Ch3. Linear Regression

ST4240, 2016/2017

Version 0.1

Alexandre Thiéry

Department of Statistics and Applied Probability

Outline

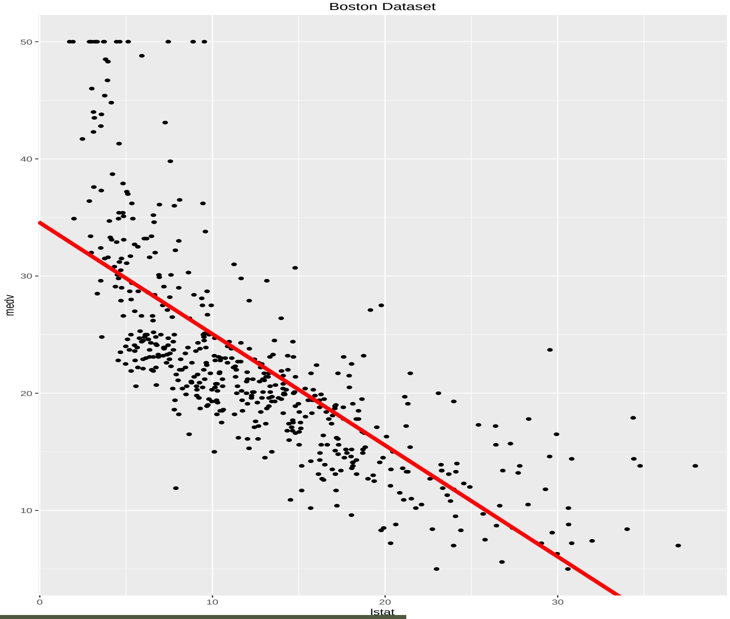
Linear Regression

- Response $y \in \mathbb{R}$
- Covariates (explanatory variables) $x = (x_0, x_1, \ldots, x_p) \in \mathbb{R}^p$
- Linear regression model

$$y = \beta_0 + \sum_{i=1}^{p} x_i \, \beta_i \; + \; (\text{noise})$$

- Training examples $\{(y_i, x_i)\}_{i=1}^n$ with $x_i = (x_{i,1}, \ldots, x_{i,p})$
- It is common to set $x_{i,0} = 1$ for the intercept β_0.

Linear Regression



One dimensional example

- Coefficients $\beta = (\beta_0, \beta_1)$ minimise

$$\textbf{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 \, x_i])^2$$

- **[Exercise]** Setting the partial derivative of the function yields that

$$\begin{cases} \beta_0 + \beta_1 \, \overline{x} & = \overline{y} \\ \beta_0 \, \overline{x} + \beta_1 \, \overline{xx} & = \overline{xy} \end{cases}$$

  with $\overline{x} = n^{-1} \sum x_i$ and $\overline{y} = n^{-1} \sum y_i$ and $\overline{xx} = n^{-1} \sum x_i^2$ and $\overline{xy} = n^{-1} \sum x_i y_i$.

General case

- Training examples $\{(y_i, x_i)\}_{i=1}^{n}$.
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ minimises

$$\textbf{RSS}(\beta) = \sum_{i=1}^{n} \{y_i - [\beta_0 + \beta_1\, x_{i,1} + \ldots + \beta_p\, x_{i,p}]\}^2 .$$

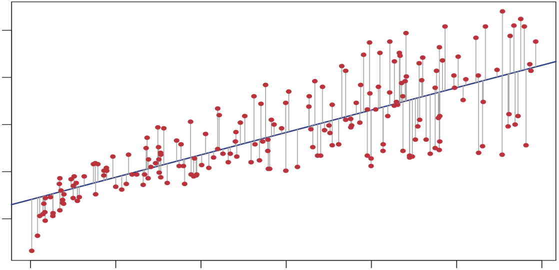
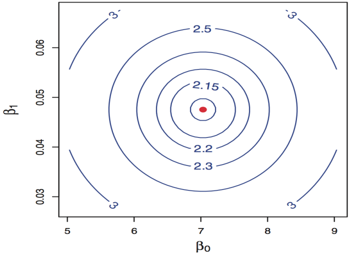
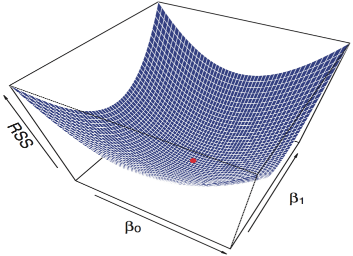


Figure: Residual Sum of Square (**RSS**)

- Ordinary Least Square (OLS) estimate $\widehat{\beta}$:

$$\widehat{\beta} = \mathbf{argmin} \left\{ \mathbf{RSS}(\beta) \ : \ \beta \in \mathbb{R}^{p+1} \right\}. \tag{1}$$





OLS: closed form expression

- One could numerically optimise **RSS**.

$$\mathbf{RSS}(\beta) = \|y - X\,\beta\|^2 = \langle y - X\,\beta, y - X\,\beta \rangle$$

- There is a closed form expression

$$\begin{aligned} \partial_\beta \mathbf{RSS} &= -2X^T\,(y - X\,\beta) \\ \partial^2_{\beta,\beta} \mathbf{RSS} &= 2\,X^T\,X. \end{aligned}$$

- **[Exercise]** the matrix $X^T\,X$ is positive semi-definite; if X has full rank, it is positive definite.
- **[Exercise]** if X has full rank,

$$\widehat{\beta} = (X^T\,X)^{-1}\,X^T\,y$$

- we have **RSS** $= \sum_i (y_i - \widehat{y}_i)^2$ where the fitted values are

$$\widehat{y} = \widehat{H}\, y \qquad \text{with} \qquad \widehat{H} \equiv X\,(X^T\,X)^{-1}\,X^T.$$

- the hat matrix $\widehat{H}$ is a projection

$$\widehat{H}^2 = \widehat{H}.$$

- $\widehat{\beta}$ well defined only if $X^T\,X \in \mathbb{R}^{p+1,p+1}$ is invertible
- [**Exercise**] it is never the case if $n \leq p$ i.e. when there is more covariates than observations

Boston Dataset: prediction versus Truth

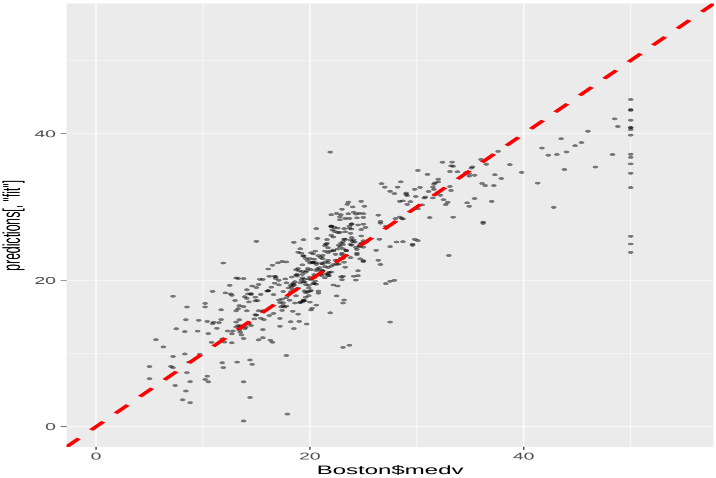


Figure: Residual Sum of Square (**RSS**)

Boston Dataset: Prediction Intervals

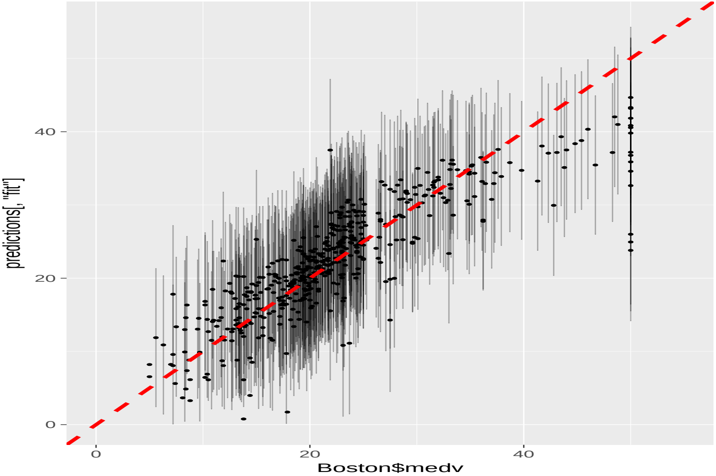


Figure: Residual Sum of Square (**RSS**)

Some remarks

The OLS estimate is not necessarily a sensible thing to consider if:

- the relationship covariates / responses is not $\approx$ linear
- the covariates are highly correlated
- the variance of the noise is not (approximately) constant
- high correlation between the error terms
- presence of outliers (square loss is highly sensitive to outliers)

Further remarks

- Sometimes useful to **transform** the covariate and/or response
- **feature engineering**: polynomials, interaction, fancier ideas..
- It is a good idea to check whether the model is **well-calibrated** by computing how many times the true data falls inside the prediction interval.
- **Categorical data** need to be taken care of!
- One can minimize a **weighted RSS** if necessary.
- **Prediction Intervals $\neq$ Confidence Intervals**

Properties of $\widehat{\beta}$

- Consider the Gaussian model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p \, x_{i,p} + \mathbf{N}\left(0, \sigma^2\right)$$

  for unknown parameter $\beta_\star = (\beta_0, \ldots, \beta_p)$.
- **[Exercise]** $\widehat{\beta}$ is an unbiased estimate of $\beta_\star$ and

$$\widehat{\beta} \sim \mathbf{N}\left(\beta_\star, \sigma^2 \, (X^T X)^{-1}\right).$$

- if $(X^T X)$ is almost singular then the variance of $\widehat{\beta}$ is very high and thus $\widehat{\beta}$ is an unreliable estimate of β. This is for example the case if:
  - the covariates are highly correlated.
  - the number of covariate is if the same order as the number of observations.

Instability

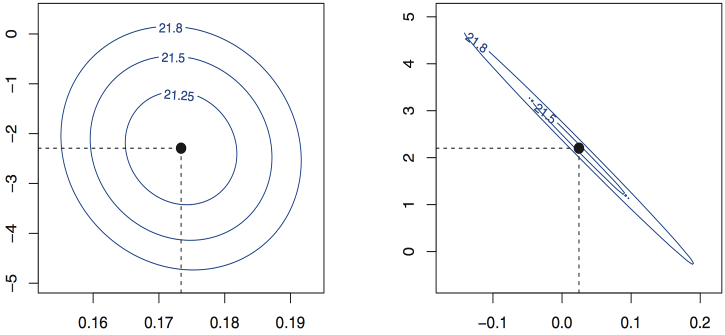


Figure: Left: no correlation. Right: high correlation

MLE connection

- Consider the Gaussian model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p \, x_{i,p} + \mathsf{\textbf{N}} \left(0, \sigma^2\right)$$

- the log-likelihood is given by

$$\ell(\beta) = -\frac{1}{2\sigma^2} \, \|y - X\beta\|^2 + (\text{irrelevant additive constants})$$

- [**Exercise**] the least square estimate is also the MLE.

Issues with the OLS estimate

- If $p \geq n$ there is not unique solution for the minimisation of the **RSS** and the OLS estimate is not well defined.
- this large p small n situation is extremely important in practice.
- if $p \leq n - 1$ but still large, even if the OLS is well-defined, it may be extremely unstable.
- When p is large, one may want to find a solution with as many zero coefficient as possible. Typically, all the coefficients of $\widehat{\beta}_{OLS}$ are non-zero.

Outline

1 Ordinary Least Square theory

2 Shrinkage methods

Penalization

- Directly penalization of the size of the coefficients
- For a regularization parameter $\lambda > 0$

$$\widehat{\beta}(\lambda) \; = \; \mathbf{argmin}\Big\{ \; \textsf{\textbf{RSS}}(\beta) + \lambda \times \Omega(\beta) \; : \; \beta \in \mathbb{R}^{p+1} \Big\}.$$

  The quantity $\Omega(\beta)$ penalises large coefficients
- Estimate $\widehat{\beta}(\lambda)$ is similar to the OLS, with the important difference that a penalization term $\lambda \times \Omega(\beta)$ is added.
- $\lambda > 0$ quantifies the amount of regularization.

LASSO and Ridge Penalization

$$\Omega_{\text{Ridge}}(\beta) \equiv \sum_{j=1}^{p} \beta_j^2 \equiv \|\beta\|_2^2$$

$$\Omega_{\text{Lasso}}(\beta) \equiv \sum_{j=1}^{p} |\beta_j| \equiv \|\beta\|_1.$$

- Recall the definition of the p-norm

$$\|v\|_p = \Big(|x_1|^p + \ldots + |x_n|^p\Big)^{1/p}$$

Normalization

- typically, the intercept coefficient β_0 is not penalized; this is because we do not want the procedures to be dependent on the location of the covariates
- shrinkage procedures do depend on the scale of the covariates
- In practice, the responses/covariates are centred/normalized

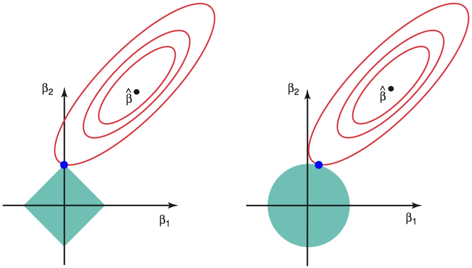
$$x_i \mapsto \frac{x_i - \bar{x}_i}{\widehat{\sigma}(x)} \qquad \text{and} \qquad y \mapsto y - \bar{y}$$

Dual view of regularisation

- One can show that $\widehat{\beta}(\lambda)$ is also solution of the following constrained optimization problem

$$\begin{cases} \text{Minimise} & \textbf{RSS}(\beta) \\ \text{Subject to} & \Omega(\beta) \leq T(\lambda) \end{cases} \tag{2}$$

for some threshold $T(\lambda)$ that depends on λ.



Ridge regression

- The coefficients are penalized by

$$\Omega_{\mathrm{Ridge}}(\beta) \equiv \sum_{j=1}^{p} \beta_j^2 .$$

- The ridge estimate $\beta_{\mathrm{Ridge}}(\lambda)$ is

$$\beta_{\mathrm{Ridge}}(\lambda) \; = \; \Big\{ \; \sum_{i=1}^{n} \big(y_i - [\beta_0 + \beta_1\, x_{i,1} + \ldots + \beta_p\, x_{i,p}]\big)^2 \\ + \lambda \times \sum_{j=1}^{p} \beta_j^2 \; : \; \beta \in \mathbb{R}^{p+1} \Big\}$$

- In practice, the covariates are first normalized; one can thus suppose $\beta_0 = 0$. For $X \in \mathbb{R}^{n,p}$ and $\beta \in \mathbb{R}^p$ the estimate $\beta_{\mathrm{Ridge}}(\lambda)$ minimizes the function

$$\beta \mapsto \|y - X\,\beta\|^2 + \lambda\; \|\beta\|_2^2$$

Ridge regression

- The estimate $\beta_{\mathrm{Ridge}}(\lambda)$ minimizes the function

$$\beta \mapsto \|y - X\,\beta\|^2 + \lambda\,\|\beta\|_2^2$$

- **[Exercise]** The gradient of this function reads

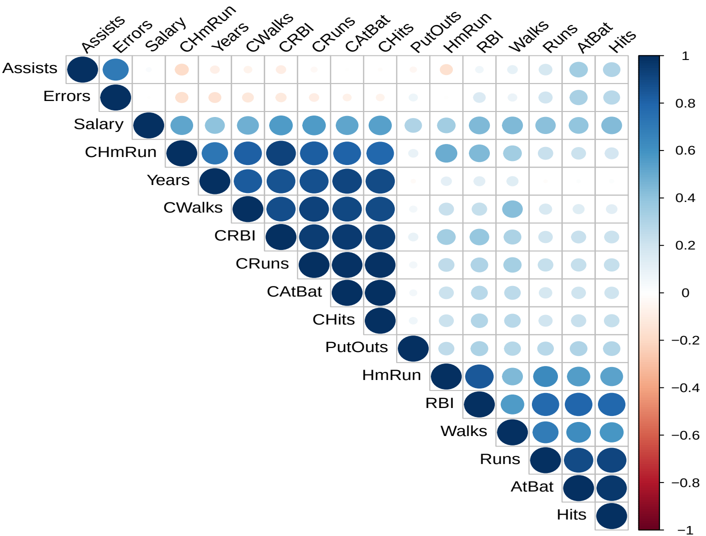
$$-2X^T\,(y - X\,\beta) + 2\lambda\,\beta$$

- **[Exercise]** Setting this gradient to zero yields that

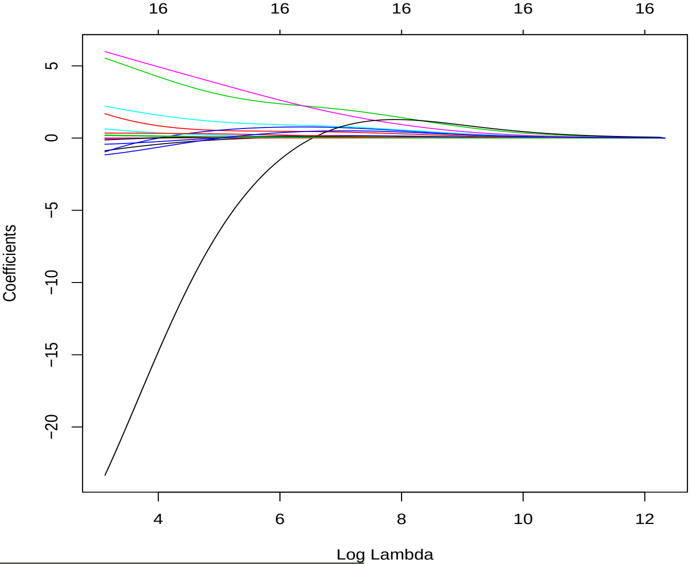
$$\widehat{\beta}(\lambda) = (X^T\,X + \lambda\,I)^{-1}\,X^T\,y.$$

- The matrix $(X^T\,X + \lambda\,I)$ is invertible if $\lambda > 0$.
- Ridge regression is well defined even for $p \geq n + 1$.

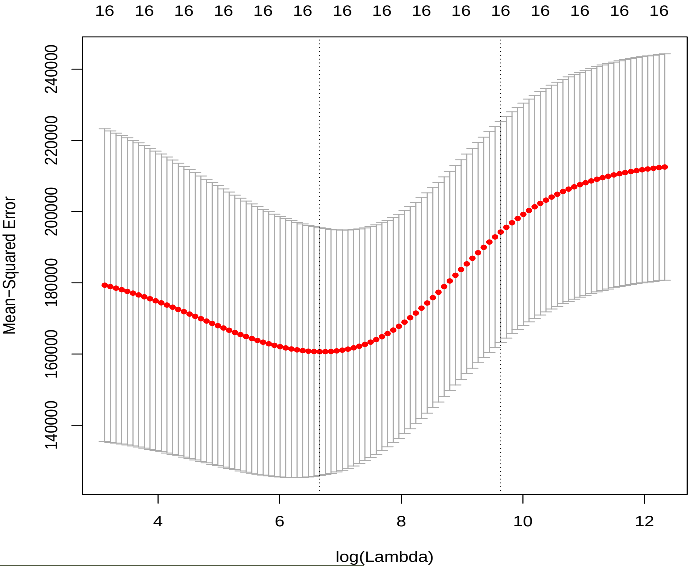
Cricket players dataset



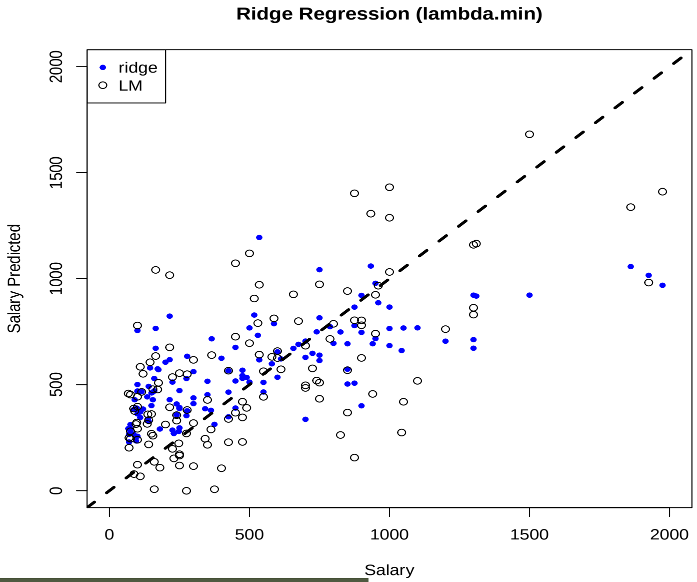
Ridge path



Ridge Cross Validation

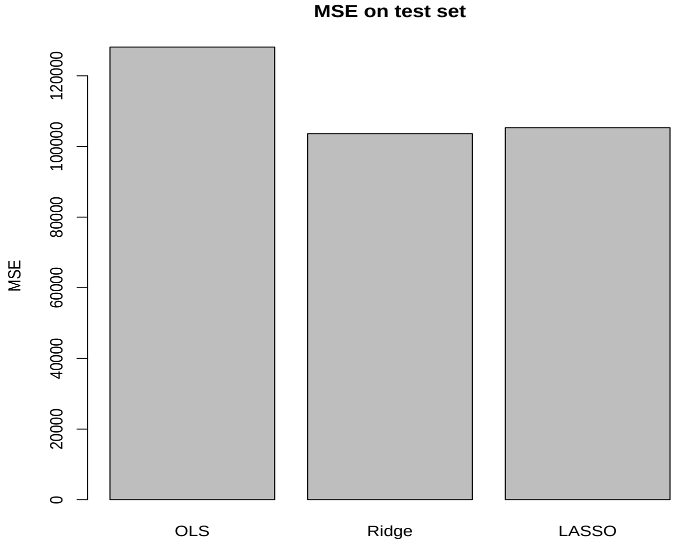


Ridge v.s. OLS



Ridge Regression (lambda.min)
ridge
LM
Salary Predicted
Salary
0
500
1000
1500
2000

Is it worth it?



Least Absolute Shrinkage and Selection Operator (LASSO)

- The coefficients are penalized by

$$\Omega_{\text{Lasso}}(\beta) \equiv \sum_{j=1}^{p} |\beta_j|.$$

- The LASSO estimate $\beta_{\text{Lasso}}(\lambda)$ is

$$\begin{aligned}\beta_{\text{Lasso}}(\lambda) \; = \; \Big\{ \; & \sum_{i=1}^{n} \left(y_i - [\beta_0 + \beta_1\, x_{i,1} + \ldots + \beta_p\, x_{i,p}]\right)^2 \\ & + \lambda \times \sum_{j=1}^{p} |\beta_j| \; : \; \beta \in \mathbb{R}^{p+1} \Big\}\end{aligned}$$

- In practice, the covariates are first normalized; one can thus suppose $\beta_0 = 0$. For $X \in \mathbb{R}^{n,p}$ and $\beta \in \mathbb{R}^p$ the estimate $\beta_{\text{Ridge}}(\lambda)$ minimizes the function

$$\beta \mapsto \|y - X\,\beta\|^2 + \lambda\; \|\beta\|_1$$

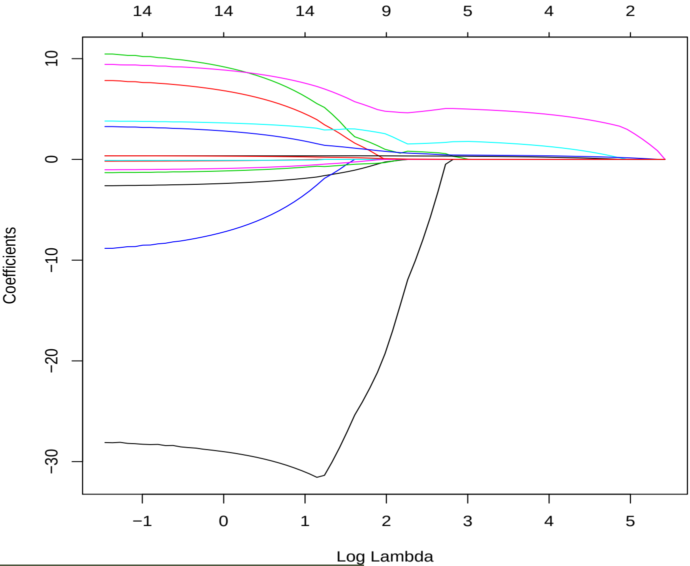
Least Absolute Shrinkage and Selection Operator (LASSO)

- The estimate $\beta_{\mathrm{Lasso}}(\lambda)$ minimizes the function

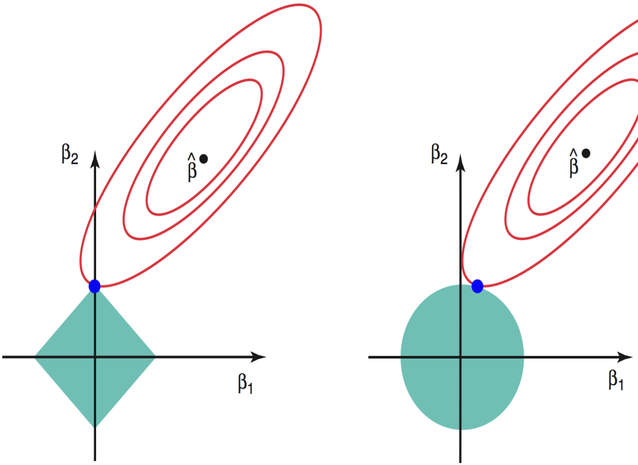
$$\beta \mapsto \|y - X\,\beta\|^2 + \lambda\ \|\beta\|_1$$

- there is no closed form
- the objective function is not differentiable
- **[Exercise]** the objective function is convex
- LASSO regression is well defined even for $p \geq n + 1$.

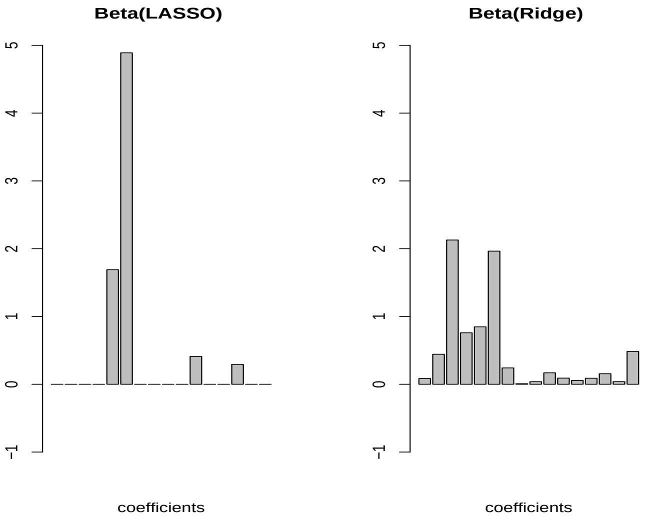
LASSO path



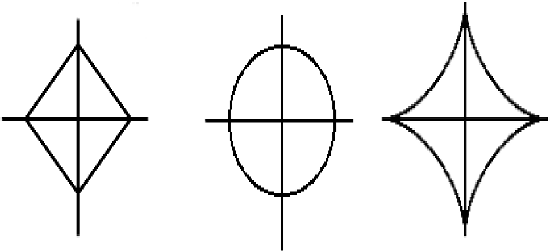
LASSO and sparsity



LASSO and sparsity

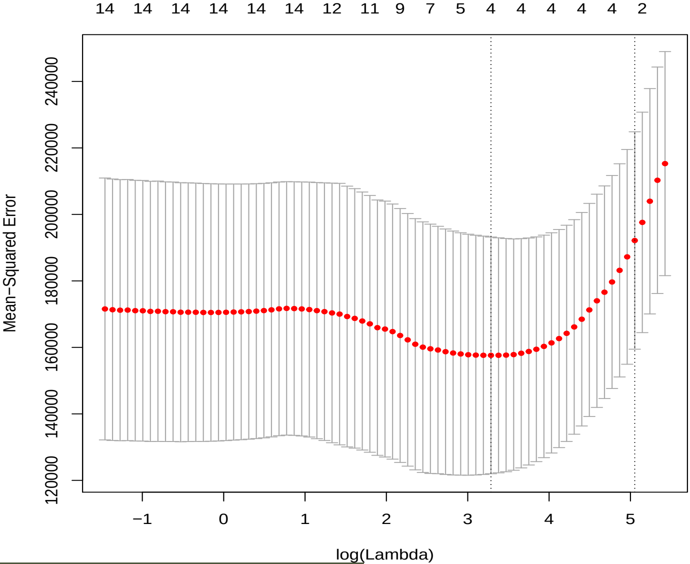


Why not a p-norm with $p < 1$?

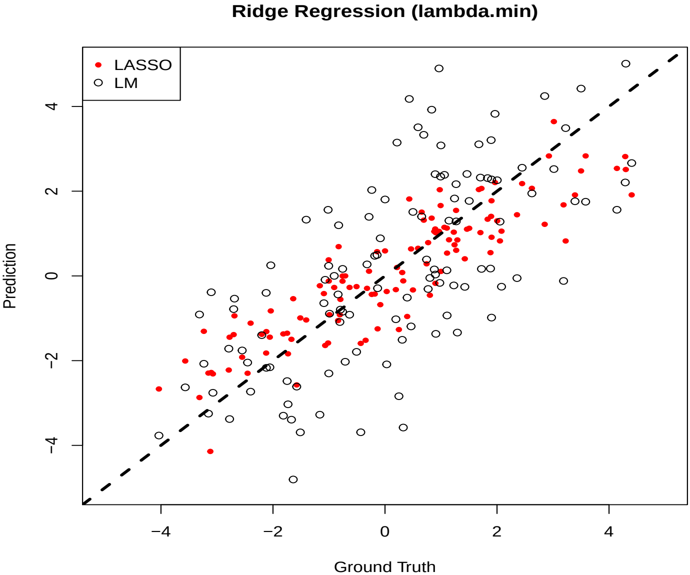


$$\beta \mapsto \|y - X\beta\|^2 + \lambda \, {\color{red}\|\beta\|_p^p}$$

Ridge Cross Validation



Ridge v.s. OLS


Ridge Regression (lambda.min)
LASSO
LM
Prediction
Ground Truth
−4
−2
0
2
4

Is it worth it?