MONASH University

Business Analytics

Week 10
Advanced regression

10 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

Regression

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

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

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

Linear regression

$$m(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}[(Y - m(X))^2]$$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

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

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

- **Subset Selection**
- **Dimension Reduction**
- **Shrinkage**

Best subset selection

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

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

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

Best subset selection

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

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

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

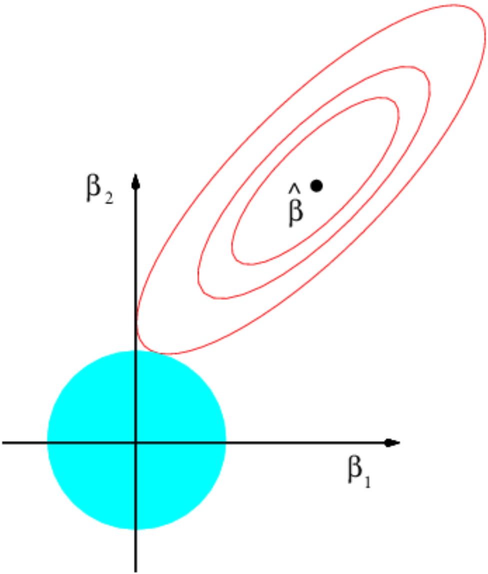
Ridge regression

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

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

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

Ridge regression: 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**.

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

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**.

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

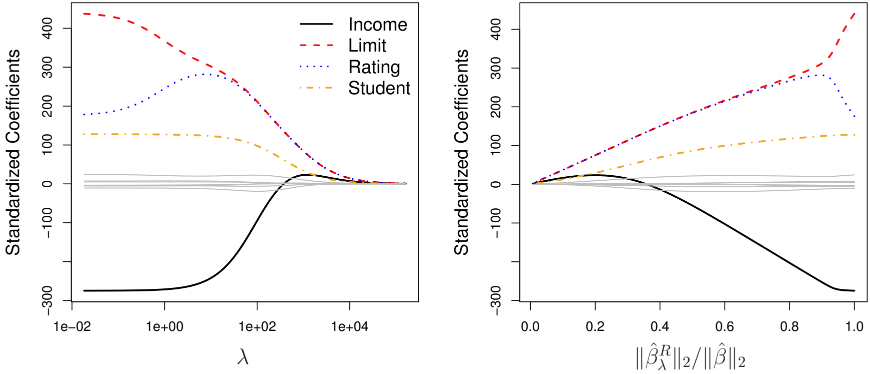
p-norm

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

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

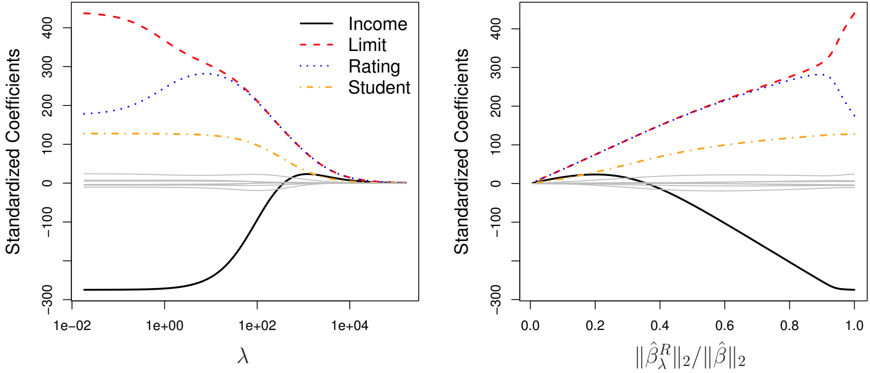
- $p = 1$: L_1 norm
- $p = 2$: L_2 norm, Euclidean norm
- $p = \infty$: L_∞ norm, uniform norm: $\|x\|_\infty = \max\{|x_1|, \ldots, |x_p|\}$.

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

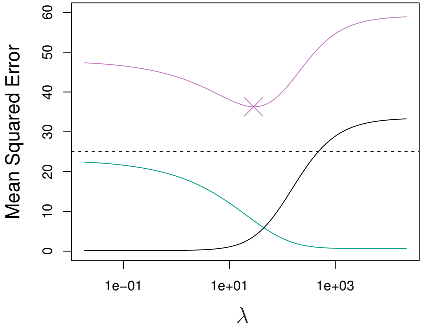


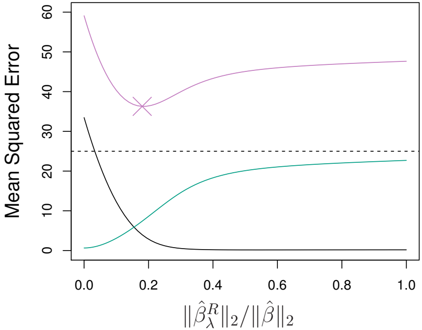
While the ridge coefficient estimates tend to **decrease in aggregate** as λ increases, individual coefficients, such as rating and income, may **occasionally increase** as λ increases.

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)

Ridge regression bias

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

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

$$\begin{aligned}
E[\beta_\lambda^R] &= E[(\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})\hat{\beta}^{ls}] \\
&= (\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})\beta \\
&\overset{\lambda \neq 0}{\neq} \beta
\end{aligned}$$

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

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

Note that $\boldsymbol{U}^{'}\boldsymbol{y}$ are the coordiantes of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$.

Ridge regression and SVD

$$\begin{aligned}\hat{\boldsymbol{y}}^{\mathrm{R}} = \boldsymbol{X}\hat{\beta}^{\mathrm{R}} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{y} \\ &= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j'\boldsymbol{y}\end{aligned}$$

where the $\boldsymbol{u_j}$ are the columns of $\boldsymbol{U}$. Note that since $\lambda \geq 0$, we have $d_j^2/(d_j^2 + \lambda) \leq 1$.

Ridge regression shrinks the coordinates by the factors $d_j^2/(d_j^2 + \lambda)$. This means that a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.

Selecting the Tuning Parameter