Business Analytics

Week 10
Advanced regression

11 October 2016

Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3	Souhaib
4	Resampling	5	Souhaib
5	Dimension reduction	6,10	Souhaib
6	Visualization		Di
7	Visualization		Di
8	Classification	4,8	Di
9	Classification	4,9	Di
	-		
10	Classification	8	Souhaib
11	Advanced regression	6	Souhaib
12	Clustering	10	Souhaib

Best subset selection

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

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

Need to consider $\binom{p}{s}$ models containing s predictors
$\rightarrow$ Computationally infeasible when p is large

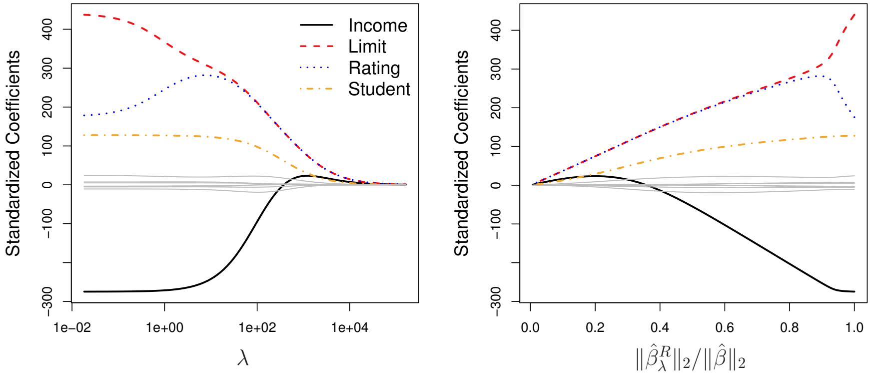
Ridge regression

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

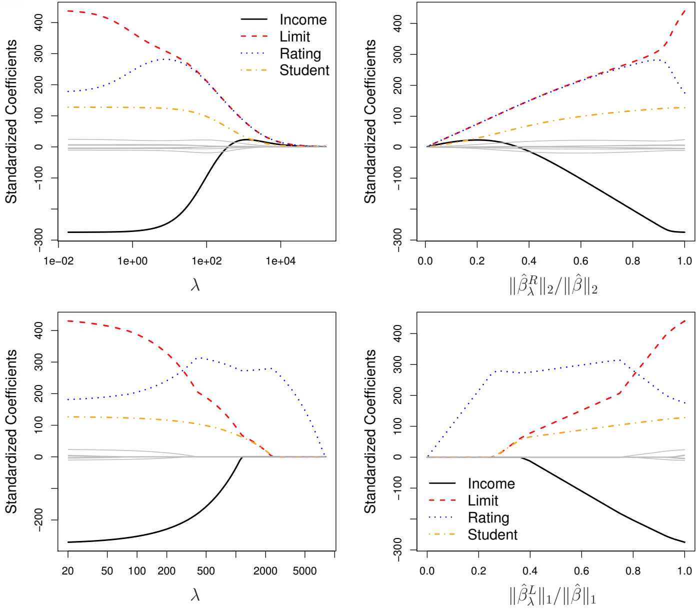
$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq s$$

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

Ridge regression: example



Another shrinkage method



Lasso regression

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

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

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

Lasso regression

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

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

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

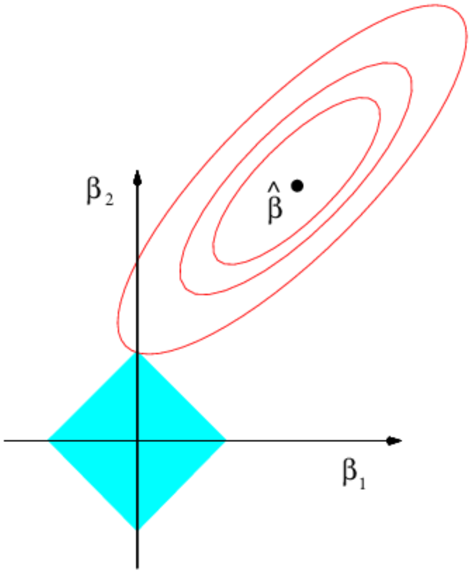
Lasso regression: another formulation

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

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

Lasso regression: geometry



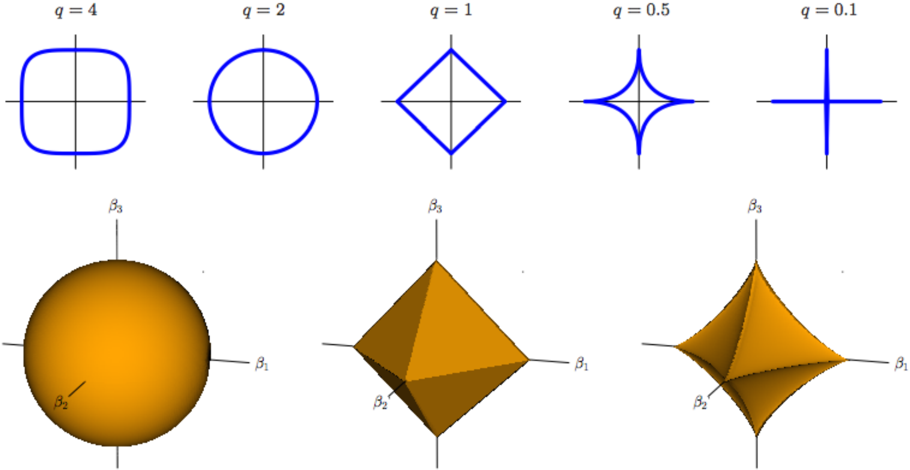
q-norm

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

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**.

$\rightarrow q = 1$: LASSO (least absolute shrinkage and selection operator)

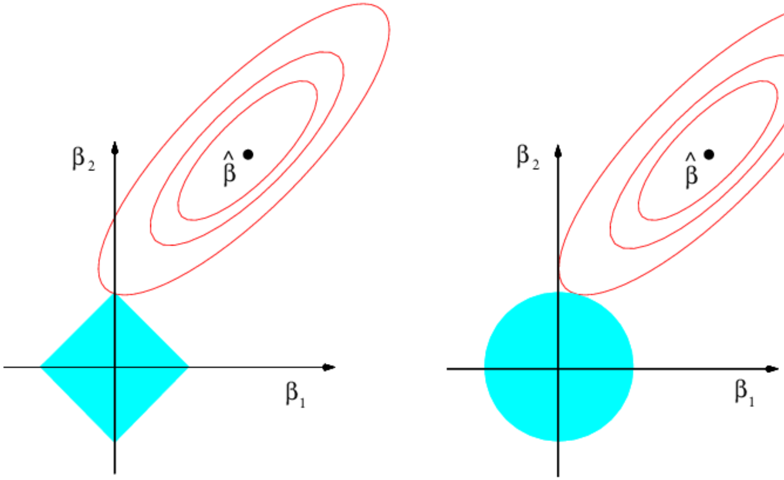
The bet on sparsity principle

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)

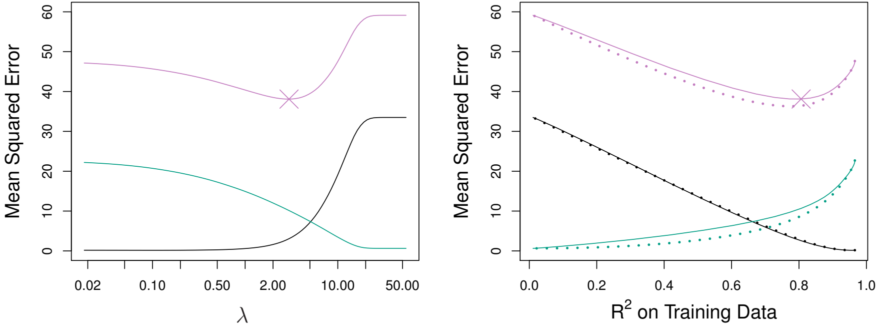
Use a procedure that does well in sparse problems, since no procedure does well in dense problems

Lasso vs ridge regression



Lasso vs ridge regression

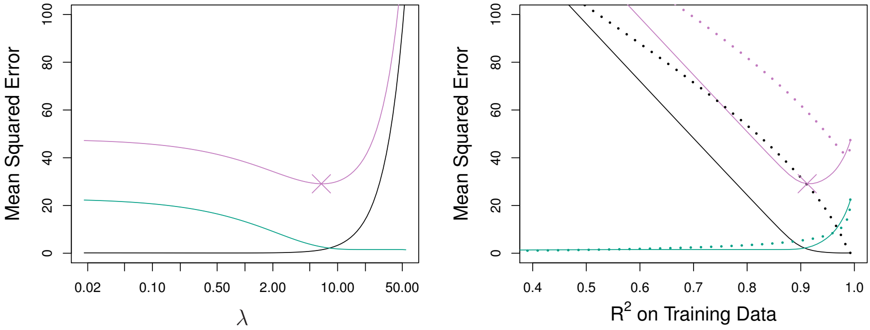
A simulated data set containing p = 45 predictors and n = 50 observations where **all 45 predictors are related to the response**.



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: least squares

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

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

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

A Simple special case: ridge

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

$$\hat{\beta}_j^R = \frac{y_j}{(1+\lambda)}$$

In ridge regression, each least squares coefficient estimate is shrunken by the **same proportion**.

A Simple special case: lasso

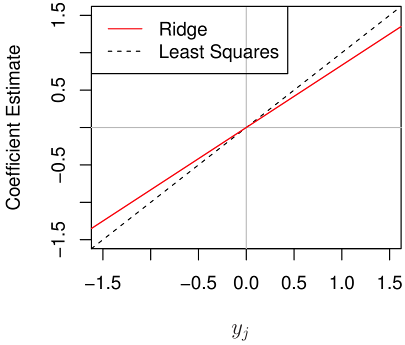
$$\underset{\beta}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

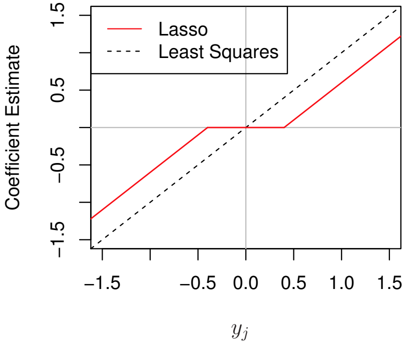
$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2}; \\ 0 & \text{if } |y_j| \le \frac{\lambda}{2}. \end{cases}$$

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

$$\lambda = 1$$





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	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$