} = (0, \dots, 0)$
- $\lambda \in (0, \infty)$ → tradeoff

- We can solve it by *data augmentation*
- $\hat{\beta}^{\text{R}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}'\boldsymbol{y}$

Ridge regression: another formulation

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **tuning parameter**.

- $\lambda = 0$? → $\hat{\beta}^{\mathrm{R}} = \hat{\beta}^{\mathrm{ls}}$ (least squares)
- $\lambda = \infty$? → $\hat{\beta}^{\mathrm{R}} = (0, \ldots, 0)$
- $\lambda \in (0, \infty)$ → tradeoff

- We can solve it by *data augmentation*
- $\hat{\beta}^{\mathrm{R}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}$

A Simple special case

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \rightarrow \hat{\boldsymbol{\beta}}^{\text{ls}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

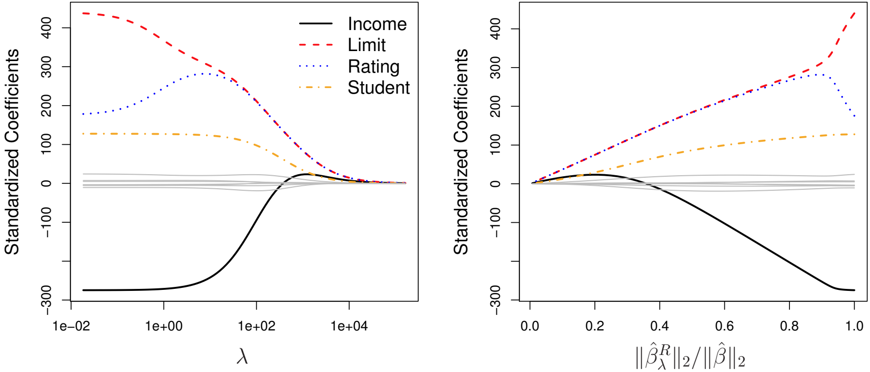
Suppose $n = p$ and $\boldsymbol{X} = \boldsymbol{I}_n = \boldsymbol{I}_p$, then

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{p}(y_j - \beta_j)^2 \quad \rightarrow \quad \hat{\beta}_j^{\text{ls}} = y_j$$

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \rightarrow \quad \hat{\beta}_j^{R} = \frac{y_j}{(1+\lambda)} = \frac{\hat{\beta}_j^{ls}}{(1+\lambda)}$$

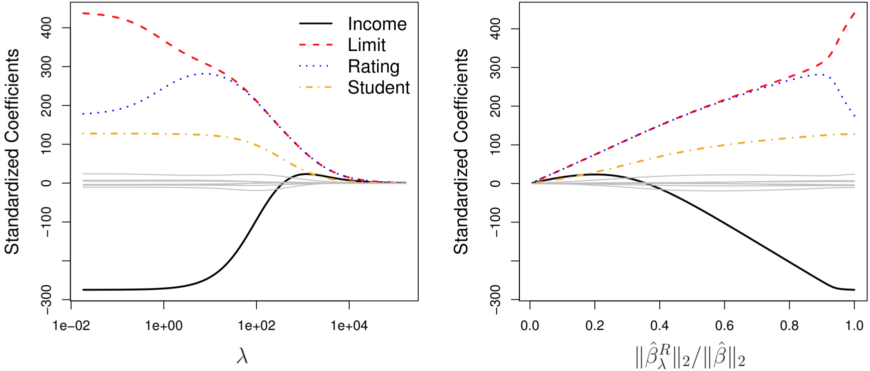
- This illustrates the essential feature of ridge regression: shrinkage. Introducing bias but reducing the variance.
- Each least squares coefficient estimate is shrunken by the **same proportion**.

Ridge regression: example



While the ridge coefficient estimates tend to **decrease in aggregate** as λ increases, individual coefficients, such as rating and income, may **occasionally increase** as λ increases.

Ridge regression: example

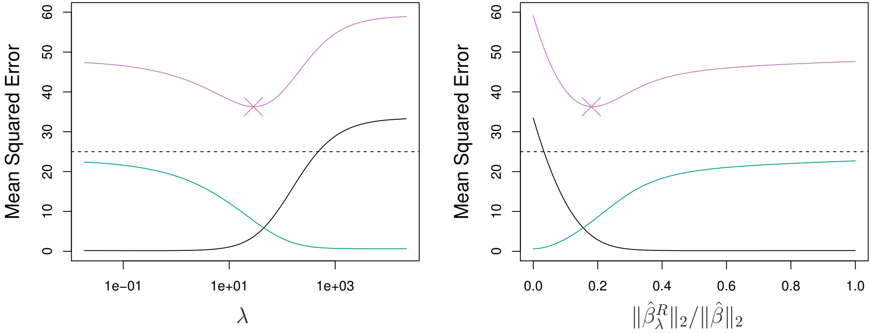


While the ridge coefficient estimates tend to **decrease in aggregate** as λ increases, individual coefficients, such as rating and income, may **occasionally increase** as λ increases.

A note on scaling

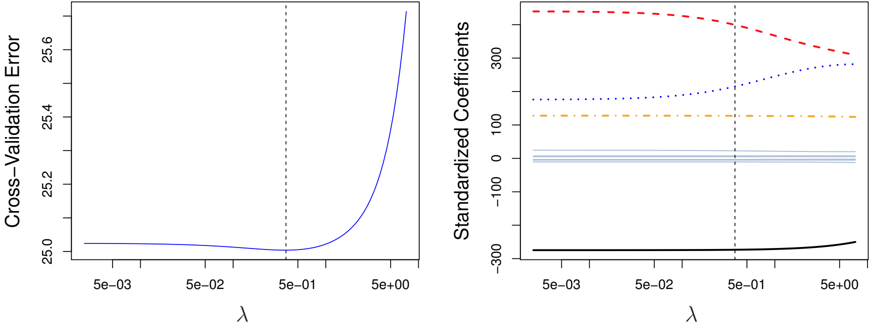
- Standard least squares coefficient estimates are **scale equivariant**
  - multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$
  - regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- The ridge regression coefficient estimates **can change substantially** when multiplying a given predictor by a constant
  - This is due to the sum of squared coefficients term in the ridge regression formulation
  - If we use thousands of dollars instead of dollars, it will **not** simply cause the ridge estimate to change by a factor of $1,000$

Ridge Regression vs Least Squares



Squared bias (black), variance (green), and test mean squared error (purple)

Selecting the Tuning Parameter



Ridge regression bias and

If $\boldsymbol{R} = \boldsymbol{X}'\boldsymbol{X}$:

$$\begin{aligned}\beta_\lambda^R &= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{R} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{R} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{R}\hat{\beta}^{ls} = [\boldsymbol{R}(\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})]^{-1}\boldsymbol{R}\hat{\beta}^{ls} \\ &= (\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})\hat{\beta}^{ls}\end{aligned}$$

If $\boldsymbol{W}_\lambda = (\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})$:

$$E[\beta_\lambda^R] = E[\boldsymbol{W}_\lambda \hat{\beta}^{ls}] = \boldsymbol{W}_\lambda \beta \overset{\lambda \neq 0}{\neq} \beta$$

$$\text{Var}(\beta_\lambda^R) = \text{Var}(\boldsymbol{W}_\lambda \hat{\beta}^{ls}) = \boldsymbol{W}_\lambda \text{Var}(\hat{\beta}^{ls})\boldsymbol{W}_\lambda' \preceq \text{Var}(\hat{\beta}^{ls})$$

Singular Value Decomposition

Singular Value Decomposition (SVD)

$$\boldsymbol{X} = \boldsymbol{UDV}'$$

- $\boldsymbol{X}$ is $n \times p$ matrix
- $\boldsymbol{U}$ is $n \times r$ matrix with orthonormal columns ($\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$)
- $\boldsymbol{D}$ is $r \times r$ diagonal matrix with diagonal entries $d_1, \geq d_2 \geq \cdots \geq d_p \geq 0$ called the singular values of $\boldsymbol{X}$.
- $\boldsymbol{V}$ is $p \times r$ matrix with orthonormal columns ($\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$).

Note: $\boldsymbol{XV} = \boldsymbol{UD}$

Least squares regression and SVD

If $\boldsymbol{X} = \boldsymbol{UDV}'$, then

$$\begin{aligned}\hat{\beta}^{\text{ls}} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^{'}\boldsymbol{y} \\ &= \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{D}\boldsymbol{U}^{'}\boldsymbol{y}\end{aligned}$$

$$\begin{aligned}\hat{\beta}^{R}_{\lambda} &= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{'} + \lambda \boldsymbol{V}\boldsymbol{V}^{'})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{'}\boldsymbol{y} \\ &= \left(\boldsymbol{V}(\boldsymbol{D}^2 + \lambda)\boldsymbol{V}^{'}\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{'}\boldsymbol{y} \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda)^{-1}\boldsymbol{V}^{'}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{'}\boldsymbol{y} \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda)^{-1}\boldsymbol{D}\boldsymbol{U}^{'}\boldsymbol{y}\end{aligned}$$

Ridge regression and SVD

$$\hat{\boldsymbol{y}}^{\text{ls}} = \boldsymbol{X}\hat{\beta}^{\text{ls}} = \boldsymbol{U}\boldsymbol{U}'\boldsymbol{y}$$

$\boldsymbol{U}'\boldsymbol{y}$ are the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$

$$\hat{\boldsymbol{y}}^{\text{R}}_{\lambda} = \boldsymbol{X}\hat{\beta}^{\text{R}}_{\lambda} = \boldsymbol{X}\boldsymbol{V}(\boldsymbol{D}^2 + \lambda)^{-1}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{y} = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{y}$$

$$= \boldsymbol{U}\,\text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right)\boldsymbol{U}'\boldsymbol{y}$$

- Since $\lambda \geq 0$, we have $d_j^2/(d_j^2 + \lambda) \leq 1$. A shrinkage is applied to the coordinates by the factors $d_j^2/(d_j^2 + \lambda)$.
- A greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.
- The derived variable $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{v}_j = \boldsymbol{u}_j d_j$ is the jth PC of $\boldsymbol{X}$. We project $\boldsymbol{y}$ onto these components with large d_j, and shrinks the coefficients of low-variance components.

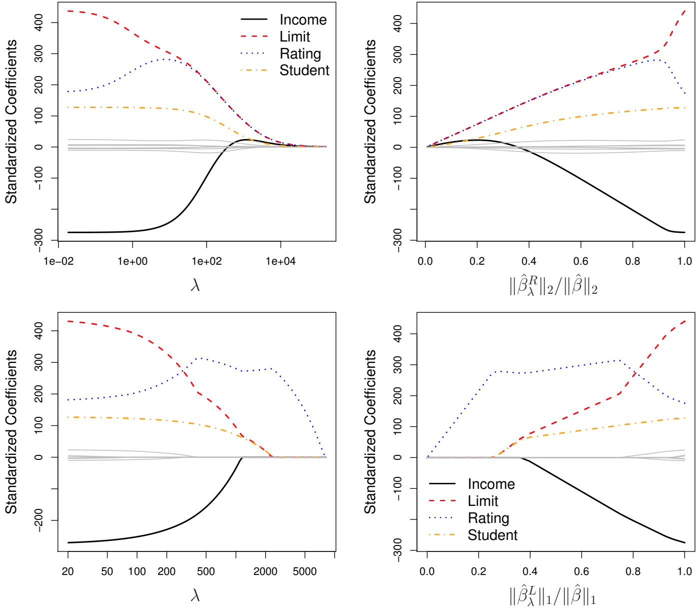
Summary

$$\hat{\boldsymbol{y}}^{\text{ls}} = \boldsymbol{U}\boldsymbol{U}'\boldsymbol{y}$$

$$\hat{\boldsymbol{y}}^{\text{R}}_{\lambda} = \boldsymbol{U}\,\text{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\boldsymbol{U}'\boldsymbol{y}$$

$$\hat{\boldsymbol{y}}^{\text{PCR}}_{k} = \boldsymbol{U}\,\text{diag}\left(\underbrace{1,\ldots,1}_{k},\underbrace{0,\ldots,0}_{p-k}\right)\boldsymbol{U}'\boldsymbol{y}$$

Another shrinkage method



LASSO regression

LASSO: Least Absolute Shrinkage and Selection Operator

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s$$

- $s = 0$? → $\hat{\beta}^{R} = (0, \ldots, 0)$
- $s = \infty$? → $\hat{\beta}^{R} = \hat{\beta}^{ls}$ (least squares)
- $s \in (0, \infty)$ → tradeoff

LASSO regression

LASSO: Least Absolute Shrinkage and Selection Operator

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s$$

- $s = 0$? → $\hat{\boldsymbol{\beta}}^{\text{R}} = (0, \ldots, 0)$
- $s = \infty$? → $\hat{\boldsymbol{\beta}}^{\text{R}} = \hat{\boldsymbol{\beta}}^{\text{ls}}$ (least squares)
- $s \in (0, \infty)$ → tradeoff

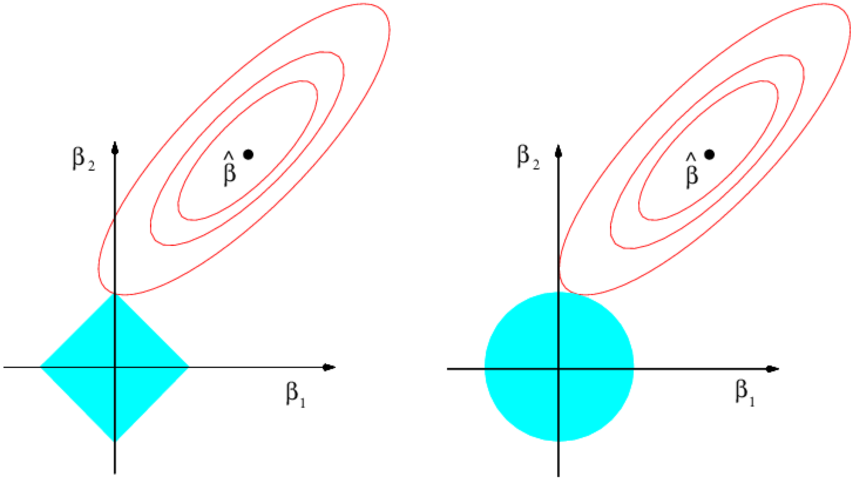
LASSO: another formulation

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is a **tuning parameter**.

- $\lambda = 0$? → $\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \hat{\boldsymbol{\beta}}^{\mathrm{ls}}$ (least squares)
- $\lambda = \infty$? → $\hat{\boldsymbol{\beta}}^{\mathrm{R}} = (0, \ldots, 0)$
- $\lambda \in (0, \infty)$ → tradeoff

LASSO vs Ridge: geometry



Sparsity

We shall say that a signal $\boldsymbol{x} \in R^n$ is **sparse**, when most of the entries of x **vanish**. Formally, we shall say that a signal is s-sparse if it has **at most** s **nonzero entries**. One can think of an s-sparse signal as having only s degrees of freedom.

- L_q regularization with $q > 1$ does not provide sparse estimate
  $\rightarrow$ e.g. ridge regression
- For $q < 1$, the solutions are sparse but the problem is **not convex** and this makes the optimisation very challenging computationally.
- The value $q = 1$ is the smallest value that yields a **convex problem**.

The bet on sparsity principle

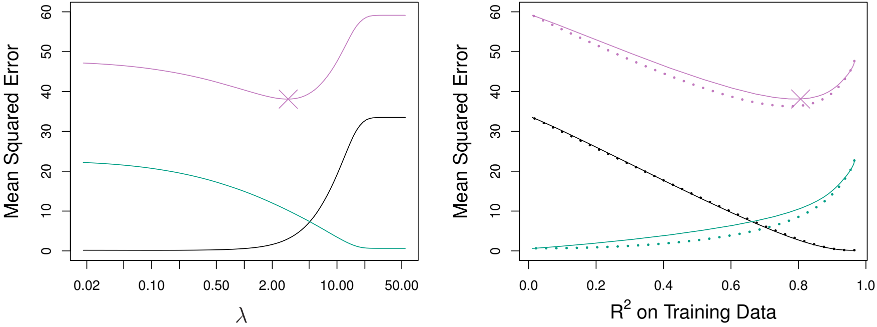
If $p >> N$ and the true model **is sparse**, so that only $k < N$ parameters are actually nonzero in the true underlying model, then it turns out that we can estimate the parameters effectively, using the lasso and related methods.

if $p >> N$, and the true model **is not sparse**, then the number of samples N is too small to allow for accurate estimation of the parameters (The amount of information per parameter is N/p)

Use a procedure that does well in sparse problems, since no procedure does well in dense problems

Lasso vs ridge regression

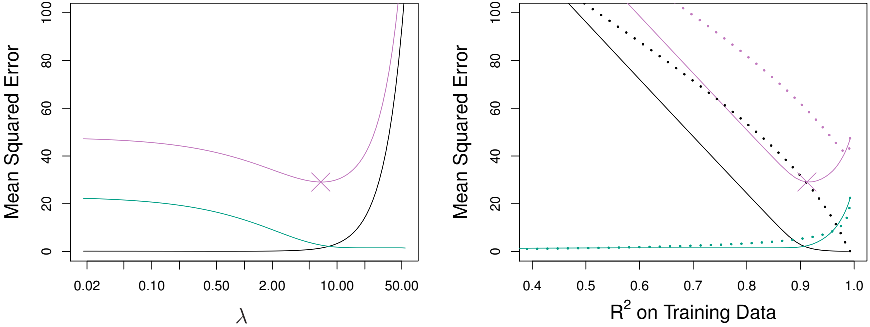
A simulated data set containing p = 45 predictors and n = 50 observations where **all 45 predictors are related to the response**.



Left: Lasso. **Right**: Lasso (solid) and ridge (dashed).

Lasso vs ridge regression

Now the response is a function of **only 2 out of 45 predictors**.



Left: Lasso. **Right**: Lasso (solid) and ridge (dashed).

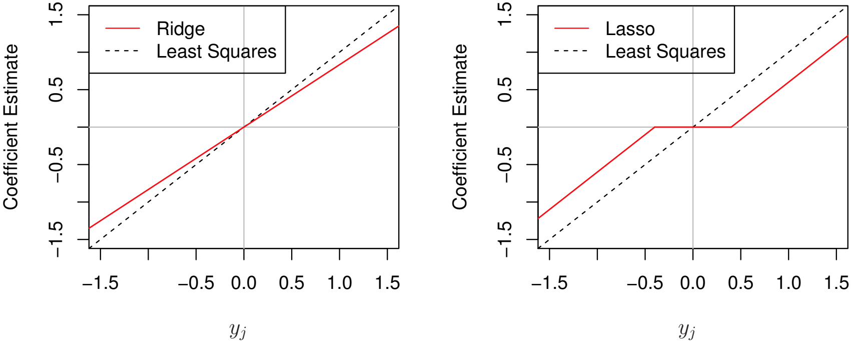
A Simple special case: lasso

$$\underset{\beta}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2}; \\ 0 & \text{if } |y_j| \leq \frac{\lambda}{2}. \end{cases}$$

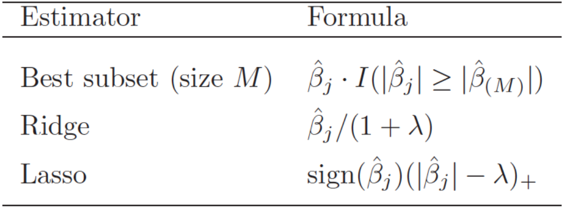
The lasso shrinks each least squares coefficient towards zero by a **constant amount**, $\lambda/2$. The least squares coefficients that are less than $\lambda/2$ in absolute value are **shrunken entirely to zero**.

A Simple special case

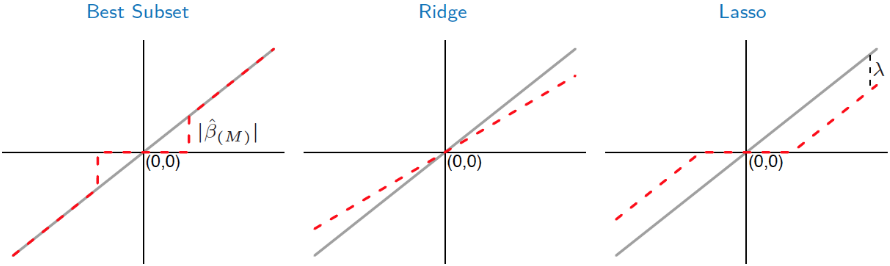


A Simple special case

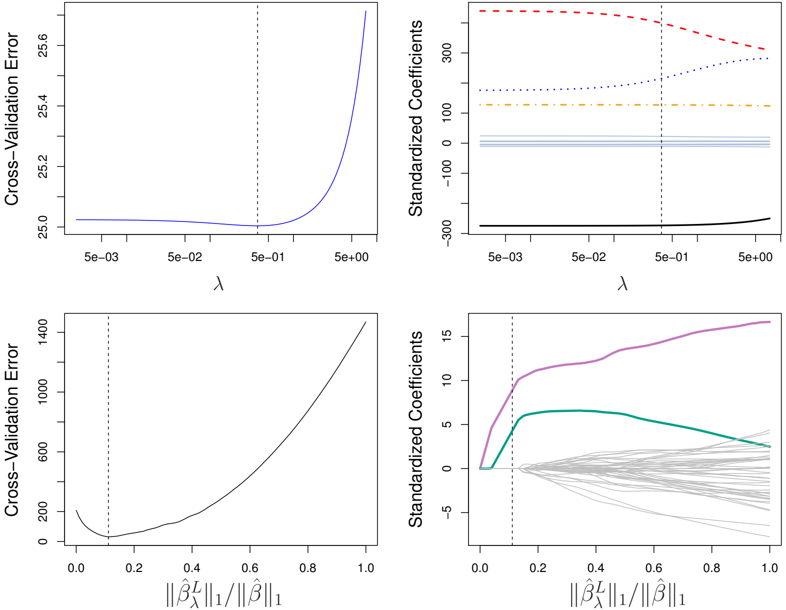
$\hat{\beta}_j$ (OLS estimate) and $\hat{\beta}_{(M)}$ (Mth largest coefficient)



Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Selecting the Tuning Parameter



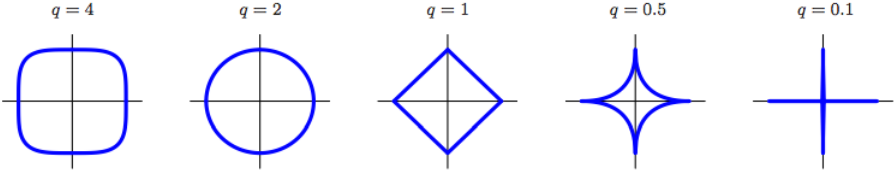
q-norm

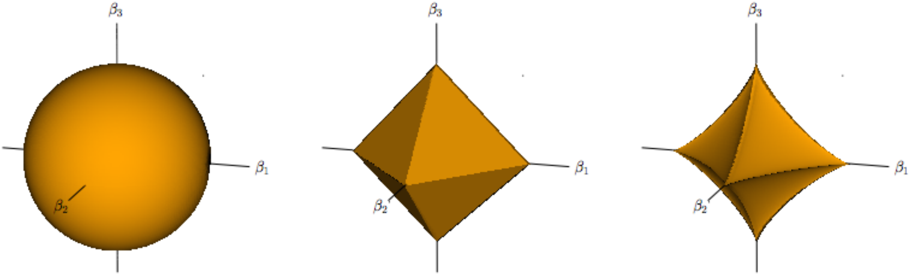
Let $q \geq 1$ be a real number. The q-norm of $\boldsymbol{x} = (x_1, \ldots, x_p)$ is given by

$$\|\boldsymbol{x}\|_q = \left(\sum_{j=1}^{p} |x_j|^q \right)^{1/q}$$

- $q = 1$: L_1 norm
- $q = 2$: L_2 norm, Euclidean norm
- $q = \infty$: L_∞ norm, uniform norm: $\|x\|_\infty = \max\{|x_1|, \ldots, |x_p|\}$.

q-norm


$q = 4$
$q = 2$
$q = 1$
$q = 0.5$
$q = 0.1$


β_3
β_1
β_2
β_3
β_1
β_2
β_3
β_1
β_2