Multicollinearity

Regularization

Ridge and Lasso

서 상수

Multicollinearity

Introduction

Problem

Detection

Remedy

Multicollinearity

Introduction

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related.

$$\text{Salary} = \beta_0 + \beta_1(Age) + \beta_2(Career) + \varepsilon$$

β_i: The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant.

Multicollinearity

Problem

Perfect multicollinearity

- Consequences of Multicollinearity

$$y_i = \alpha + \beta X_i + \gamma Z_i + u_i$$

  - Least Squares Estimator for β

$$\hat{\beta} = \frac{S_{zz}S_{xy} - S_{xz}S_{zy}}{S_{xx}S_{zz} - S_{xz}^2}$$

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2 S_{zz}}{S_{xx}S_{zz} - S_{xz}^2}$$

where $S_{xx} \equiv \sum (X_i - \bar{X})^2$, $S_{zz} \equiv \sum (Z_i - \bar{Z})^2$, $S_{xz} \equiv \sum (X_i - \bar{X})(Z_i - \bar{Z})$,

$S_{xy} \equiv \sum (X_i - \bar{X})(Y_i - \bar{Y})$, and $S_{zy} \equiv \sum (Z_i - \bar{Z})(y_i - \bar{y})$.

- Consequences of Perfect Multicollinearity

Suppose that $Z_i = a + bX_i$. Then,

$$S_{xz} \equiv \sum (X_i - \bar{X})(Z_i - \bar{Z}) = \sum (X_i - \bar{X})(a + bX_i - a - b\bar{X}) = bS_{xx}$$

$$S_{zz} \equiv \sum (Z_i - \bar{Z})^2 = \sum (a + bX_i - a - b\bar{X})^2 = b^2 S_{xx}$$

Thus,

$$\hat{\beta} = \frac{S_{zz}S_{xy} - S_{xz}S_{zy}}{S_{xx}(b^2 S_{xx}) - (bS_{xx})^2} = \frac{S_{zz}S_{xy} - S_{xz}S_{zy}}{0} \text{ : Not Computable}$$

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2 S_{zz}}{S_{xx}S_{zz} - S_{xz}^2} = \frac{\sigma^2 S_{zz}}{S_{xx}(b^2 S_{xx}) - (bS_{xx})^2} = \frac{\sigma^2 S_{zz}}{0} = \infty$$

Multicollinearity

Problem

Perfect multicollinearity

```
> X
     X1 X2
[1,]  1  5
[2,]  2 10
[3,]  3 15
[4,]  4 20
[5,]  5 25
```

```
> t(X)%*%X
     X1   X2
X1   55  275
X2 275 1375
```

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$(X'X)^{-1}$ *incoumputable*

```
> solve(t(X)%*%X)
Error in solve.default(t(X) %*% X) :
  Lapack routine dgesv: system is exactly singular: U[2,2] = 0
```

Theorem 4.2.7 A Unifying Theorem

If A is an $n \times n$ matrix, then the following statements are equivalent.

1. The reduced row echelon form of A is I_n.
2. A is expressible as a product of elementary matrices.
3. A is invertible.
4. $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
5. $A\mathbf{x} = \mathbf{b}$ is consistent for every vector $\mathbf{b}$ in $\mathbb{R}^n$.
6. $A\mathbf{x} = \mathbf{b}$ has exactly one solution for every vector $\mathbf{b}$ in $\mathbb{R}^n$.
7. The column vectors of A are linearly independent.
8. The row vectors of A are linearly independent.
9. $\det(A) \neq 0$.

Multicollinearity

Problem

Near multicollinearity

- Consequences of Near (Imperfect) Multicollinearity

  Suppose that $Z_i \approx a + bX_i$. Then,

  $$S_{xz} \approx bS_{xx} \quad \text{and} \quad S_{zz} \approx b^2 S_{xx}$$

  Thus,

  $$\hat{\beta} \approx \frac{S_{zz}S_{xy} - S_{xz}S_{zy}}{0}$$

  $$\text{Var}(\hat{\beta}) \approx \frac{\sigma^2 S_{zz}}{0} \to \infty$$

  - t-test for H_0: $\beta = 0$

    $$t = \frac{\hat{\beta}}{\sqrt{\hat{\text{Var}}(\hat{\beta})}} \approx \frac{\hat{\beta}}{\infty} \to 0$$

$(X'X)^{-1}$ *coumputable*

$$\det(X'X) \approx 0$$

$$var(\hat{\beta}) = \delta^2 (X'X)^{-1} \approx \infty$$

Unable to reject H_0, not because the variable has no effects but because the sample is not good enough to isolate the effect of the variable.

Multicollinearity

Problem

Near multicollinearity

$$\text{Salary} = \beta_0 + \beta_1(Age) + \beta_2(Career) + \varepsilon$$

	Coef	S.E	t Stat	P-value
Intercept	19074	51499	0.37	0.72
Age	3886	2093	1.85	0.10
Career	2023	1928	1.04	0.32

Even though R^2 is high, model reliability is low

Multicollinearity

Detection

Correlation

	X1	X2	X3
X1	1		
X2	0.91	1	
X3	0.4	-0.2	1

Multicollinearity

Detection

Condition Number

$$CN(X_1, X_2, \ldots, X_k) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

where λ_{max} (λ_{min}) is the maximum (minimum) eigenvalue of (X′X) matrix after normalization which makes $\lambda_{max} = 1$. [CN is sometimes defined without the square root]

- If (X′X) matrix is diagonal (no multicollinearity at all), then $\lambda_{min} = 1$, thus, CN = 1.
- If (X′X) matrix is singular (perfect multicollinearity), then $\lambda_{min} = 0$, thus, CN $\rightarrow \infty$.
- Belsley proposes the following *guideline*:
  - If CN<10, weak multicollinearity
  - If 10<CN<30, moderate to strong multicollinearity
  - If CN>30, severe multicollinearity

Belsley, D.A. E. Kuh and R.H. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, NY, 1980.

Multicollinearity

Detection

Theil's m

$$m = R^2 - \sum_{j=1}^{k} (R^2 - R^2_{-j})$$

where R^2 is from the regression of y on the other explanatory variables (X_1, X_2, … X_k), and R^2_{-j} is from the regression of y on (X_1, …, X_{j-1}, X_{j+1}, … X_k).

- If X_j is perfectly collinear with other explanatory variables, $R^2 = R^2_{-j}$.
- $(R^2 - R^2_{-j})$ is the 'exclusive' explanation of y by X_j (beyond all the other explanatory variables). If there exists no overlapped influence (all the explanatory variables are independent), then $\sum (R^2 - R^2_{-j}) = R^2$ so that $m = 0$.
- Thus, roughly, $0 \le m \le R^2 \le 1$.

Multicollinearity

Detection

Variance Inflation Factor

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

1) Create regression models for each X variable

$$X_1 = \beta_0^* + \beta_1^* X_2 + \beta_2^* X_3 + \beta_3^* X_4 + \varepsilon^*$$
$$X_2 = \beta_0^{**} + \beta_1^{**} X_1 + \beta_2^{**} X_3 + \beta_3^{**} X_4 + \varepsilon^{**}$$
$$\vdots$$

2) Find VIF by R^2 from each regression

$$VIF = \frac{1}{1 - R_k^2}$$

Multicollinearity

Detection

Variance Inflation Factor

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

$$VIF = \frac{1}{1 - R_k^2}$$

Higher R^2, Higher VIF

	VIF
X1	3.1
X2	1.42
X3	12.05
X4	1.91

Multicollinearity

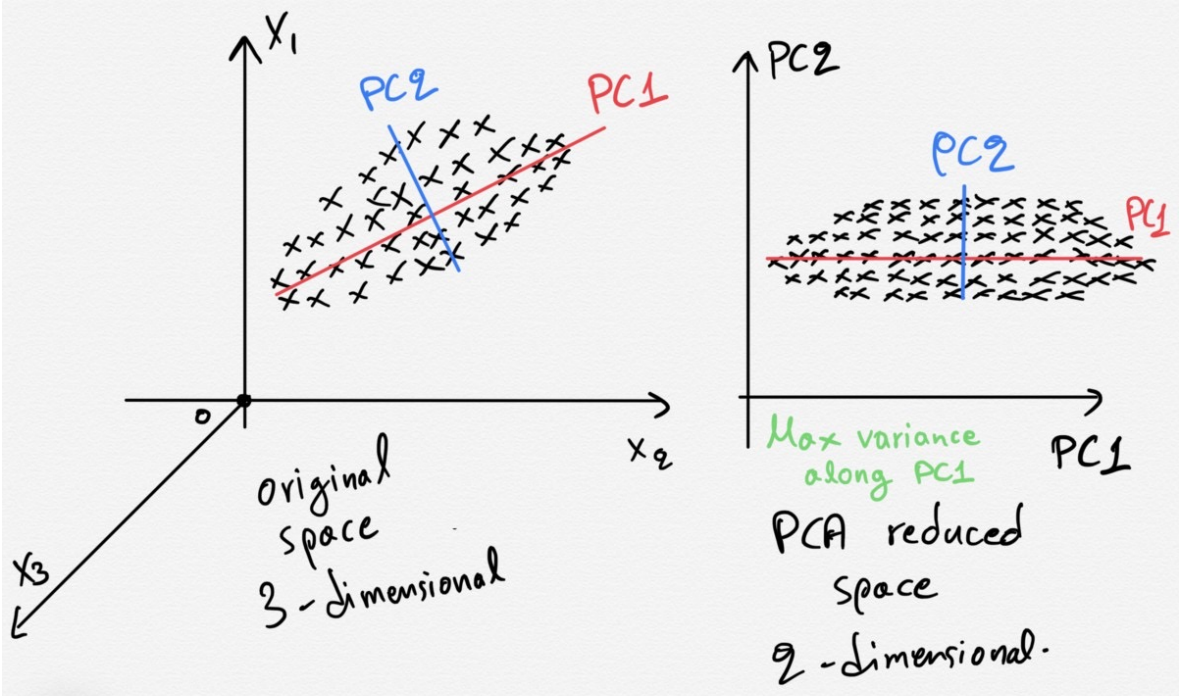
Remedy

Do nothing

Remove correlated variable

PCA

Regularization



Multicollinearity

Remedy

Regularization

Ridge(L2) regression

Lasso(L1) regression

Review

What is a good model?

Interpretation

Minimize training error

$$MSE = (Y - \hat{Y})^2$$

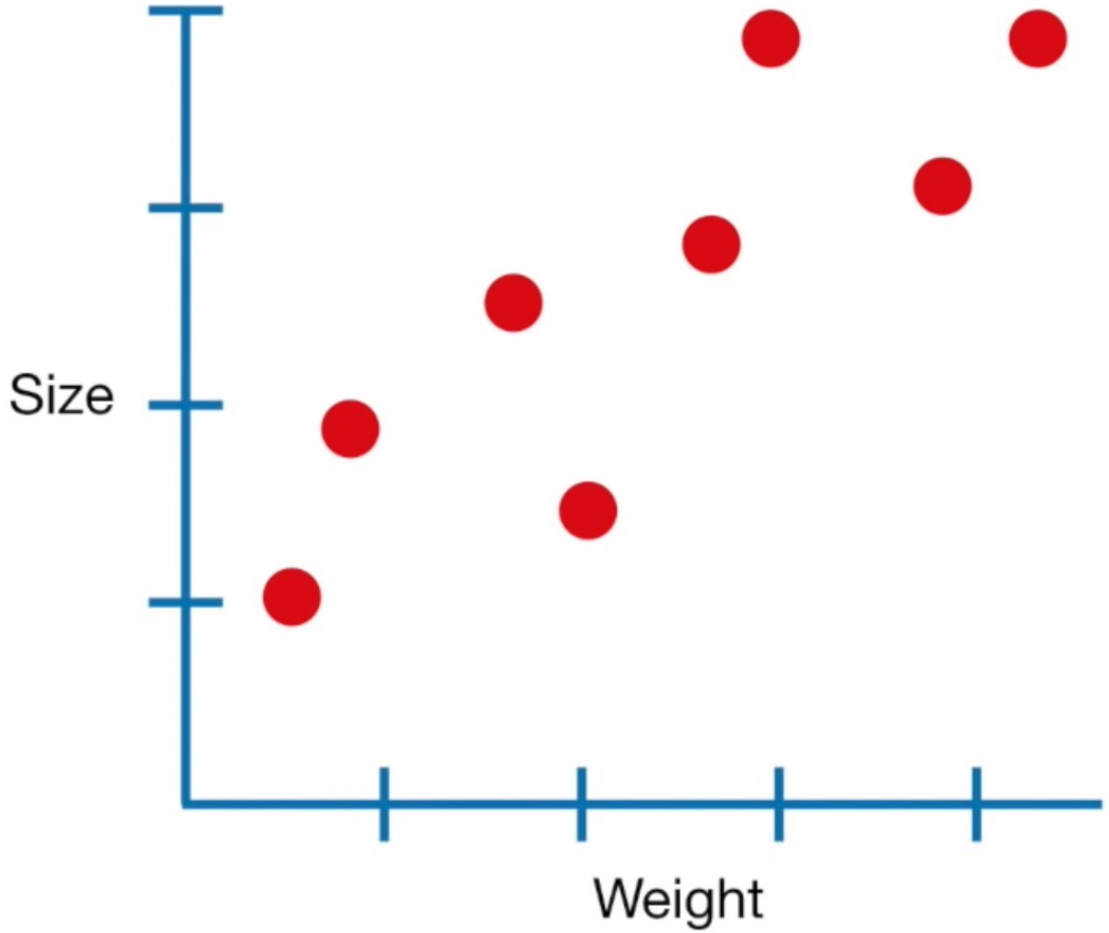
Prediction

Minimize test error

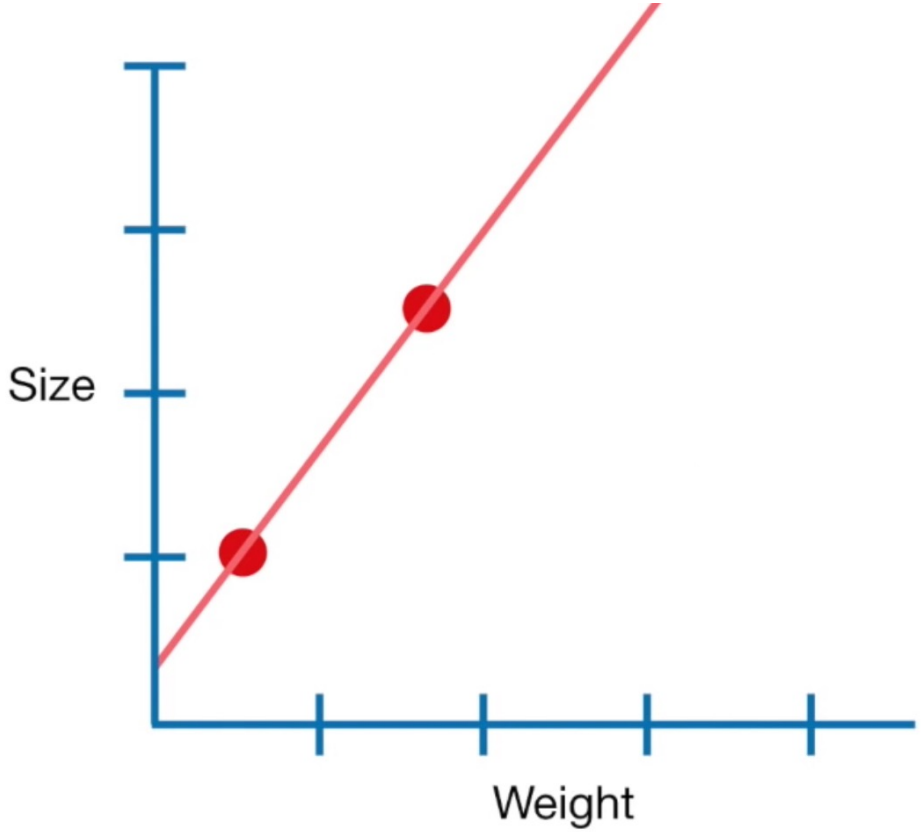
$$\begin{aligned}\text{Expected MSE} &= E\left[(Y - \hat{Y})^2 | X\right] \\ &= \sigma^2 + (E[\hat{Y}] - \hat{Y})^2 + E[\hat{Y} - \mathrm{E}[\hat{Y}]]^2 \\ &= \sigma^2 + \mathrm{B}ias^2(\hat{Y}) + \mathrm{Var}(\hat{Y}) \\ &= \text{Irreducible Error} + \mathrm{B}ias^2 + \text{Variance}\end{aligned}$$

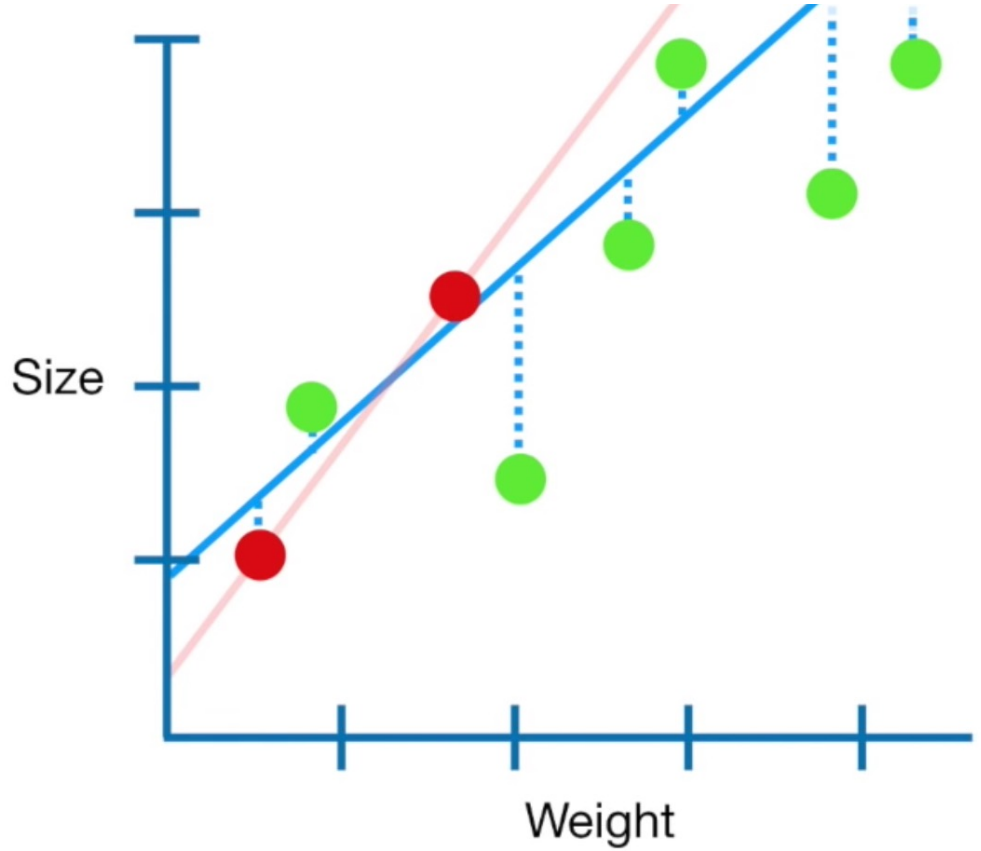
Review

What is a good model?



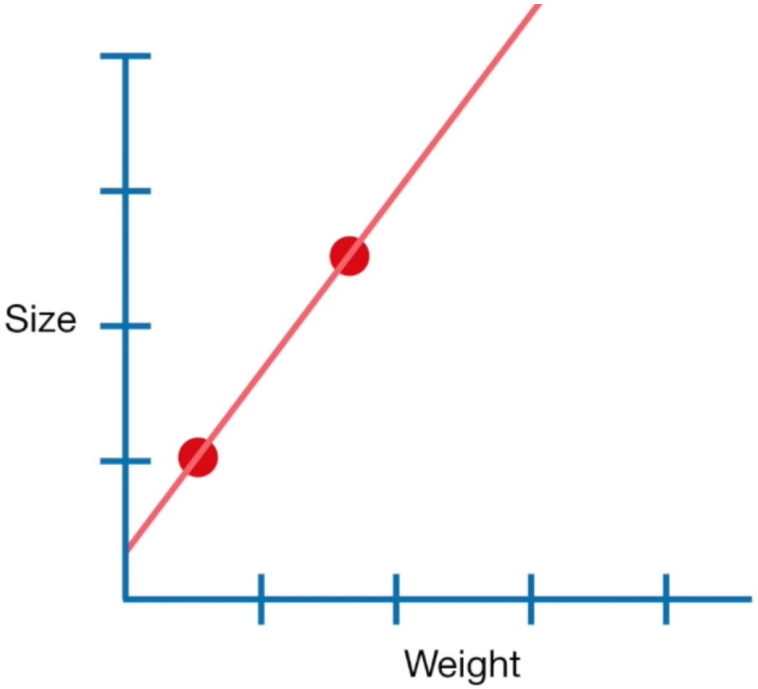
Regularization

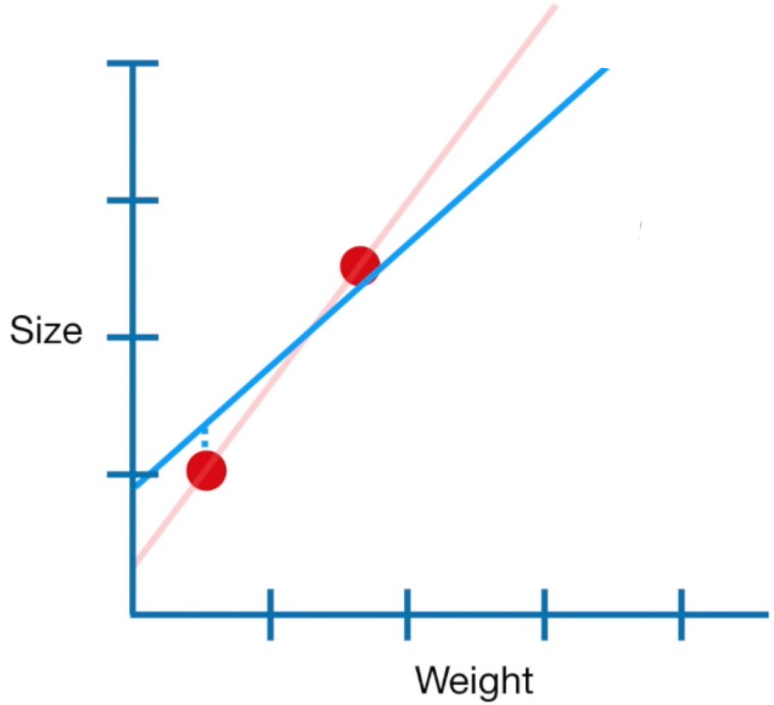




Regularization

$$Size = \beta_0 + \beta_1(Weight)$$





$$\text{SSE} + \lambda \times \beta_i^2$$

Ridge

$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1}^{} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$$

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

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

$$= (X'X + \lambda I)^{-1} X'Y$$

Ridge

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1} X'Y$$

```
> X
     X1 X2
[1,]  1  5
[2,]  2 10
[3,]  3 15
[4,]  4 20
[5,]  5 25
```

```
> t(X)%*%X
     X1   X2
X1   55  275
X2 275 1375
```

```
> solve(t(X)%*%X + 2*diag(2))
                X1           X2
X1   0.48079609 -0.09601955
X2 -0.09601955  0.01990223
```

```
> solve(t(X)%*%X)
Error in solve.default(t(X) %*% X) :
  Lapack routine dgesv: system is exactly singular: U[2,2] = 0
```

$$MSE(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

$$= \sum_{i=1}^{n} {y_i}^2 - 2\sum_{i=1}^{n} y_i(\beta_1 x_{i1} + \beta_2 x_{i2}) + \sum_{i=1}^{n} (\beta_1 x_{i1} + \beta_2 x_{i2})^2$$

$$= \sum_{i=1}^{n} {y_i}^2 - 2\left(\sum_{i=1}^{n} y_i x_{i1}\right)\beta_1 - 2\left(\sum_{i=1}^{n} y_i x_{i2}\right)\beta_2 + \sum_{i=1}^{n} (\beta_1^2 x_{i1}^2 + \beta_2^2 x_{i2}^2 + 2\beta_1\beta_2 x_{i1} x_{i2})$$

$$= \left(\sum_{i=1}^{n} x_{i1}^2\right)\beta_1^2 + \left(\sum_{i=1}^{n} x_{i2}^2\right)\beta_2^2 + \left(2\sum_{i=1}^{n} x_{i1} x_{i2}\right)\beta_1\beta_2$$

$$-2\left(\sum_{i=1}^{n} y_i x_{i1}\right)\beta_1 - 2\left(\sum_{i=1}^{n} y_i x_{i2}\right)\beta_2 + \sum_{i=1}^{n} {y_i}^2$$

$$= A\beta_1^2 + B\beta_1\beta_2 + C\beta_2^2 + D\beta_1 + E\beta_2 + F$$ Conic equation (2차원의 경우)

$$A\beta_1^2 + B\beta_1\beta_2 + C\beta_2^2 + D\beta_1 + E\beta_2 + F = 0$$

Discriminant of conic equation (판별식): B^2-4AC

B^2-4AC = 0 → parabola (포물선)
B^2-4AC > 0 → hyperbola (쌍곡선)
B^2-4AC < 0 → ellipse (타원)
B = 0 and A=C → circle (원)

$$MSE(\beta_1, \beta_2) = \left(\sum_{i=1}^{n} x_{i1}^2\right)\beta_1^2 + \left(\sum_{i=1}^{n} x_{i2}^2\right)\beta_2^2 + \left(2\sum_{i=1}^{n} x_{i1}x_{i2}\right)\beta_1\beta_2 - 2\left(\sum_{i=1}^{n} y_i x_{i1}\right)\beta_1 - 2\left(\sum_{i=1}^{n} y_i x_{i2}\right)\beta_2 + \sum_{i=1}^{n} {y_i}^2$$

$$B^2 - 4AC = \left(2\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2 - 4\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2$$

$$= 4\left\{\left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2 - \sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2\right\} < 0$$

By Cauchy-Schwartz inequality

Cauchy-Schwartz Inequality

$$X = [x_1, \dots, x_n]$$
$$Y = [y_1, \dots, y_n]$$

$$\sum x_i^2 \sum y_i^2 \geq [\sum x_i y_i]^2$$

The Cauchy–Schwarz inequality states that for all vectors u and v of an inner product space it is true that

$$|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle,$$ **(Cauchy-Schwarz inequality [written using only the inner product])**

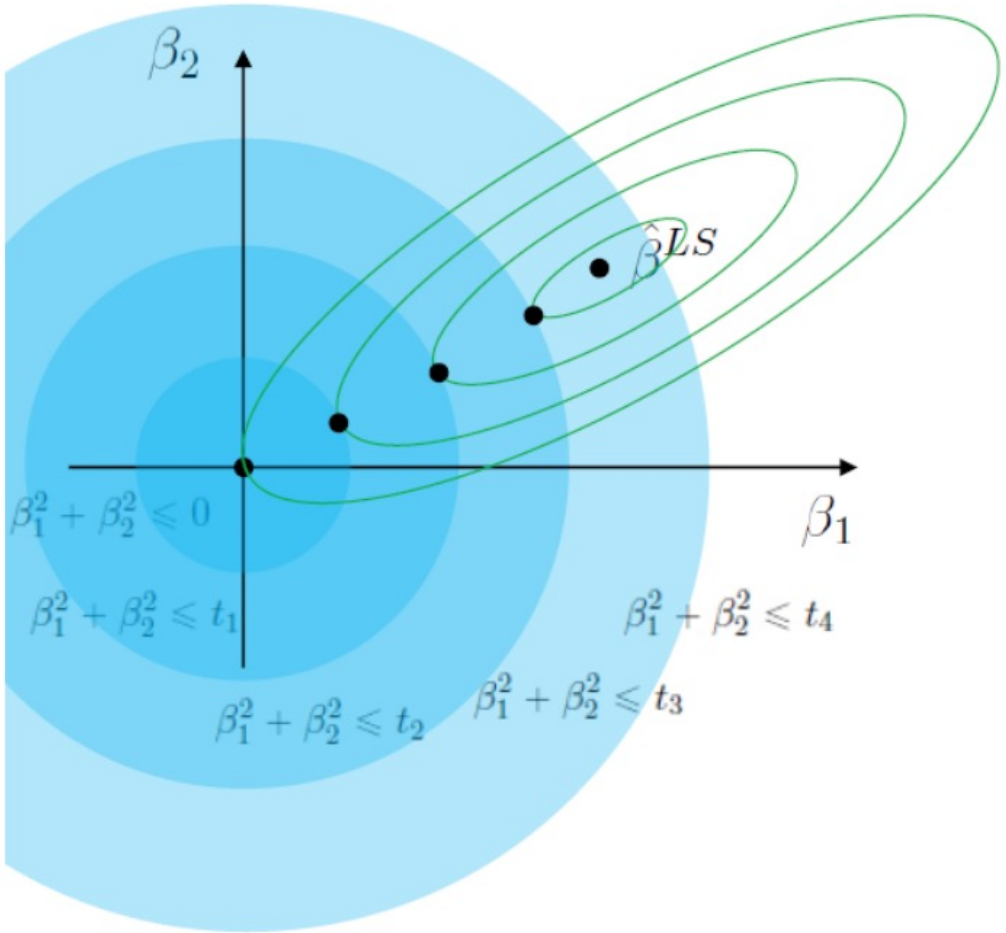
where $\langle \cdot, \cdot \rangle$ is the inner product. Examples of inner products include the real and complex dot product; see the examples in inner product. Every inner product gives rise to a norm, called the *canonical* or *induced norm*, where the norm of a vector $\mathbf{u}$ is denoted and defined by:

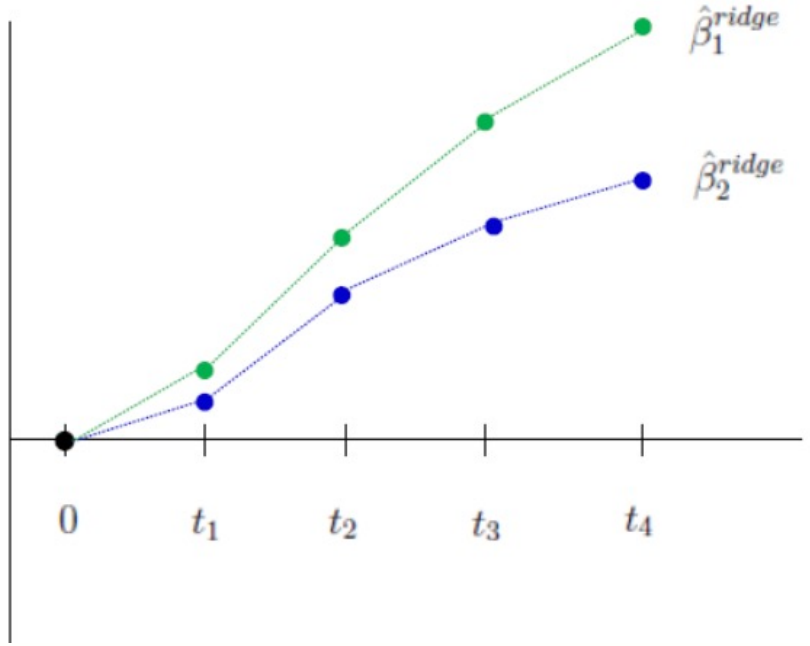
$$\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$$

so that this norm and the inner product are related by the defining condition $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$, where $\langle \mathbf{u}, \mathbf{u} \rangle$ is always a non-negative real number (even if the inner product is complex-valued). By taking the square root of both sides of the above inequality, the Cauchy–Schwarz inequality can be written in its more familiar form:[3][4]

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$ **(Cauchy-Schwarz inequality [written using norm and inner product])**

Moreover, the two sides are equal if and only if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent.[5][6]



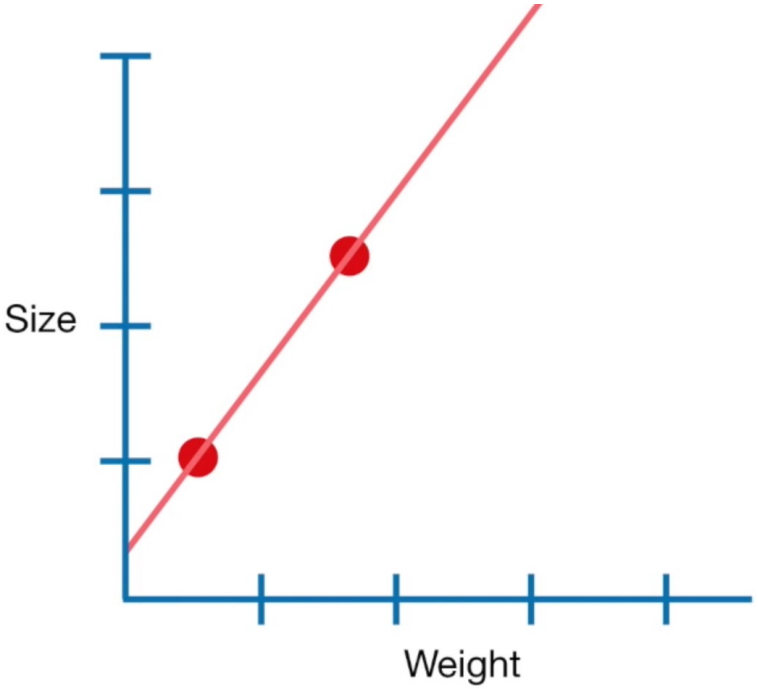


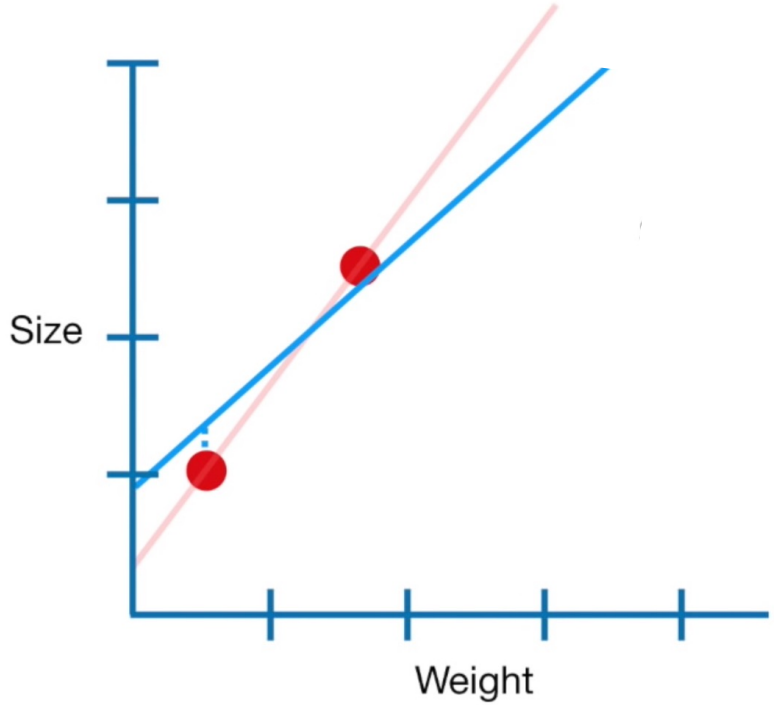
$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

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

Regularization

$$Size = \beta_0 + \beta_1(Weight)$$





$$\text{SSE} + \lambda \times |\beta_i|$$

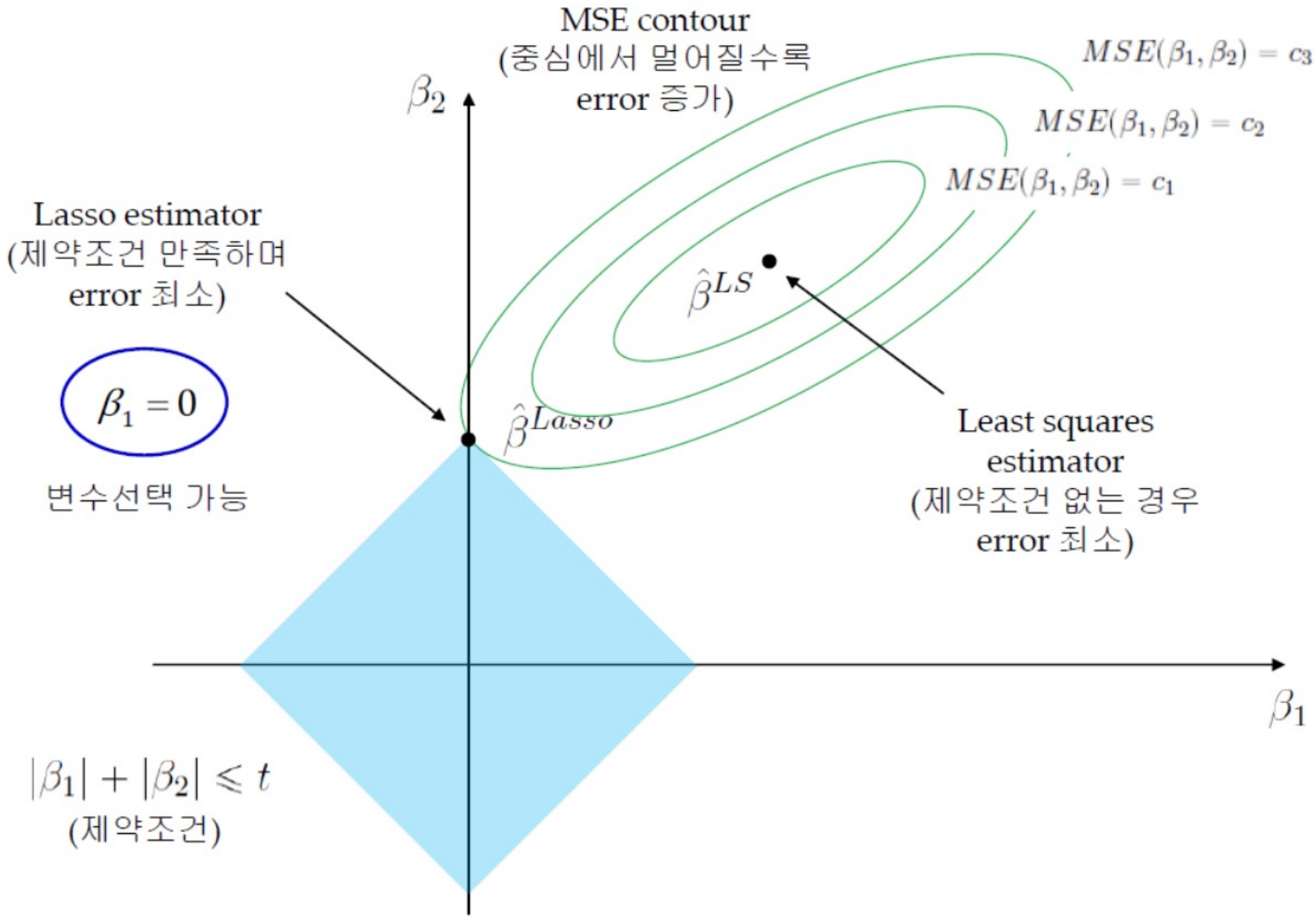
Lasso

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

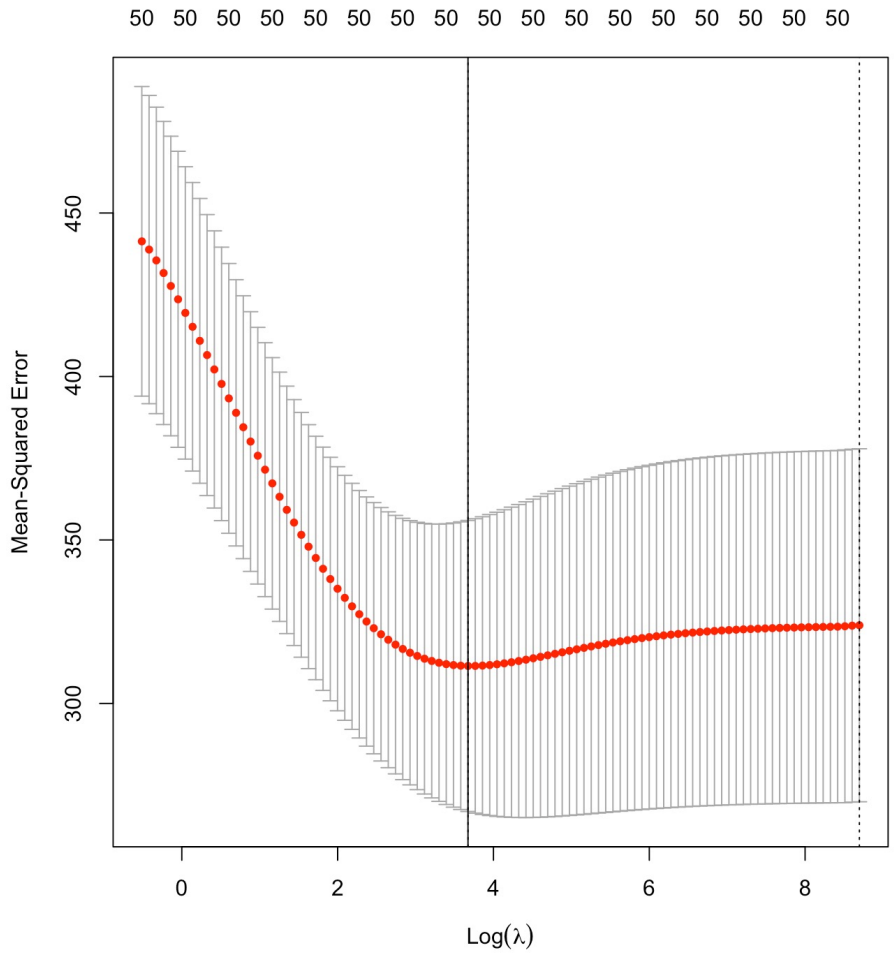
$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

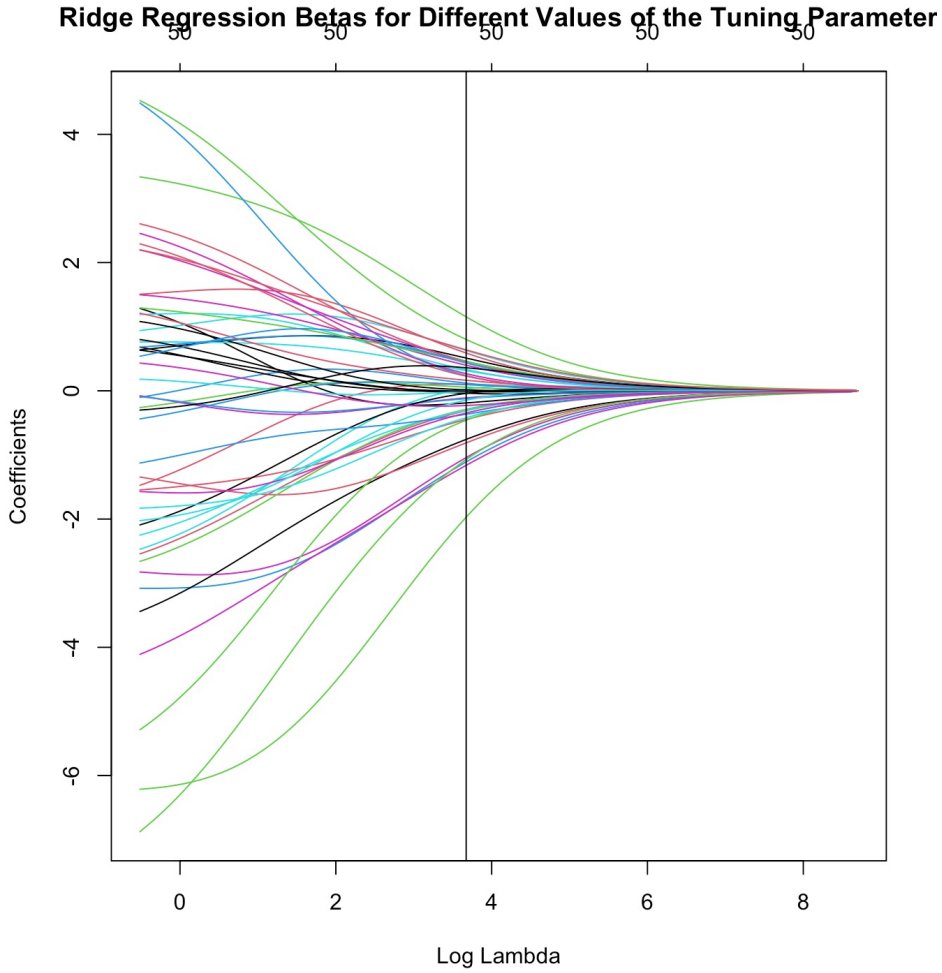
$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad \Rightarrow \quad \hat{\beta}^{lasso} = \ ?$$

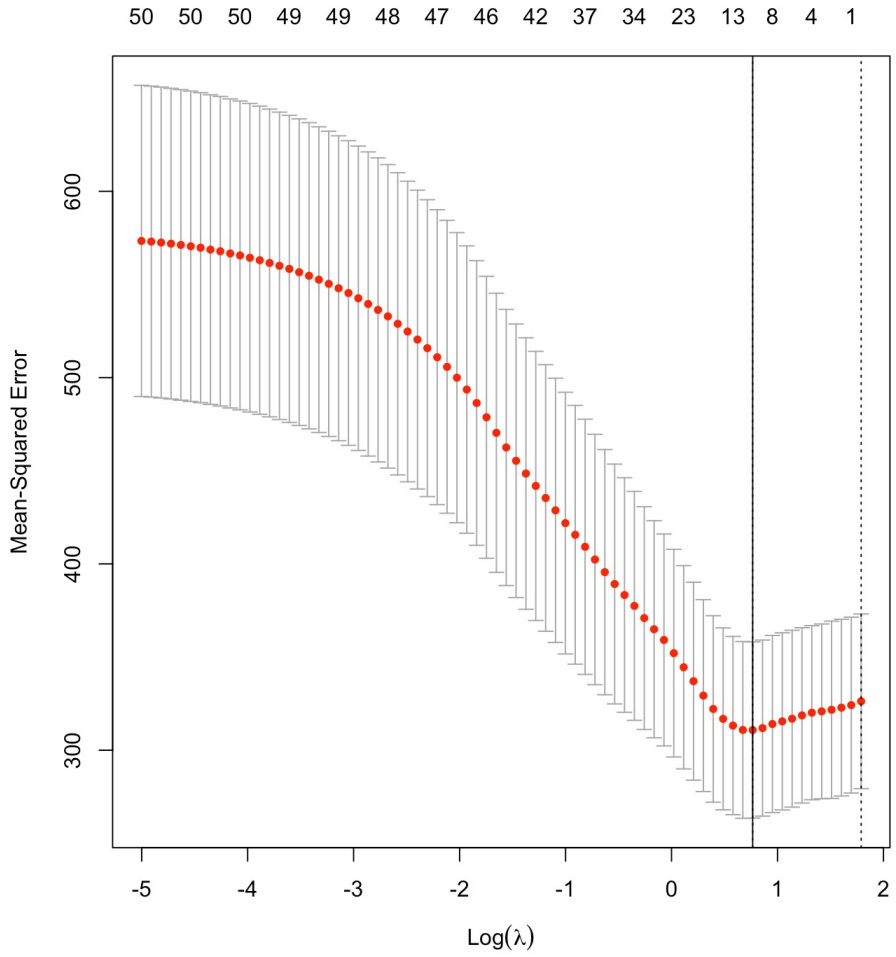

MSE contour
(중심에서 멀어질수록
error 증가)
$MSE(\beta_1, \beta_2) = c_3$
$MSE(\beta_1, \beta_2) = c_2$
$MSE(\beta_1, \beta_2) = c_1$
β_2
Lasso estimator
(제약조건 만족하며
error 최소)
$\hat{\beta}^{LS}$
$\hat{\beta}^{Lasso}$
$\beta_1 = 0$
변수선택 가능
Least squares
estimator
(제약조건 없는 경우
error 최소)
β_1
$|\beta_1| + |\beta_2| \leqslant t$
(제약조건)

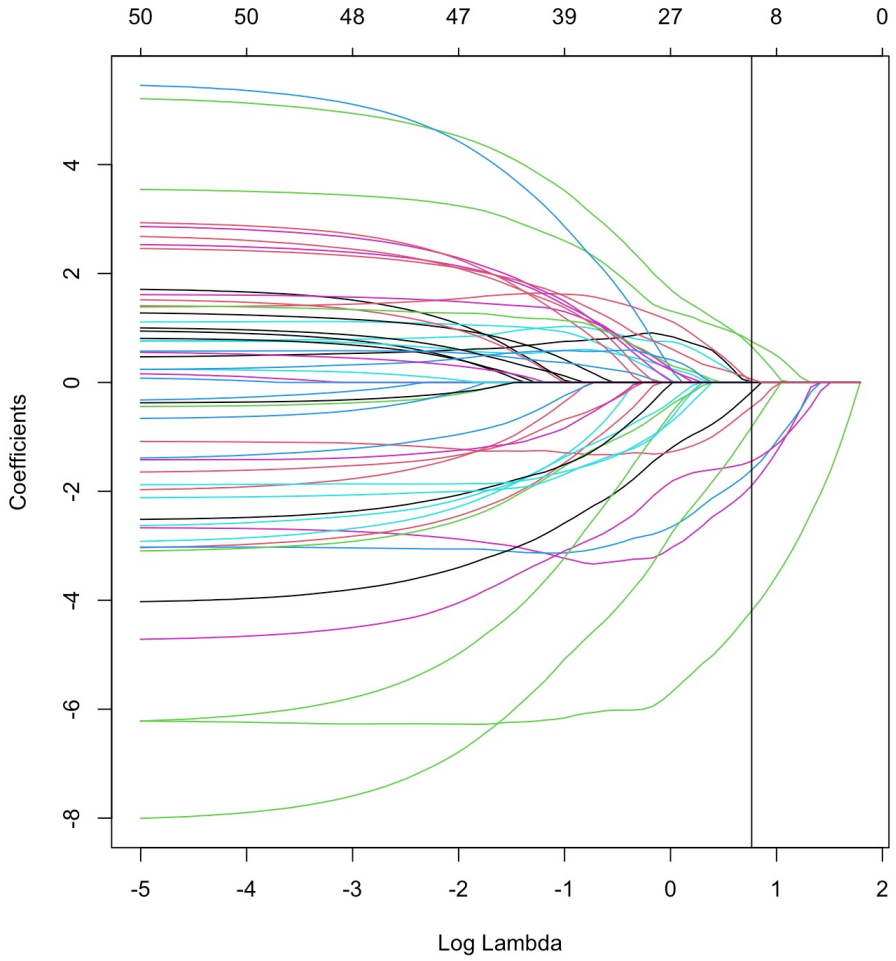
Cross-Validation





Cross-Validation





HW ??

Shrinkage Methods of Linear Regession : Ridge and LASSO

연세대학교 통계데이터사이언스 석사과정 이청파 (leechungpa@naver.com)

Review

Test (=generalization) **error** : predictoin error over an independent test sample (X, Y)

$$Err_\tau = E_{X,Y}[L(Y, \hat{f}_\tau(X))|\tau]$$

Expected prediction error : The randomness in the training set τ is averaged over.

$$Err = E_{X,Y}[Err_\tau] = E_{X,Y}[L(Y, \hat{f}_\tau(X))]$$

Training error : optimistic estimate of the test error Err_τ

$$\bar{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}_\tau(x_i))$$

In-sample error

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(y_i^0, \hat{f}_\tau(x_i))|\tau]$$

Linear model case : $\hat{f}_\tau(x) = x^T\hat{\beta}$ **where** $Y = f(X) + \epsilon$

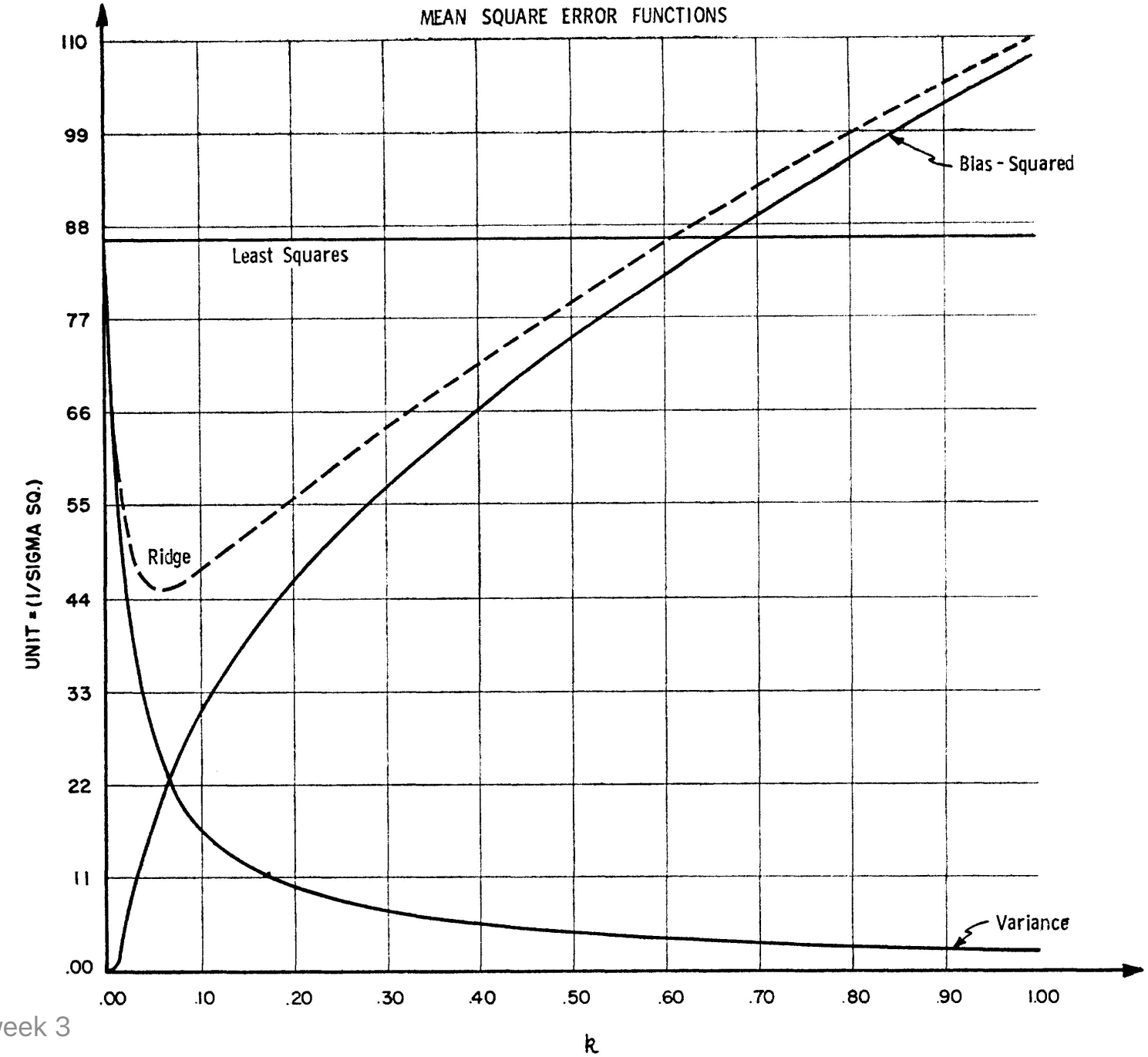
Expected prediction error of a linear model with L^2 loss function at $X = x_0$:

$$\begin{aligned} Err(x_0) &= E_Y[L(Y, \hat{f}_\tau(X))|X = x_0] \\ &= \sigma_\epsilon^2 + \left\{f(x_0) - E[x_0^T\hat{\beta}]\right\}^2 + \left\|X(X^TX)^{-1}x_0\right\|_2^2\,\sigma_\epsilon^2 \end{aligned}$$

In above linear regeression case, we can decompose the average squared bias.

$$E_{x_0}\left\{f(x_0) - E[x_0^T\hat{\beta}]\right\}^2 = E_{x_0}\underbrace{\left\{f(x_0) - x_0^T\hat{\beta}^{\text{BLUE}}\right\}^2}_{\text{model bias}} + E_{x_0}\underbrace{\left\{x_0^T\hat{\beta}^{\text{BLUE}} - E[x_0^T\hat{\beta}]\right\}^2}_{\text{estimation bias}}$$

- $\hat{\beta}^{\text{BLUE}}$: BLUE of linear regression ($\hat{\beta} = \arg\min_\beta \left\|y - X^T\beta\right\|_2^2$)
- When the linear model is ordinary least squres method, the estimation basis is 0.


MEAN SQUARE ERROR FUNCTIONS
110
99
88
77
66
55
44
33
22
11
.00
UNIT = (1/SIGMA SQ.)
Least Squares
Ridge
Bias - Squared
Variance
.00
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
k

Pros and cons of subset selection

Subset selection : best-subset, foward-stepwise, backward-stepwise selection, etc

$$\hat{\beta}^{\text{best-subset}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right\} \qquad \text{where } \|u\|_p = \sum_{i=1}^{N} |u_i|^0 \ (= \sum_{i=1}^{N} I(u_i \neq 0))$$

- Produces an interpretable model
- (Possibly) lower prediction error than full model

However,

- Subset selection is discrete process.
    - Variables are either retainded or discarded.
- Correlated variables often exhibits high variance.
    - When many there are many correlated vaiables, a widely large positive coefficient on one variable can be canceled by a similaraly large negative coefficient on its correlated.
    - High variance results to high prediction error than full model.

$\Longrightarrow$ More continuous and lower variability shrinkage methods are needed.

Ridge

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 \right\} \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t$$

$$\iff \quad (\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{ridge}} = \underset{\beta\in\mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 \right\} \quad \text{subject to } \|\beta\|_2^2 \le t$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (=0)$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{ridge}} = \underset{\beta\in\mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (=0)$$

Cf. standard l^p-norm : $\|u\|_p = \left(\sum_{i=1}^{N} |u_i|^p \right)^{1/p}$

Ridge shrinks the regression coefficients by imposing a penalty on their size.

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- $\lambda \geq 0$: complexity parameter
  - controls the amount of shirnkage
  - large λ shrinks coefficients toward 0
  - λ and t have a one-to-one correspondence
- β_0 is not included in the penalty term.
  - Penalization of the intercept would make the regression depend on the origin. (Consider simple regression with a negative slope.)

If we standarize X and y,

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} \quad \Longrightarrow \quad \hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (= 0)$$

- Standarizeds the inputs before solving.
  - The solutions are not equivariant under scaling of the inputs.
  - So X is $N \times p$ matrix, not $N \times (p+1)$.
  - $\hat{\beta}_0^{\text{ridge}}$ must be $\bar{y}$ $(= 0)$. (DIY)
- Ridge was first introduced in statisitcs to make $X^T X$ nonsingular (to solve linear regression), even if $X^T X$ is not of full rank.

Interpretations of Ridge : Singular Value Decomposition

X is decomposed by orthogonal matrices U, V and a diagonal matirx D.

$$X = UDV^T \quad \Longrightarrow$$

$$\begin{aligned} X\hat{\beta}^{\text{ridge}} &= X(X^TX + \lambda I)^{-1}X^Ty \\ &= UD(DD + \lambda I)^{-1}DU^Ty \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^Ty \end{aligned}$$

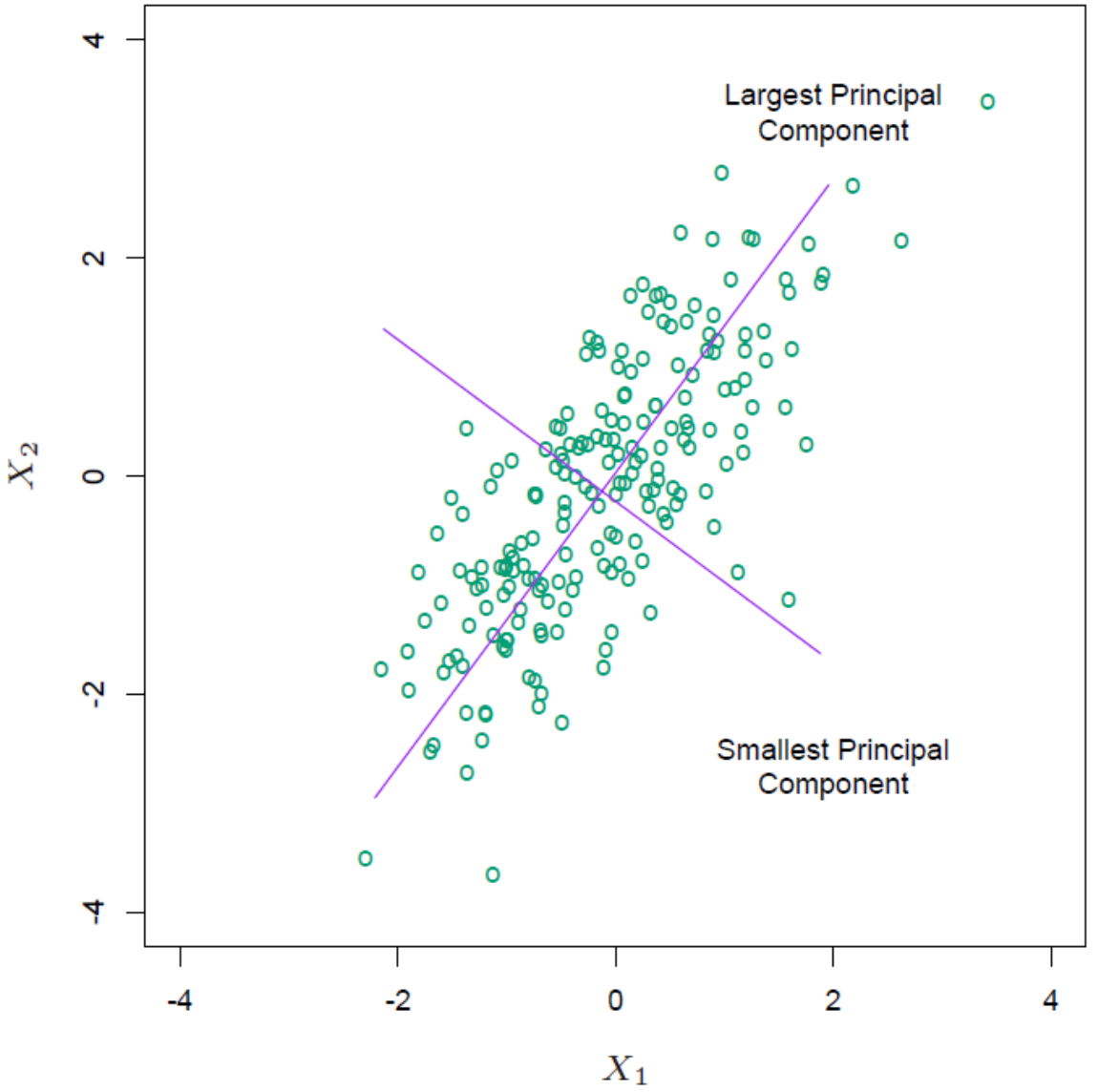
- $\mathbf{u}_j$ are the columns of U, which spans the column space of X.
- Comparing to linear regression ($X\hat{\beta} = UU^Ty$), $\frac{d_j^2}{d_j^2+\lambda}$ shrinkages j^{th} coordinate.

$$X^T X = VDDV^T$$

By using the pricipal componets of the variables X, there are eigen- vectors $\mathbf{v}_i$ also called the i^{th} principal component direction of X.

$$Var(\mathbf{z}_i) = Var(X\mathbf{v}_i) = \frac{d_i^2}{N}$$

The small singular values d_j has small variance, and ridge regression shrinks these directions the most.



The configuration of the data allow us to determine its gradient more accurately in the long direction than the short.

Ridge regression protects against the potentially high variance of gradients estimated in the short directions.

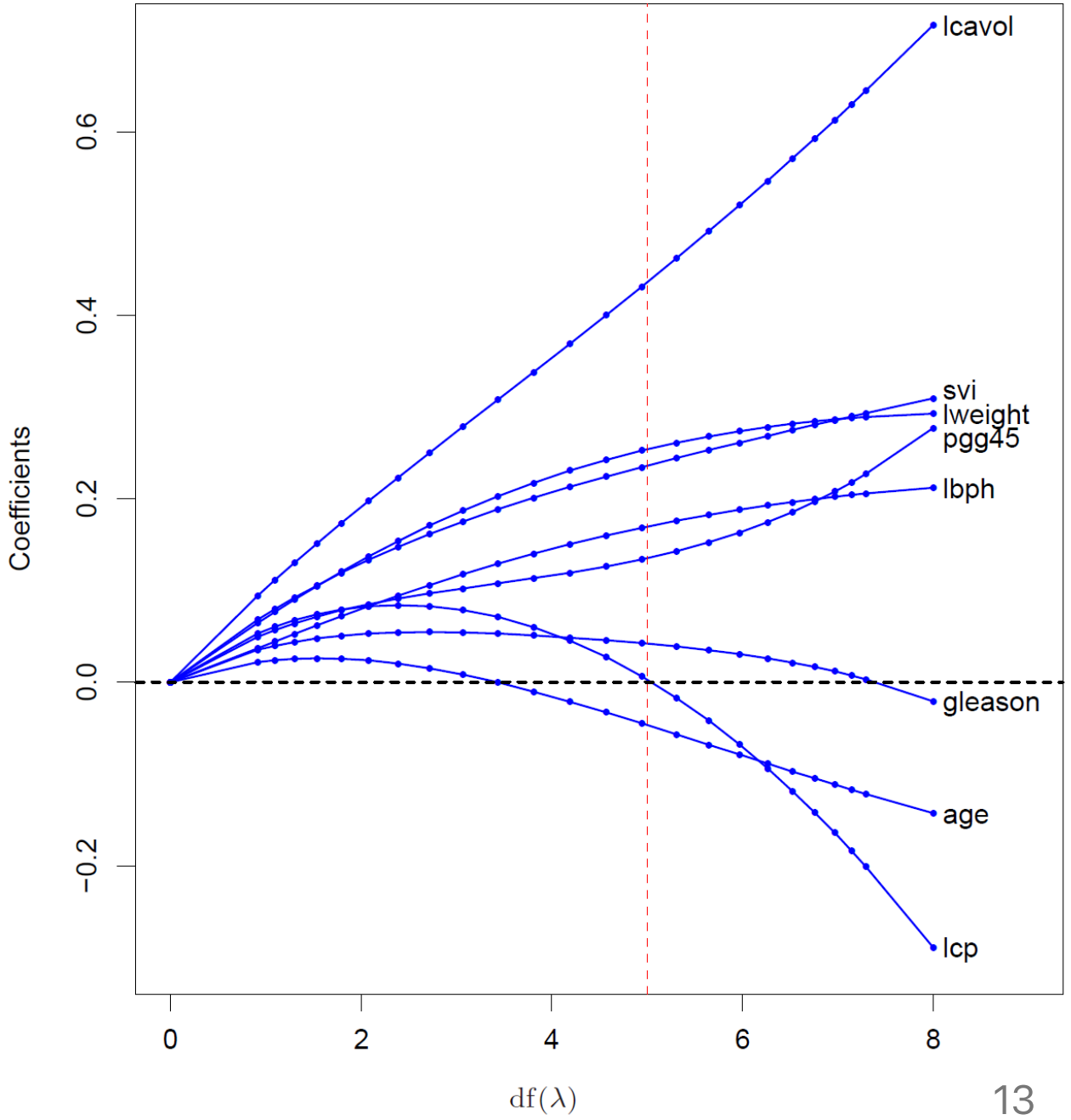
> **Implicit assumption of ridge :**
> **The response tend to vary most in the directions of high variance of inputs.**

More details about PCA will be on next week.

Effective degrees of freedom

$$
\begin{aligned}
df(\lambda) &\overset{\text{def}}{=} trace[H_\lambda^{\text{ridge}}] \\
&= trace[X(X^TX + \lambda I)^{-1}X^T] \\
&= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}
\end{aligned}
$$

- $df(\lambda)$ of linear regression is the number of variables p.
  - $df(\lambda) = p$ when no regularization $\lambda = 0$.



LASSO

Least Absolute Shrinkage and Selection Operator

$$(\hat{\beta}_0^{\text{lasso}}, \hat{\beta}^{\text{lasso}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \quad \hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 \right\} \quad \text{subject to } \|\beta\|_1 \leq t$$

$$\hat{\beta}_0^{\text{lasso}} = \bar{y} \ \ (=0)$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \quad \hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (=0)$$

If we standarize X and y,

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (= 0)$$

- Standarizeds the inputs before solving, likewise Ridge.
- LASSO use L^1 penalty instead of L^2
  - The solution of LASSO is nonlinear in the y. (No closed form)
  - The shrinkage method of LASSO makes the coefficient exactly 0, unlike Ridge.
  - The lasso does a kind of continuous subset selection.

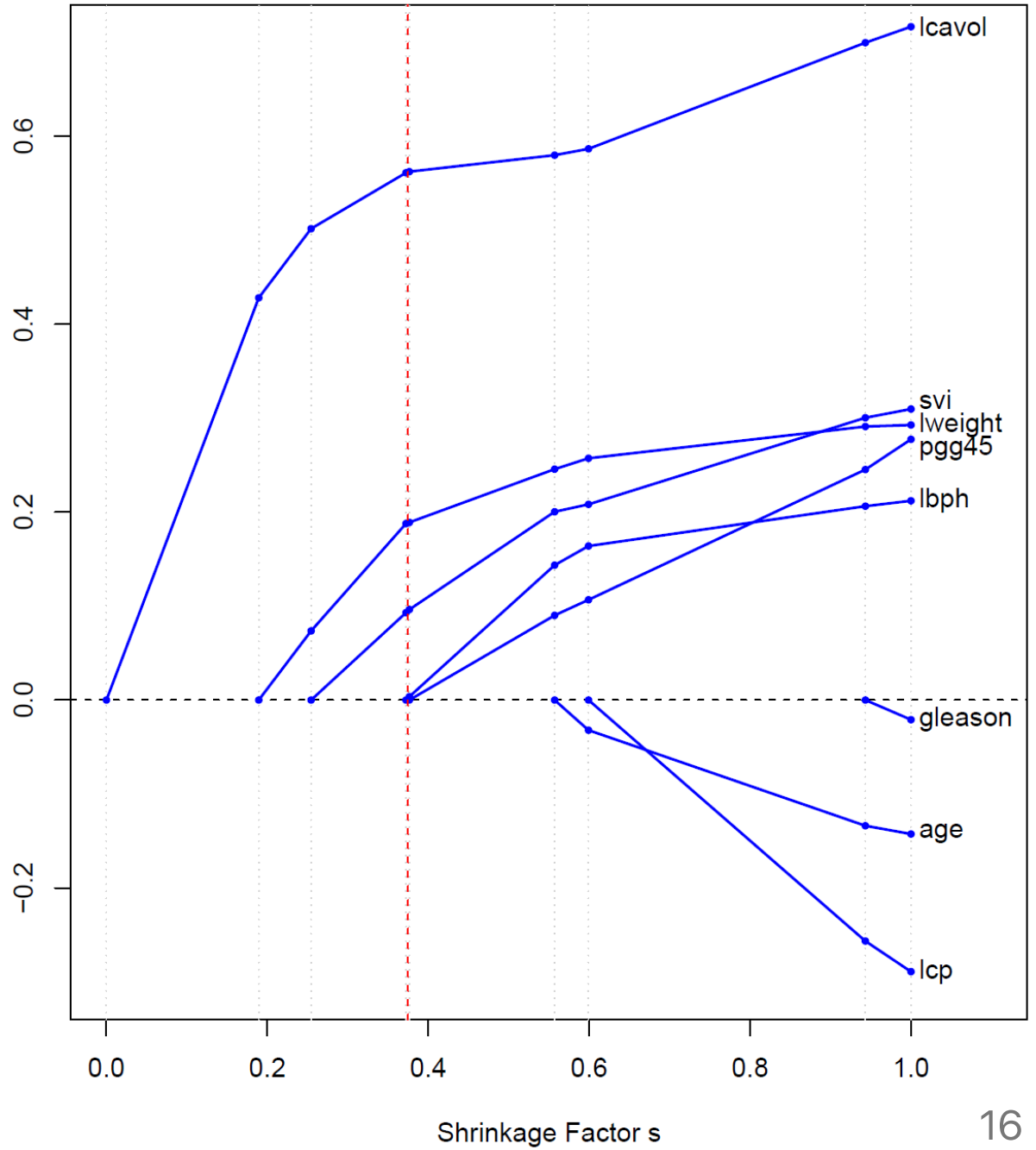
Interpretations of LASSO : Standardized shrinkage factor

Recall LASSO subject to $\|\beta\|_1 \leq t$.

If we choose t larger than $\|\hat{\beta}\|_1$ where $\hat{\beta}$ is LSE, then $\hat{\beta}^{\text{lasso}} = \hat{\beta}$.

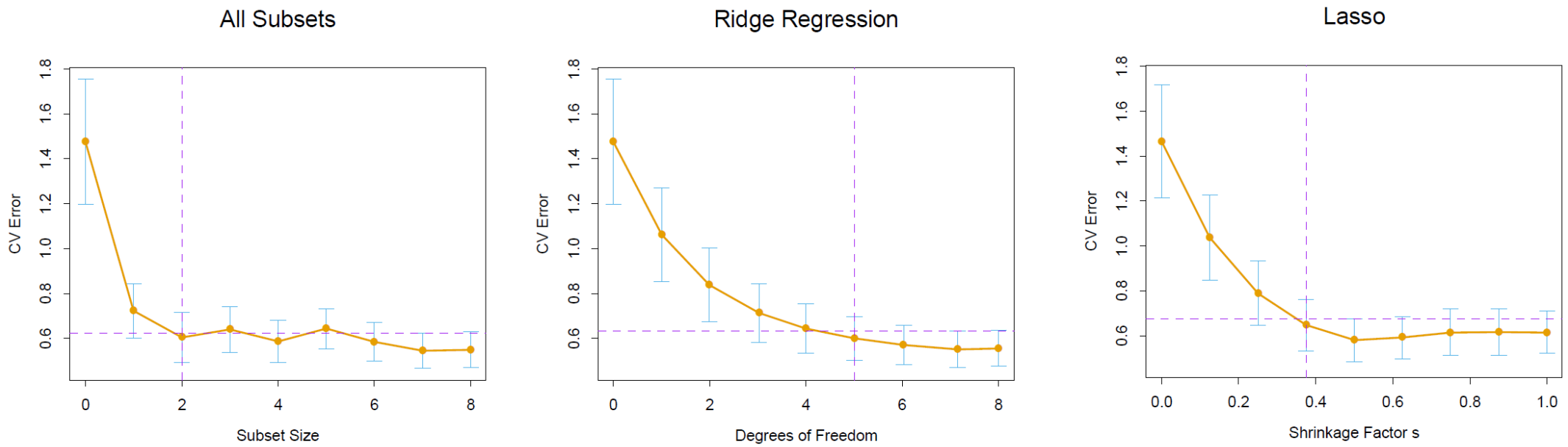
$$s \overset{\text{def}}{=} \frac{t}{\|\hat{\beta}\|_1}$$

Using maximum, standardize the shrinkage factor like above.



Compare 3 methods : subset selection, ridge, lasso

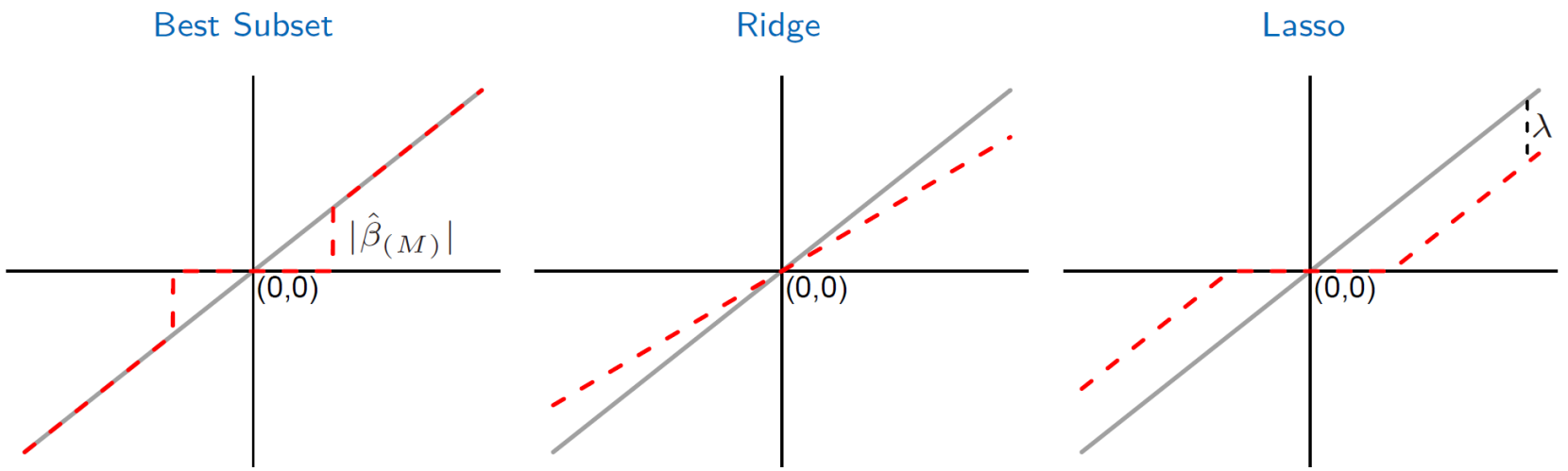
Above 3 methods should adaptively choose λ (or p, t) to minimize an estimate of expected prediction error.



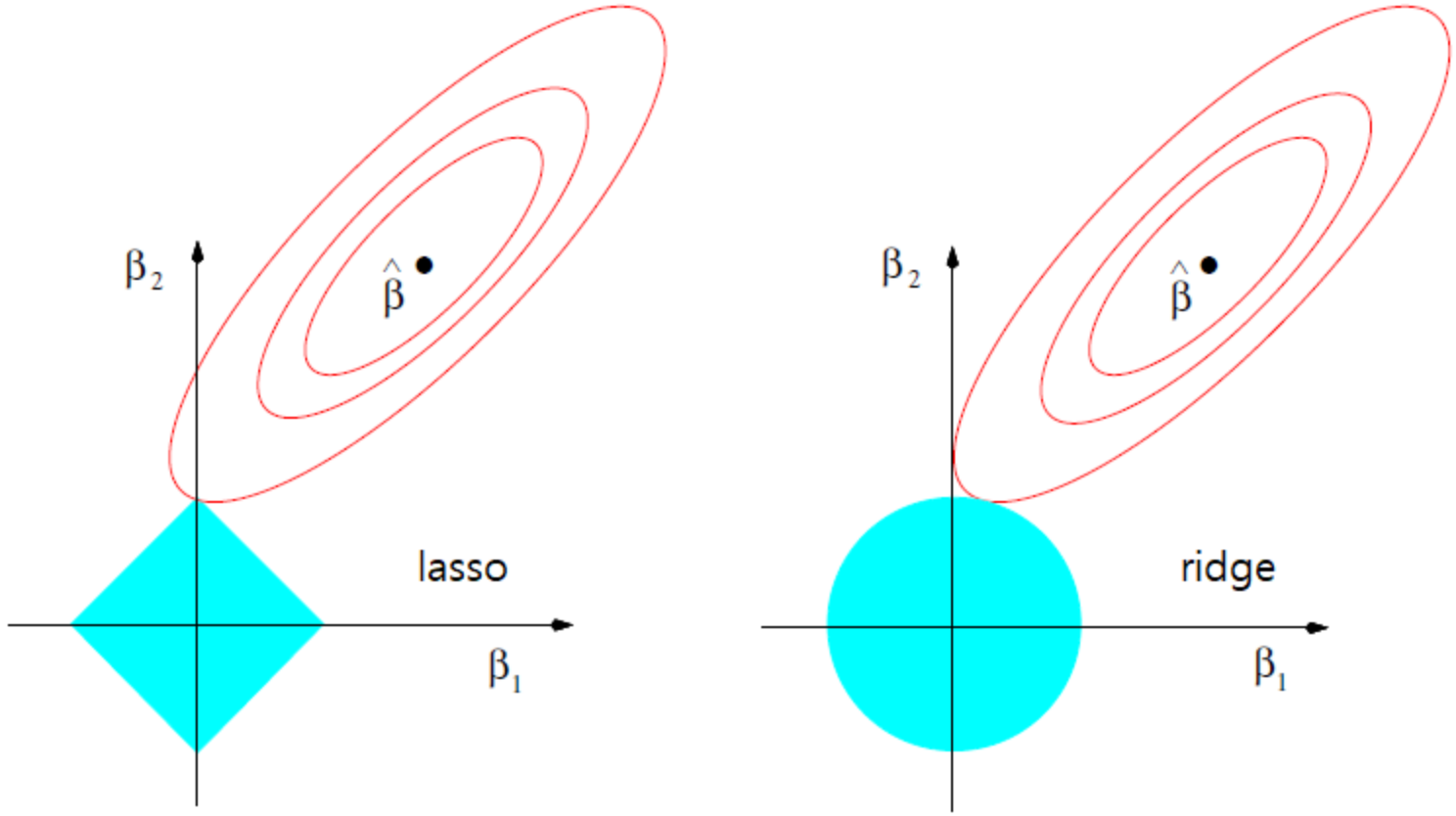
Case of an orthonormal input matrix X :

In an orthonomral case, there are explicit solutions.

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



Case of a non-orthogonal input matrix X :



The solid blue areas are the constraint regions, while the red ellipses are the contours of the least squares error function.

Generalize version of regularization regression

For given $q \geq 0$,

$$(\hat{\beta}_0^{q-\text{norm}}, \hat{\beta}^{q-\text{norm}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j|^q \leq t$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \quad \hat{\beta}^{\text{q-norm}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right\}$$

$$\hat{\beta}_0^{\text{q-norm}} = \bar{y} \ \ (= 0)$$

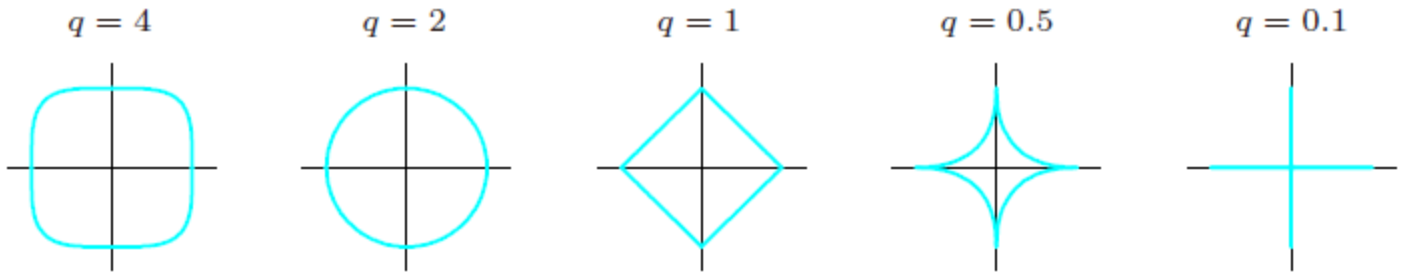
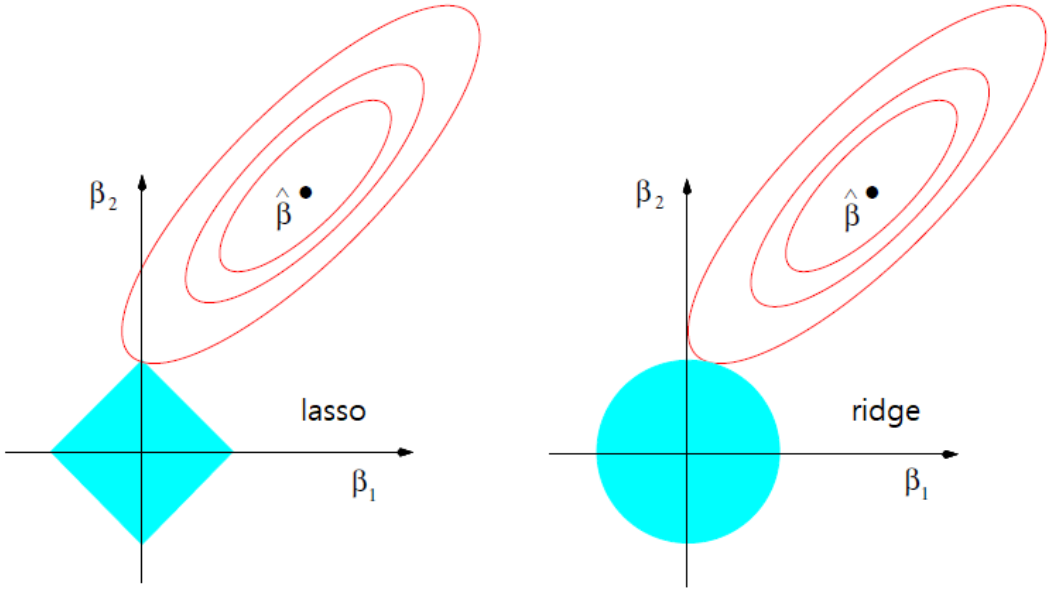


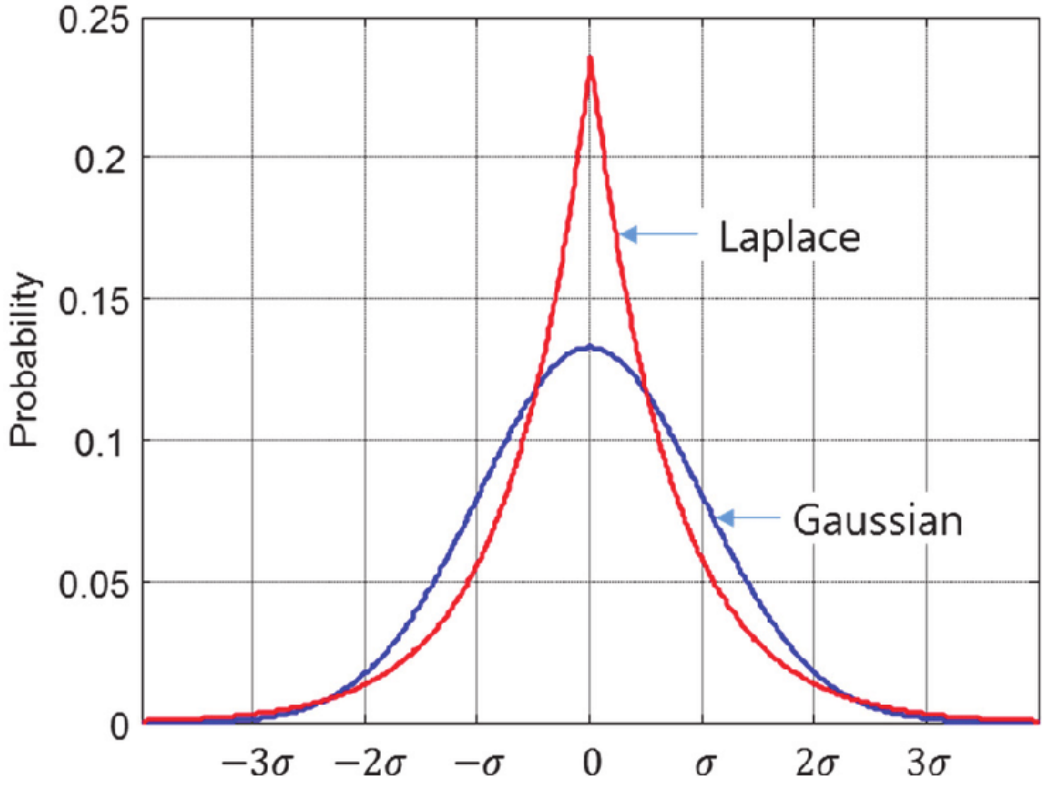
FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Bayesian approaches of Generalize version of ridge and lasso

- log-prior : $\log \beta^{\text{q-norm}} \sim \lambda \|\beta\|_q^q$
  - If $q = 0$ (best-subset selection), $\beta^{\text{q-norm}} \sim exp\{\lambda \|\beta\|_0\}$
  - If $q = 1$ (LASSO), $\beta^{\text{q-norm}} \sim exp\{\lambda \|\beta\|_1\}$ (Laplace)
    ($q = 1$ is the smallest q such that the constraint region is convex.)
  - If $q = 2$ (ridge), $\beta^{\text{q-norm}} \sim exp\left\{\lambda \|\beta\|_2^2\right\}$ (Gaussian)
- likelihood : $(X, y)|\beta \sim exp\left\{\|y - X\beta\|_2^2\right\}$ (Gaussian)
- log-posterior : $\log \beta^{\text{q-norm}}|(X, y) \sim \left\{\|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q\right\}$

$$\Longrightarrow \log \hat{\beta}^{\text{q-norm}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\log \beta^{\text{q-norm}}|(X, y)\}$$





- If $q = 1$ (LASSO), $\beta^{\text{q-norm}} \sim exp\left\{\lambda \|\beta\|_1\right\}$ (Laplace)
- If $q = 2$ (ridge), $\beta^{\text{q-norm}} \sim exp\left\{\lambda \|\beta\|_2^2\right\}$ (Gaussian)

$\Longrightarrow$ 3 methods are Bayes estimates with different priors.

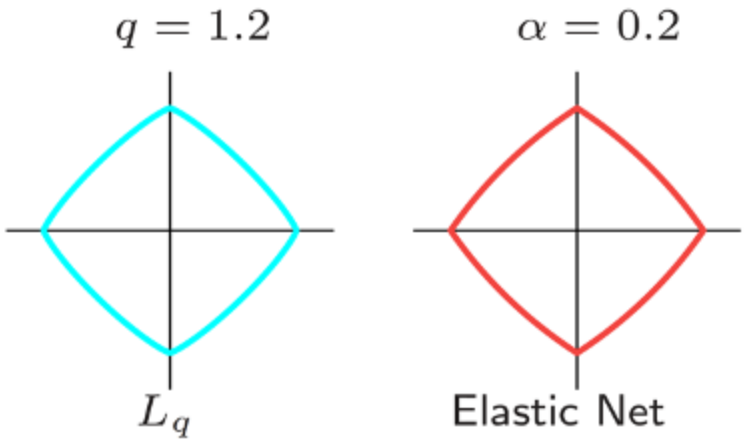
Elastic net

$$(\hat{\beta}_0^{\text{elastic}}, \hat{\beta}^{\text{elastic}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \text{subject to } \sum_{j=1}^{p} (\alpha \beta_j + (1-\alpha)|\beta_j|) \leq t$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \quad \hat{\beta}^{\text{elastic}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_2^2 + (1-\alpha)\|\beta\|_1 \right) \right\}$$

$$\hat{\beta}_0^{\text{elastic}} = \bar{y} \ \ (=0)$$

- Values of $q \in (1, 2)$ suggest a compromise between the lasso and ridge regression.
  - Although $\|\beta\|_q^q$ for $q \in (1, 2)$ is differentiable at 0, the regularization does not share the ability of lasso, setting coefficients exactly to zero.
- The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.



- Above contours are the constant value of $q = 1.2$ norm regularization (left), and the elastic-net for $\alpha = 0.2$ (right).
- The elastic-net has sharp (non-differentiable) corners, while the $q = 1.2$ penalty does not.
- It also has considerable computational advantages over the L^q penalties.

Homework

Exercise. 3.29

Suppose we fit a ridge regression with a given shrinkage parameter $\lambda \in \mathbb{R}^+$ on a single variable x_1. (Notice that x_1 is a $N \times 1$ vector.)

1. (Essential) Show that the coefficient must be $\frac{X^T y}{X^T X + \lambda}$ where $X = x_1$.
2. (Essential) We now include an exact copy $x_2 = x_1$, so our new design matrix would be $X = [x_1 | x_2]$. Using this matrix, re-fit our ridge regression. Show that both coefficients are identical, and derive their value.
3. (Extra) Show in general that if m copies of a variable x_j, are included in a ridge regression, so X would be $[x_1 | x_2 | \cdots | x_m]$, their coefficients are all the same.

References

- The Elements of Statistical Learning (2008) Trevor Hastie, Robert Tibshirani, Jerome Friedman
- Ridge Regression: Biased Estimation for Nonorthogonal Problems (1970) Arthur E. Hoerl, Robert W. Kennard