VE472 Lecture 6

Jing Liu

UM-SJTU Joint Institute

Summer

- Although a simple training-test split is done randomly, only one such split is actually used in the above approach. For a small n, the resulting λ_{opt} based on the above approach might vary a lot from one random split to another.
- Leave-one-out cross-validation (LOOCV) can remove this layer of variability.

split: n-1 as trainging set, 1 as test set

Algorithm 1: Determining λ_{opt} using LOOCV

Input : Data matrix $\mathbf{X}$, data vector $\mathbf{y}$, set of parameters $\mathcal{S}_\lambda$

Output : Optimal shrinkage parameter λ_{opt}

Function `LamLOOCV`$(\mathbf{X}, \mathbf{y}, \mathcal{S}_\lambda)$**:**

    for $i \leftarrow 1$ **to** n **do**

        $\mathbf{X}_{-i,} \leftarrow \mathbf{X}[-i,]$; `/* Create` n `Training sets, the` i`th set */`

        $\mathbf{y}_{-i} \leftarrow \mathbf{y}[-i]$; `/* consists all cases except the` i`th case */`

    end for

每一行作为test set的时候的error

$$\lambda_{\text{opt}} \leftarrow \underset{\lambda \in \mathcal{S}_\lambda}{\arg\min} \left\{ \sum_{i=1}^{n} \left(\mathbf{y}[i] - \mathbf{X}[i,] \left(\mathbf{X}_{-i,}^{\mathrm{T}} \mathbf{X}_{-i,} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}_{-i,}^{\mathrm{T}} \mathbf{y}_{-i} \right)^2 \right\} ;$$

without ith row

    return λ_{opt} ;

end

- Notice the last algorithm is only realistic for relatively small n since

$$\left(\mathbf{X}_{-i,}^{\mathrm{T}} \mathbf{X}_{-i,} + \lambda \mathbf{I}\right)^{-1}$$

n大了算起来太慢
就算用LU, etc也太慢了

need to be done n times for every λ, that is, one for every $i = 1, 2, \ldots, n$.

Q: Why does the following give a more efficient way to implement Algorithm 1.

Theorem 0.1

Let $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ *and* $\mathbf{D}$ *be a diagonal matrix of* $n \times n$ *with*

$$\mathbf{D}_{ii} = \frac{1}{1 - \eta_{ii}}, \qquad \textit{where} \qquad \eta_{ii} = \mathbf{X}_{i,}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1}\mathbf{X}_{i,}^{\mathrm{T}}$$

and $\mathbf{A}_{-i} = \mathbf{X}_{-i,}^{\mathrm{T}}\mathbf{X}_{-i,} + \lambda \mathbf{I}$, *then the following holds*

只要算一遍inverse

$$\boxed{\sum_{i=1}^{n}\left(y_i - \mathbf{X}_{i,}\mathbf{A}_{-i}^{-1}\mathbf{X}_{-i,}^{\mathrm{T}}\mathbf{y}_{-i}\right)^2} = \left\|\mathbf{D}\left(\mathbf{I} - \mathbf{H}\right)\mathbf{y}\right\|^2$$

前一页的sum of squares

where $\|\cdot\|$ *denotes the usual Euclidean norm.*

Algorithm 2: Determining λ_{opt} using LOOCV with Woodbury identity

Input : Data matrix $\mathbf{X}$, data vector $\mathbf{y}$, set of parameters $\mathcal{S}_\lambda$

Output : Optimal shrinkage parameter λ_{opt}

Function `LamLOOCVWoodbury(`$\mathbf{X}, \mathbf{y}, \mathcal{S}_\lambda$`)`**:**

`/* Create functions of` λ `needed for the optimisation */`

$\mathbf{A}^{-1}(\lambda) \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}$; `/* matrix-valued function of` λ `*/`

$\mathbf{H}(\lambda) \leftarrow \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}$; `/* matrix-valued function of` λ `*/`

for $i \leftarrow 1$ **to** n **do**

$\eta_{ii}(\lambda) \leftarrow \mathbf{X}[i,]\mathbf{A}^{-1}\mathbf{X}[i,]^{\mathrm{T}}$; `/* real-valued functions of` λ `*/`

end for

$$\mathbf{D}(\lambda) \leftarrow \operatorname{diag}\left(\frac{1}{1-\eta_{11}(\lambda)}, \frac{1}{1-\eta_{22}(\lambda)}, \cdots, \frac{1}{1-\eta_{nn}(\lambda)}\right);$$

`/* matrix-valued function of` λ `*/`

$$\lambda_{\text{opt}} \leftarrow \underset{\lambda \in \mathcal{S}_\lambda}{\arg\min} \left\{ \left\| \mathbf{D}(\lambda)\left(\mathbf{I} - \mathbf{H}(\lambda)\right)\mathbf{y} \right\|^2 \right\};$$

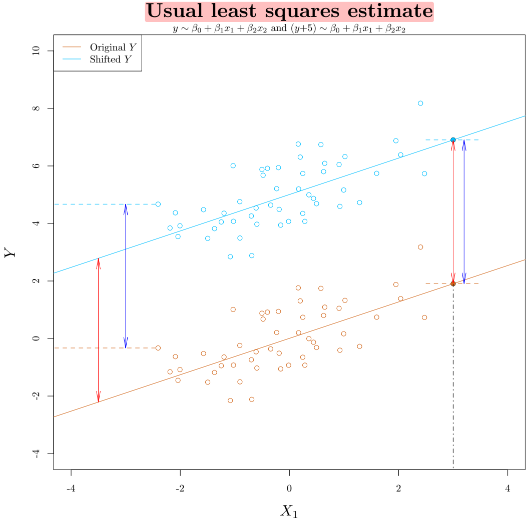
return λ_{opt} ;

end

- Notice the following desirable property that the least squares estimator has.

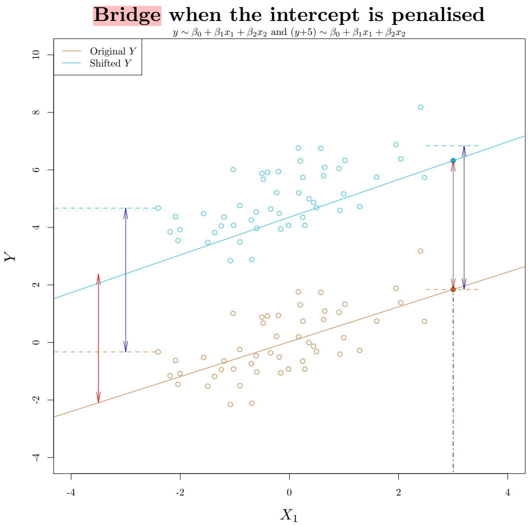
所有的y加了5

slope is the same, intercept changed



- However, it isn't true in our current formulation of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$, i.e. penalising $\hat{\beta}_0$

shift y ->
slope intercept
都变了



- The above simulation illustrates a desirable property of $\hat{\boldsymbol{\beta}}_{\text{lse}}$ that our current formulation of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ does not have, namely, altering the centre of $\mathbf{y}$,

目标：slope不change

$$\bar{y} = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{y}, \qquad \text{where } \mathbf{1} \text{ denotes the vector of ones,}$$

y bar: mean of all components

only alters the intercept component of $\hat{\boldsymbol{\beta}}_{\text{lse}}$, while it alters $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ completely.

- To understand why, let $\mathbf{y}_{\mathrm{c}} = \mathbf{y} - \bar{y}\mathbf{1}$, then for any centre $\bar{y}$ we can rewrite

yc centered at 0

$$\begin{aligned}\mathbf{b}_{\text{lse}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{y}_{\mathrm{c}} + \bar{y}\mathbf{1}\right)\\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}_{\mathrm{c}} + \bar{y}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1}\end{aligned}$$

first colum of XTX (X第一列都是1，因为有intercept)

- Note $\mathbf{X}^{\mathrm{T}}\mathbf{1}$ is the 1st column of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, so $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1} = \begin{bmatrix}1 & \mathbf{0}_{1\times k}\end{bmatrix}^{\mathrm{T}}$ and

$$\underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2\right\} = \boxed{\underset{\mathbf{b}_{\mathrm{c}}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y}_{\mathrm{c}} - \mathbf{X}\mathbf{b}_{\mathrm{c}}\|^2\right\}} + \boxed{\begin{bmatrix}\bar{y} & \mathbf{0}_{1\times k}\end{bmatrix}^{\mathrm{T}}}$$

intercept = 0的部分

changing according to changing of intercept

- However, in terms of our current formulation of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$, we have

$$\begin{aligned}\mathbf{b}_{\text{ridge}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{y}_{\mathrm{c}} + \bar{y}\mathbf{1}\right) \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}_{\mathrm{c}} + \bar{y}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1}\end{aligned}$$

- Note the vector $\mathbf{X}^{\mathrm{T}}\mathbf{1}$ is not the 1st column of $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}$, thus in general,

$$\underline{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1} \neq \begin{bmatrix}1 & \mathbf{0}_{1\times k}\end{bmatrix}^{\mathrm{T}}}$$

hence the estimate will change more than just the first component,

$$\begin{aligned}\mathbf{b}_{\text{ridge}} &= \underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2\right\} \\ &\neq \underset{\mathbf{b}_{\mathrm{c}}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y}_{\mathrm{c}} - \mathbf{X}\mathbf{b}_{\mathrm{c}}\|^2 + \lambda\|\mathbf{b}_{\mathrm{c}}\|^2\right\} + \begin{bmatrix}\bar{y} & \mathbf{0}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\end{aligned}$$

- A small change in our formulation of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ will allow us to rectify it.

- Suppose $\mathbf{X} = \begin{bmatrix}\mathbf{1} & \mathbf{X}_{-0}\end{bmatrix}$ and $\mathbf{b}^{\mathrm{T}} = \begin{bmatrix}b_0 & \mathbf{b}_{-0}^{\mathrm{T}}\end{bmatrix}$, then we have the following

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 &= \|\mathbf{y}_{\mathrm{c}} + \bar{y}\mathbf{1} - b_0\mathbf{1} - \mathbf{X}_{-0}\mathbf{b}_{-0}\|^2 \\ &= \|\mathbf{y}_{\mathrm{c}} - (b_0 - \bar{y})\,\mathbf{1} - \mathbf{X}_{-0}\mathbf{b}_{-0}\|^2 = \|\mathbf{y}_{\mathrm{c}} - \mathbf{X}\mathbf{b}_{\mathrm{c}}\|^2\end{aligned}$$

  from which we can conclude $\mathbf{b}_{\mathrm{c}}^{\mathrm{T}} = \begin{bmatrix}b_0 - \bar{y} & \mathbf{b}_{-0}^{\mathrm{T}}\end{bmatrix}$ for any centre $\bar{y}$.
- Note $\|\mathbf{b}_{\mathrm{c}}\|$ varies as $\bar{y}$ varies due to the 1st component, so does the penalty

$$\lambda\|\mathbf{b}_{\mathrm{c}}\|^2$$

  for the same $\mathbf{b}$ and λ, which suggests an approach of rectifying the difference

$$\begin{aligned}\mathbf{b}_{\mathrm{ridge}} &= \underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2\right\} \\ &\neq \underset{\mathbf{b}_{\mathrm{c}}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y}_{\mathrm{c}} - \mathbf{X}\mathbf{b}_{\mathrm{c}}\|^2 + \lambda\|\mathbf{b}_{\mathrm{c}}\|^2\right\} + \begin{bmatrix}\bar{y} & \mathbf{0}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\end{aligned}$$

Q: Can you guess what this approach is?

- Consider the following alternative formulation $\hat{\boldsymbol{\beta}}_{\text{ridge}}$

$$\mathbf{b}_{\text{ridge}} = \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - b_0 \mathbf{1} - \mathbf{X}_{-0}\mathbf{b}_{-0}\|^2 + \lambda \|\mathbf{b}_{-0}\|^2 \right\}$$

which differs from our original formulation by exempting b_0 from penalty.

- Differentiating the above objective function with respect to b_0, we have

$$-2\mathbf{1}^{\mathrm{T}}\mathbf{y} + 2nb_0 + 2\mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}\mathbf{b}_{-0}$$

setting which to zero, we have

center of y

intercept

$$b_0 = \frac{1}{n}\left(\mathbf{1}^{\mathrm{T}}\mathbf{y} - \mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}\mathbf{b}_{-0}\right) = \bar{y} - \underbrace{\frac{1}{n}\left(\mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}\right)}_{?}\mathbf{b}_{-0}$$

row vector of sample mean of each variable

which means if we apply a mean-centring shift to $\mathbf{X}_{-0}$, then the value of b_0 that attains the minimum is given by $b_0 = \bar{y}$ independent of $\mathbf{b}_{-0}$ and λ.

- Let us denote the mean-centring shift in $\mathbf{X}_{-0}$ by

$$z_{ij} = x_{ij} - \bar{x}_j \qquad \text{for} \quad j = 1, 2, \cdots, k \quad \text{and} \quad i = 1, 2, \cdots, n$$

where $\bar{x}_j$ is the mean of the jth column of $\mathbf{X}_{-0}$, in matrix notation, we have

$$\mathbf{Z}_{-0} = \mathbf{X}_{-0} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}} \qquad \text{where} \quad \bar{\mathbf{x}}^{\mathrm{T}} = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}$$

so
$$\begin{aligned}
\mathbf{b}_{\text{ridge}} &= \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - b_0\mathbf{1} - \mathbf{X}_{-0}\mathbf{b}_{-0}\|^2 + \lambda\|\mathbf{b}_{-0}\|^2 \right\} \\
&= \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - b_0\mathbf{1} - \left(\mathbf{Z}_{-0} + \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)\mathbf{b}_{-0}\|^2 + \lambda\|\mathbf{b}_{-0}\|^2 \right\} \\
&= \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - \left(b_0 + \bar{\mathbf{x}}^{\mathrm{T}}\mathbf{b}_{-0}\right)\mathbf{1} - \mathbf{Z}_{-0}\mathbf{b}_{-0}\|^2 + \lambda\|\mathbf{b}_{-0}\|^2 \right\} \\
&= \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y}_{\mathrm{c}} - \mathbf{Z}_{-0}\mathbf{b}_{-0}\|^2 + \lambda\|\mathbf{b}_{-0}\|^2 \right\}
\end{aligned}$$

where the minimiser $b_0 = \bar{y} - \frac{1}{n}\left(\mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}\right)\mathbf{b}_{-0}$, and $\mathbf{y} = \mathbf{y}_{\mathrm{c}} - \bar{y}\mathbf{1}$ are used.

- Therefore, we can conclude $\mathbf{b}_{-0}$ is invariant under mean-centring

$$
\begin{aligned}
\mathbf{b}_{\text{ridge}} &= \underset{\mathbf{b}_{-0}\in\mathbb{R}^k}{\arg\min}\left\{\|\mathbf{y}-b_0\mathbf{1}-\mathbf{X}_{-0}\mathbf{b}_{-0}\|^2+\lambda\|\mathbf{b}_{-0}\|^2\right\} \\
&= \underset{\mathbf{v}_{-0}\in\mathbb{R}^k}{\arg\min}\left\{\|\mathbf{y}-v_0\mathbf{1}-\mathbf{Z}_{-0}\mathbf{v}_{-0}\|^2+\lambda\|\mathbf{v}_{-0}\|^2\right\}
\end{aligned}
$$

where $v_0=\bar{y}$ and $\mathbf{b}_{-0}=\mathbf{v}_{-0}$, and prediction is also invariant

$$
\begin{aligned}
\hat{y}_i = b_0+\mathbf{x}_{i,}^{\mathrm{T}}\mathbf{b}_{-0} &= \left(v_0-\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{b}_{-0}\right)+\mathbf{x}_{i,}^{\mathrm{T}}\mathbf{b}_{-0} \\
&= v_0+\left(\mathbf{x}_{i,}^{\mathrm{T}}-\bar{\mathbf{x}}^{\mathrm{T}}\right)\mathbf{b}_{-0} \\
&= v_0+\mathbf{z}_{i,}^{\mathrm{T}}\mathbf{v}_{-0}
\end{aligned}
$$

since the last expression is the prediction for Y in terms of the centred data

$$\mathbf{z}_{i,}=\mathbf{x}_{i,}-\bar{\mathbf{x}}$$

- Note this invariant property is actually true for any shift $\mathbf{Z}_{-0}=\mathbf{X}_{-0}-\mathbf{1}\boldsymbol{\alpha}^{\mathrm{T}}$.

- Recall what we have defined

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_{-0} \end{bmatrix}; \qquad \mathbf{b}^{\mathrm{T}} = \begin{bmatrix} b_0 & \mathbf{b}_{-0}^{\mathrm{T}} \end{bmatrix}; \qquad \mathbf{y} = \mathbf{y}_{\mathrm{c}} + \bar{y}\mathbf{1}$$
$$\mathbf{Z}_{-0} = \mathbf{X}_{-0} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}; \qquad \mathbf{v}^{\mathrm{T}} = \begin{bmatrix} v_0 & \mathbf{v}_{-0}^{\mathrm{T}} \end{bmatrix}; \qquad \mathbf{v}_{-0} = \mathbf{b}_{-0}$$

Q: Have we rectified our first formulation of $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ with our second formulation

$$\begin{aligned}
\mathbf{b}_{\mathrm{ridge}} &= \underset{b_0 \in \mathbb{R}, \mathbf{b}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - b_0\mathbf{1} - \mathbf{X}_{-0}\mathbf{b}_{-0}\|^2 + \lambda\|\mathbf{b}_{-0}\|^2 \right\} \\
&= \underset{v_0 \in \mathbb{R}, \mathbf{v}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y} - v_0\mathbf{1} - \mathbf{Z}_{-0}\mathbf{v}_{-0}\|^2 + \lambda\|\mathbf{v}_{-0}\|^2 \right\} \\
&= \underset{v_0 = \bar{y}, \mathbf{v}_{-0} \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{y}_{\mathrm{c}} - \mathbf{Z}_{-0}\mathbf{v}_{-0}\|^2 + \lambda\|\mathbf{v}_{-0}\|^2 \right\} \\
&= \begin{bmatrix} \bar{y} & \left(\left(\mathbf{Z}_{-0}^{\mathrm{T}}\mathbf{Z}_{-0} + \lambda\mathbf{I} \right)^{-1} \mathbf{y}_{\mathrm{c}} \right)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}
\end{aligned}$$

Q: Is the above estimator invariant under scaling of $\mathbf{y}$ or columns of $\mathbf{Z}_{-0}$?

no

- Let $\mathbf{M}$ be an invertible diagonal matrix of $k \times k$, and $\mathbf{U} = \mathbf{Z}_{-0}\mathbf{M}^{-1}$, then

$$\underline{\left(\mathbf{Z}_{-0}^{\mathrm{T}}\mathbf{Z}_{-0} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}_{\mathrm{c}}} = \left((\mathbf{U}\mathbf{M})^{\mathrm{T}}\mathbf{U}\mathbf{M} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}_{\mathrm{c}}$$
$$= \left(\mathbf{M}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{M} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}_{\mathrm{c}} \neq \underline{\left(\mathbf{U}^{\mathrm{T}}\mathbf{U} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}_{\mathrm{c}}}$$

since $\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{M} \neq \mathbf{M}\mathbf{U}^{\mathrm{T}}\mathbf{U}$ and $\mathbf{M}$ is not orthogonal in general, which means our second formulation is not invariant under column scaling of $\mathbf{Z}_{-0}$.

Q: Should we scale the the data matrix $\mathbf{Z}_{-0}$? If so, what should we use for $\mathbf{M}$?

- Unless the independent variables have the same units, to be "fair" we scale the columns of the data matrix so that the sample variances become 1,

$$\mathbf{U} = \mathbf{Z}_{-0}\,\mathrm{diag}\left(s_1, s_2, \cdots, s_k\right)^{-1}, \quad \text{where} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_{ij} - \bar{x}_j\right)^2$$

which is the usual formulation of $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ that is widely implemented.

Algorithm 3: Ridge Estimation

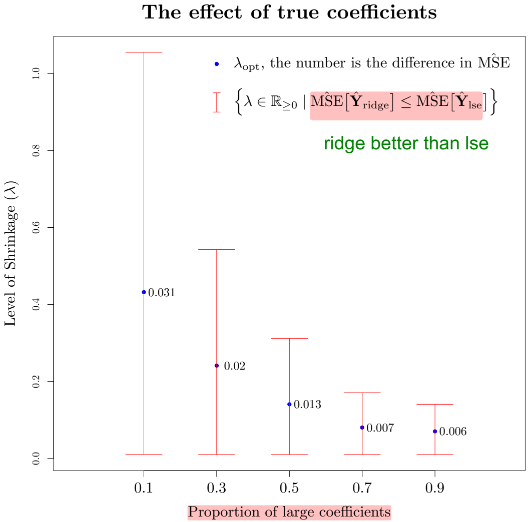
Input : Data matrix $\mathbf{X}_{n\times(k+1)}$, data vector $\mathbf{y}_{n\times 1}$, set of parameters $\mathcal{S}_\lambda$
Output : Vector of estimated parameters $\mathbf{b}_{\text{ridge}}$

Function `RIDGE(`$\mathbf{X}, \mathbf{y}, \mathcal{S}_\lambda$`)`:
  $\mathbf{X}_{-0} \leftarrow \mathbf{X}[, -1]$; `/* remove the column of ones */`
  $\mathbf{Z}_{-0} \leftarrow \mathbf{X}_{-0} - \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{X}_{-0}$; `/* mean-centre the columns of` $\mathbf{X}_{-0}$ `*/`
  for $j \leftarrow 1$ **to** k **do**

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\Big(\mathbf{Z}_{-0}[i,j]\Big)^2} ;$$ `/* column standard deviation */`

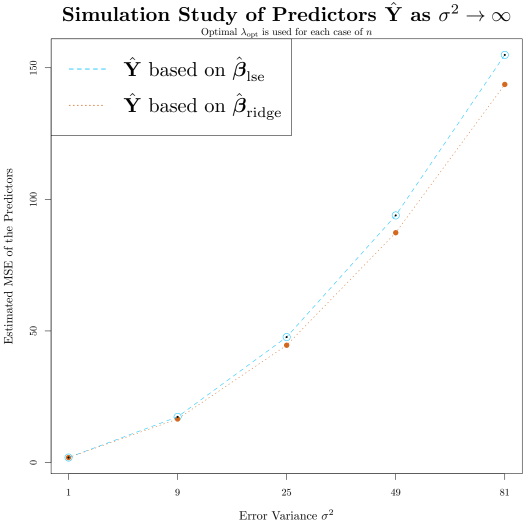
  end for
  $\mathbf{U} \leftarrow \mathbf{Z}_{-0}\,\mathrm{diag}\,(s_1, s_2, \ldots, s_k)^{-1}$; `/* scale the columns of` $\mathbf{Z}_{-0}$ `*/`
  $\mathbf{b}_{\text{ridge}}[1] \leftarrow \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{y}$; `/* set the mean of` $\mathbf{y}$ `as` b_0 `*/`
  $\mathbf{y}_{\text{c}} \leftarrow \mathbf{y} - \mathbf{b}_{\text{ridge}}[1]\cdot\mathbf{1}$; `/* mean-centre` $\mathbf{y}$ `*/`
  $\lambda_{\text{opt}} \leftarrow$ `LamLOOCVWoodbury(`$\mathbf{U}, \mathbf{y}_{\text{c}}, \mathcal{S}_\lambda$`)` ; `/* find the optimal` λ `*/`
  $\mathbf{b}_{\text{ridge}}[-1] \leftarrow \left(\mathbf{U}^{\mathrm{T}}\mathbf{U} + \lambda_{\text{opt}}\mathbf{I}\right)^{-1}\mathbf{y}_{\text{c}}$; `/* Slope estimates */`
  return $\mathbf{b}_{\text{ridge}}$;
end

- $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ only slightly outperform $\hat{\boldsymbol{\beta}}_{\text{lse}}$ for a narrow range of λ if $\boldsymbol{\beta}$ is mostly big.



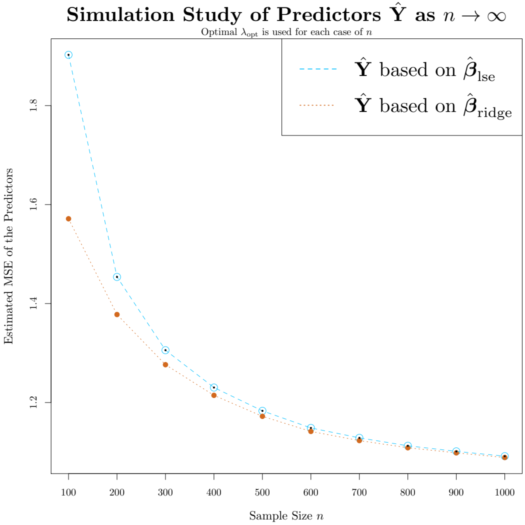
small proportion of large coefficients -> a large number of lamba that we can choose to outperform b_lse

- As σ^2 grows, the tradeoff between variance and bias becomes more notable.



noise多，用ridge好

- However, as n increases, the performance of $\hat{\boldsymbol{\beta}}_{\text{lse}}$ catches up with $\hat{\boldsymbol{\beta}}_{\text{ridge}}$.



n很大，both converge to real beta

- Our motivation for $\hat{\boldsymbol{\beta}}_{\text{ridge}}$, i.e. bypass variable selection and achieve a smaller

$$\text{MSE}\left(\hat{Y}_i\right) = \text{MSE}\left(\mathbf{x}_{i,}^{\text{T}}\hat{\boldsymbol{\beta}}\right)$$

ridge其实适用的是medium size data
n或者k太大不行

becomes weak if n as well as k is relatively large.

- Although it is not exclusive to big/complex datasets, rank deficiency issue,

$$\det\left(\mathbf{X}^{\text{T}}\mathbf{X}\right) \approx 0$$

is often more prominent in more complex datasets.

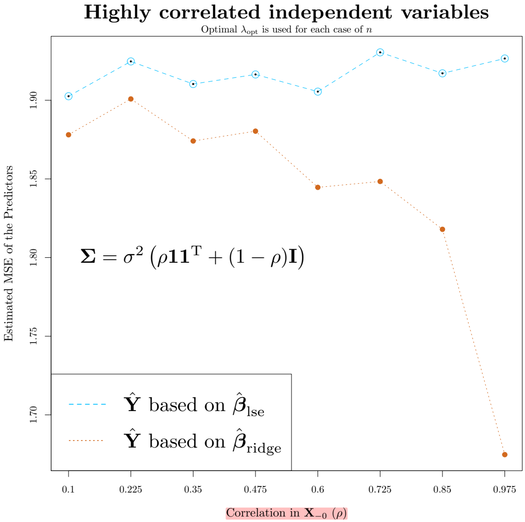
- We will show later on the motivation for using the ridge estimate for large n

$$\mathbf{b}_{\text{ridge}} = \begin{bmatrix} \bar{y} \\ \left(\mathbf{U}^{\text{T}}\mathbf{U} + \lambda\mathbf{I}\right)^{-1}\mathbf{U}^{\text{T}}\mathbf{y}_c \end{bmatrix}$$

is the fact that the matrix $\mathbf{U}^{\text{T}}\mathbf{U} + \lambda\mathbf{I}$ is always invertible for $0 < \lambda < \infty$.

- Highly Correlated

有rank deficiency:
用ridge更好



Recap: Regression

- Linear least squares regression has zero bias but suffers from high variance.

$$\mathrm{MSE}(\hat{Y}_i) = \mathbb{E}\left[\left(\hat{Y}_i - Y_i\right)^2\right] = \mathrm{Var}\left[\hat{Y}_i\right] + \left(\underbrace{\mathbb{E}\left[\hat{Y}_i\right] - \mathbb{E}\left[Y_i\right]}_{\text{Bias}}\right)^2 + \sigma^2$$

- Ridge regression reduces mean square error without doing variable selection.
- Various things, the size of k and n, the relative size of the true coefficients, and the relationship between independent variables, decides when we prefer ridge over linear least square regression when building a predictive model.
- It is used for building predictive models, where k is relatively large while n is relatively small. It is particularly good if the independent variables are highly correlated, and there is a subset of true coefficients that are relatively small.

Q: What should we do in terms of regression if n is very large? Very large k?