VE472 Lecture 5

Jing Liu

UM-SJTU Joint Institute

Summer

- It can be shown that the least squares estimator is unbiased and consistent, furthermore, it is the best in the sense described by the following theorem.

Theorem 0.1 (Gauss-Markov Theorem)

Given $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ *and* $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$, *then the least squares estimator, if it exists,*

$$\hat{\boldsymbol{\beta}}_{\mathrm{lse}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

is the best linear unbiased estimator of $\boldsymbol{\beta}$ *in the sense of minimum variance for*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

also unbiased

that is, for any estimator of the form $\tilde{\boldsymbol{\beta}} = \mathbf{A}_{(k+1)\times n}\mathbf{y}$ *such that* $\mathbb{E}\big[\tilde{\boldsymbol{\beta}}\big] = \boldsymbol{\beta}$, *then*

$$\mathrm{Var}\big[\boldsymbol{\alpha}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}\big] \geq \mathrm{Var}\big[\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{lse}}\big] \qquad \text{where} \quad \boldsymbol{\alpha} \in \mathbb{R}^{k+1}$$

smallest variance

- This is one of the main reasons why $\hat{\boldsymbol{\beta}}_{\mathrm{lse}}$ is widely used for small data.

Q: What happens if $\mathbf{X}$ is not full rank?

Q: Will the last theorem still hold when $\mathbf{X}$ is not full rank? no

Q: Why do we expect the least squares estimate is no longer unique?

$$\arg\min_{\mathbf{b}\in\mathbb{R}^{k+1}} \|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2$$

一种方法: min ||b|| (L4)
minimum norm solution:
eliminate unecessary virables

- In this situation, we say $\boldsymbol{\beta}$ is non-identifiable, and additional constraint(s) need to be introduced to reach a unique estimate of $\boldsymbol{\beta}$.
- The matrix $\mathbf{X}$ being less than full rank is not specific to big data but it does become more prominent when $\mathbf{X}$ becomes more complex in modern era.
- The simplest type of constraint is to set some of the coefficients β_j to zero, that is, we select or exclude some variables from the k independent variables.
- Variable selection is traditionally done according to some criterion, e.g.

AIC/BIC

- When building a predictive model, variable selection is done according to

$$\mathrm{MSE}(\hat{Y}_i) = \mathbb{E}\left[\left(\hat{Y}_i - Y_i\right)^2\right]$$

which is different from the MSE of an estimator $\hat{\theta}$ of $\theta \in \mathbb{R}$

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

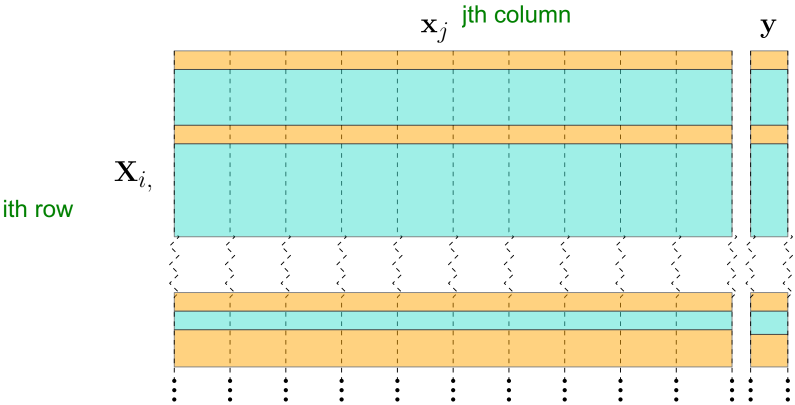
since $\mathrm{MSE}(\hat{Y}_i)$, a.k.a. prediction error, involves two random variables.

- In the context of regression analysis, it can be shown that

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

- In practice, we have to estimate $\mathrm{MSE}(\hat{Y}_i)$, e.g. via simple training-test split.

- Simple training-test split is widely used when n is sufficiently large.



Randomly selected training set, denoted by $\mathbf{X}^\dagger$ and $\mathbf{y}^\dagger$

Randomly selected test set, denoted by $\mathbf{X}^*$ and $\mathbf{y}^*$

- Let $n-m$ and m denote the number of cases in the training and test set.

- Given a simple training-test split, we "train" various linear models

$$\mathbf{y}^\dagger = \mathbf{X}_\ell^\dagger \mathbf{b}_\ell + \boldsymbol{\varepsilon}_\ell$$

where only some of the k independent variables, say $p \leq k$, are included, e.g.

$$y_i = \beta_0 + \beta_1 x_{i1} + 0 x_{i2} + \cdots + \beta_{k-1} x_{i,k-1} + 0 x_{ik} + e_i$$

- The MSE of the ℓth model is estimated by computing the followings

$$\mathbf{b}_\ell = \underset{\mathbf{b} \in \mathbb{R}^{p+1}}{\arg\min} \left\| \mathbf{y}^\dagger - \mathbf{X}_\ell^\dagger \mathbf{b} \right\|^2$$

$$\boxed{\hat{\mathrm{MSE}} = \frac{1}{m} \left\| \mathbf{y}^* - \mathbf{X}_\ell^* \mathbf{b}_\ell \right\|^2}$$

.calculate using test set

beta_k = 0 / 1

- Note there are 2^k models in total, and each $\mathbf{b}_\ell$ defines just one of them, the model with the smallest $\hat{\mathrm{MSE}}$ is considered to be the best predictive model.

problem: might have too many models

Q: Is there any way to reduce MSE without conducting variable selection?

- Note the following two aspects of this approach when comes to big datasets:

1. When the dataset is complex, i.e. a big k value, the number of models

$$2^k$$

grows really quickly, and significantly outstrip computing power we have.

2. When the dataset is large, i.e., as n grows, $\hat{\boldsymbol{\beta}}_{\mathrm{lse}}$ might not be the best

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

tolerate some bias, reduce variance

- The Gauss-Markov theorem only guarantees this approach would lead us to the minimum $\hat{\mathrm{MSE}}$ estimators of $\boldsymbol{\beta}$ amongst all linear unbiased estimators.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x}_{i,}^{\mathrm{T}} \hat{\boldsymbol{\beta}}$$

where $\mathbf{x}_{i,}^{\mathrm{T}}$ denotes the row vector $\begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik} \end{bmatrix}$ that we predict Y_i with

- However, unbiasedness is not relevant if the dataset is large enough, being consistent will guarantee the quality of the prediction for sufficiently large n.
- In other words, for a large dataset, we can tolerate some bias as long as it is consistent, so we should find a predictive model that minimises the estimated

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

without restricting ourselves to unbiased estimators to form a predictor.

- Just like you have seen earlier, depending on the size of dataset, n, there might be a significant improvement in $\hat{\mathrm{MSE}}$ if we give up unbiasedness.
- Shrinkage methods, which assign the importance of each X_j in predicting y

$$\mathbb{E}\left[Y_i\right] = \mathbf{x}_{i,}^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

as well as estimating the coefficient β_j, offer ways to avoid variable selection as well as estimators which lead to smaller $\hat{\mathrm{MSE}}$s by having some small bias.

Q: Where can we get some bias?

- Recall the following property of variance for fixed scalars w_1, α_1, w_2 and α_2

$$\begin{aligned}\operatorname{Var}\left[w_1 Z_1 \alpha_1 + w_2 Z_2 \alpha_2\right] = w_1^2 \operatorname{Var}\left[Z_1\right] \alpha_1^2 + w_2^2 \operatorname{Var}\left[Z_2\right] \alpha_2^2 \\ + 2 w_1 w_2 \operatorname{Cov}\left[Z_1, Z_2\right] \alpha_1 \alpha_2\end{aligned}$$

Z1 Z2如果not corelated, Cov = 0

where Z_1 and Z_2 are two arbitrary random variables with finite variance.

- Notice the variance shrinks to 0 as the "size" of w_1 and w_2 shrinks.

- Given the linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the variance of following predictor,

w: weight vector

$$\hat{Y}_i = w_0 \hat{\gamma}_0 1 + w_1 \hat{\gamma}_1 x_{i1} + w_2 \hat{\gamma}_2 x_{i2} + \cdots + w_k \hat{\gamma}_k x_{ik}$$

which will be used as a prediction for Y_i given x_{ij} values, shrinks to zero in a similar fashion to the case above if the ℓ_2 norm of $\mathbf{w}$ shrinks to zero.

- Conceptually, $\mathbf{w}$ can be thought as a vector of some kind of weighting factor for each X_j in predicting Y_i, and $\hat{\boldsymbol{\gamma}}$ is some estimator of $\boldsymbol{\beta}$ with respect to $\mathbf{w}$.

- In practice, we do not separate w_j from $\hat{\gamma}_j$, for $j = 0, \ldots, k$,

$$\hat{Y}_i = w_0\hat{\gamma}_0 1 + w_1\hat{\gamma}_1 x_{i1} + w_2\hat{\gamma}_2 x_{i2} + \cdots + w_k\hat{\gamma}_k x_{ik}$$

- Instead of explicitly introducing $\mathbf{w}$, we form an estimator $\hat{\boldsymbol{\beta}}$ that takes $\|\hat{\boldsymbol{\beta}}\|$ into account for shrinking it shrinks the variance, and use the linear predictor

$$\hat{Y}_i = \mathbf{x}_{i,}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

- One natural approach is still to minimise $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ with respect to $\mathbf{b}$, but with the constraint $\|\mathbf{b}\|^2 \leq c_\lambda$, where c_λ defines the level of shrinkage.
- This approach is known as ridge regression,

$$\boxed{\begin{aligned} &\underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min}\, \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\ &\text{s. t.} \qquad \|\mathbf{b}\|^2 \leq c_\lambda \end{aligned}}$$

and its solution is known as the ridge estimate, often denoted by $\mathbf{b}_{\text{ridge}}$.

Theorem 0.2

The constrained minimisation problem in ridge regression for some $c_\lambda > 0$

$$\begin{aligned} &\arg\min_{\mathbf{b}\in\mathbb{R}^{k+1}} \|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2 \\ &\text{s. t.} \qquad \|\mathbf{b}\|^2 \le c_\lambda \end{aligned}$$

can be converted into the following penalised/regularised least squares problem

$$\arg\min_{\mathbf{b}\in\mathbb{R}^{k+1}} \left\{ \|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2 \right\} \qquad \textit{where} \quad 0 \le \lambda < \infty$$

||b|| 越小，variance越小，同时不能让bias太大

where $0 \le \lambda < \infty$ *depends on* c_λ*, and the solution of this problem is given by*

$$\boxed{\mathbf{b}_{\text{ridge}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

X^TX+lamdal is always invertible

- Recall our motivation of using $\mathbf{b}_{\text{ridge}}$ is to avoid variable selection, and attain a smaller $\mathrm{MSE}(\hat{Y})$ by giving up unbiasedness and relying on consistency.

Theorem 0.3

Given $\mathbb{E}\left[\boldsymbol{\varepsilon}\right] = 0$ *and* $\mathrm{Var}\left[\boldsymbol{\varepsilon}\right] = \sigma^2 \mathbf{I}$, *then the ridge estimator,*

$$\boxed{\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

for the unknown parameter $\boldsymbol{\beta}$ *in the linear model,*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is biased and the bias is given by

$$-\lambda \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\beta}$$

but $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ *is consistent in the following sense,*

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} - \boldsymbol{\beta}\right\| \geq \delta\right] = 0 \qquad \textit{for all} \quad \delta > 0$$

where n *denotes the number of cases in the dataset.*

- Note all it does in comparison to $\mathbf{b}_{\text{lse}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is that it creates a "ridge" down the diagonal of the matrix that needs to be "inverted"

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

- Of course, this distorts the original least squares estimate depending on λ

$$\underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min}\left\{\|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2\right\} \qquad \text{where} \quad 0 < \lambda < \infty$$

- It is clear that we have the following limiting cases of the ridge estimate

lse is special case of ridge

$\mathbf{b}_{\text{ridge}} \to \mathbf{b}_{\text{lse}}$	as	$\lambda \to 0$
$\mathbf{b}_{\text{ridge}} \to \mathbf{0}$	as	$\lambda \to \infty$

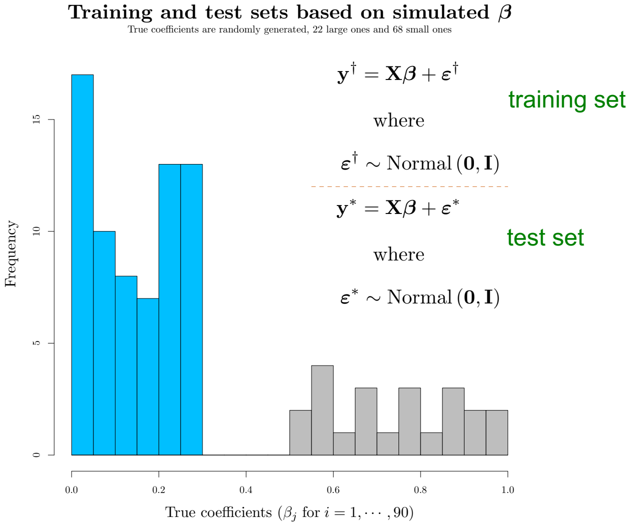
lambda太大bias太大

- The two limiting cases represent the zero bias and the zero variance solution.

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

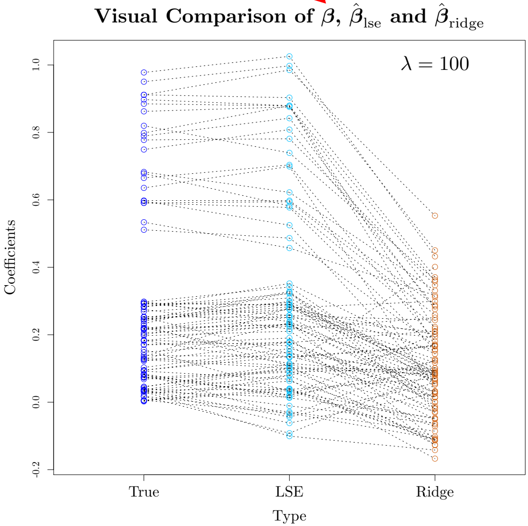
- It is reasonable to expect there is some optimal λ that achieves the balance.

- Training and test sets are simulated with the following true coefficients.



- Notice the shrinkage and the bias in the ridge estimates of the coefficients.

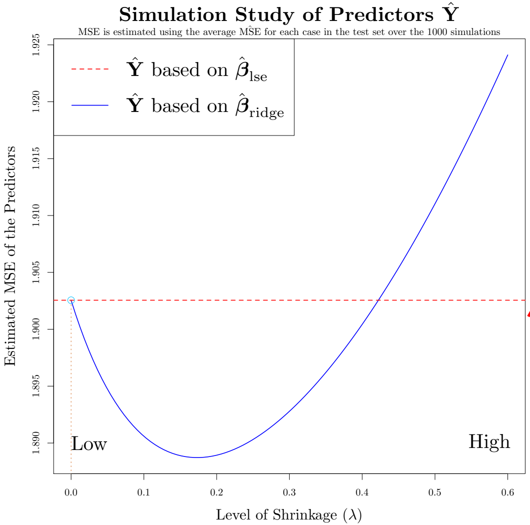
true beta



big coefficients are pushed down (penalty on norm of b)

- As we expected, there is an optimal λ according to the 1000 simulations.

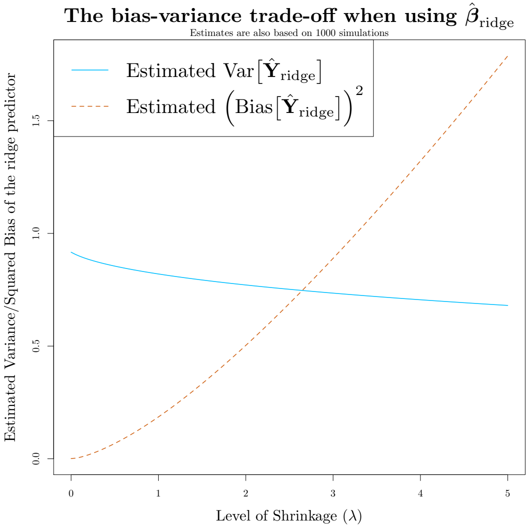
找best lambda
的方法只有试



use b_lse

use ridge

- As λ increases, the variance shrinks but the squared bias increases sharply.



bias变大很快，
variance慢慢变小

- The simulation result is a direction consequence of the following theorem.

Theorem 0.4 reason of the previous slide

Let $\hat{\boldsymbol{\theta}}_\ell$ for $\ell = 1, 2$ denote two distinct estimators of vector $\boldsymbol{\theta}$ with second order moments and generalised mean square errors:

generalized MSE

$$\mathbf{M}_\ell = \mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}\right)^{\mathrm{T}}\right] \quad \text{and} \quad \mathrm{GMSE}\left[\hat{\boldsymbol{\theta}}_\ell\right] = \mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}\right)^{\mathrm{T}} \mathbf{A} \left(\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}\right)\right]$$

where $\mathbf{A} \succeq 0$, that is, $\mathbf{A}$ is a positive semi-definite matrix, then $\mathrm{M}_1 - \mathrm{M}_2 \succeq 0$ if and only if

$$\mathrm{GMSE}\left[\hat{\boldsymbol{\theta}}_1\right] - \mathrm{GMSE}\left[\hat{\boldsymbol{\theta}}_2\right] \geq 0 \quad \text{for all} \quad \mathbf{A} \succeq 0$$

positive definite matrix

If $\boldsymbol{\beta}$ is fixed but unknown and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$, then there is an optimal $\lambda_{\mathrm{opt}} > 0$ so that

$$\mathbf{M}_{\mathrm{lse}} - \mathbf{M}_{\mathrm{ridge}} \succ 0 \quad \text{and} \quad \boxed{\mathrm{GMSE}\left[\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}(\lambda_{\mathrm{opt}})\right] < \mathrm{GMSE}\left[\hat{\boldsymbol{\beta}}_{\mathrm{lse}}\right]} \quad \text{for all} \quad \mathbf{A} \succeq 0$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{lse}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}(\lambda) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

- Notice the last Theorem only guarantees the existence of the optimal λ_{opt}.
- Suppose the data is big in the sense that k is large, but n is relatively small.

Q: How to determine the shrinkage parameter λ_{opt}?

- Given a dataset $\left\{\mathbf{X}_{n\times(k+1)}, \mathbf{y}_{n\times 1}\right\}$, and a simple training-test split

$$\left\{\mathbf{X}^{\dagger}_{(n-m)\times(k+1)}, \mathbf{X}^{*}_{m\times(k+1)}, \mathbf{y}^{\dagger}_{(n-m)\times(k+1)}, \mathbf{y}^{*}_{m\times(k+1)}\right\}$$

for any value of $0 \leq \lambda < \infty$, we can estimate $\mathrm{MSE}(\hat{Y}_i)$ for arbitrary i by

试出来

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

$$\hat{\mathrm{MSE}} = \frac{1}{m}\left\|\mathbf{y}^{*} - \mathbf{X}^{*}\mathbf{b}_{\text{ridge}}\right\|^2$$

- Of course, we can not only do this for a value of λ, but a set of values of λ, thus choose the λ that numerically minimises $\hat{\mathrm{MSE}}$ to be λ_{opt}.