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

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

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

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,

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

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

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,

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

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

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

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

$$(X'X)^{-1} \; coumputable$$

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

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

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, $CN \to \infty$.
- 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

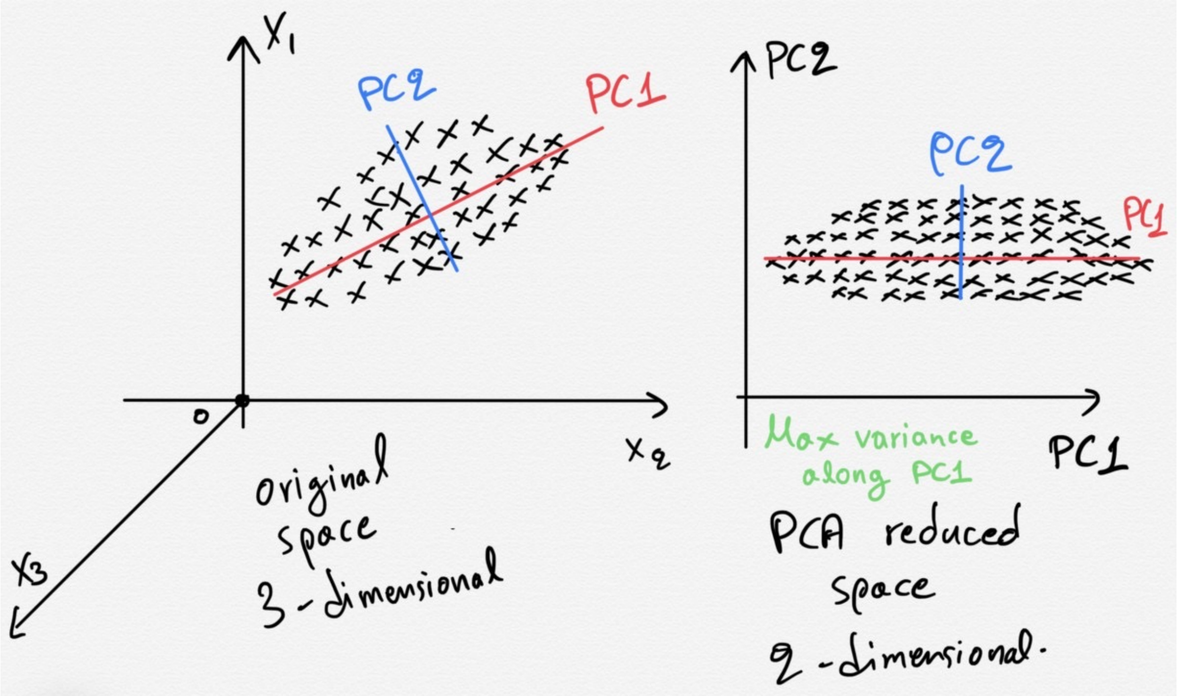
Remedy

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

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

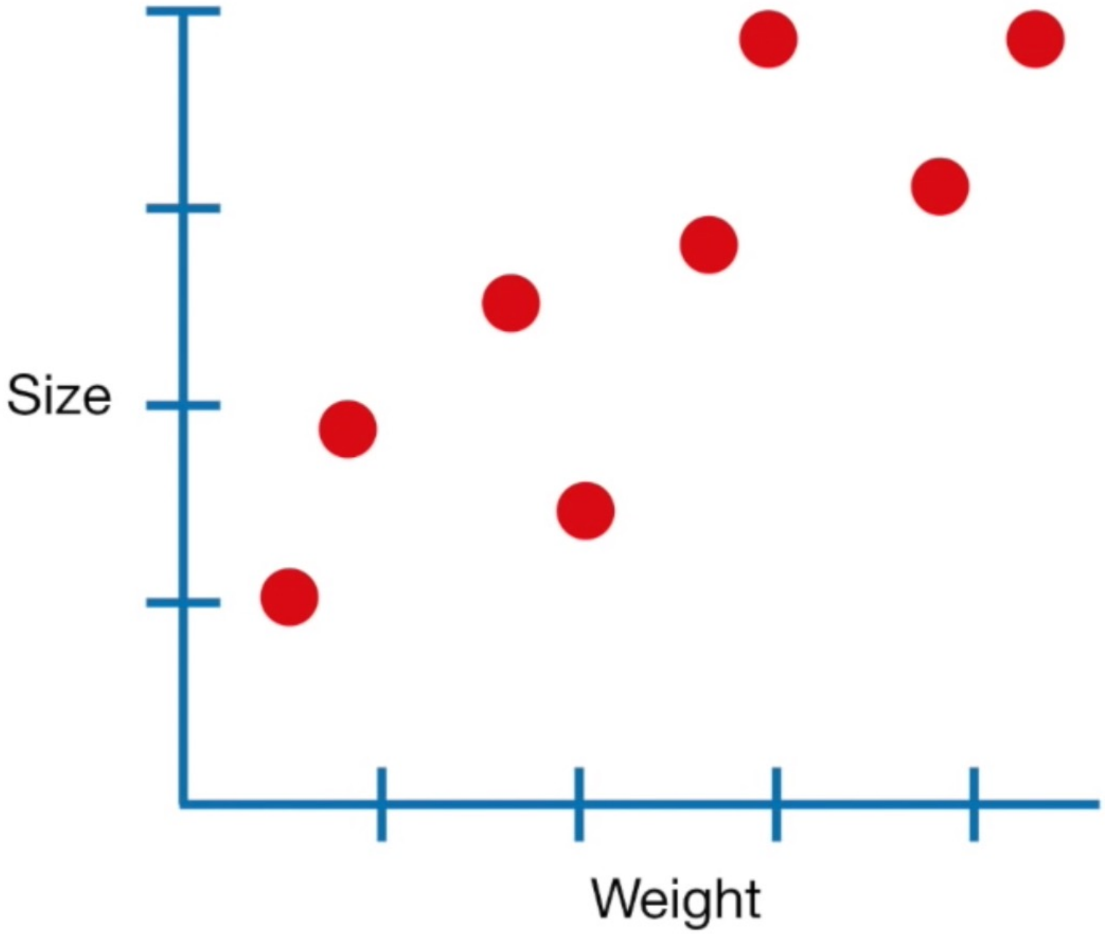
Prediction

Minimize test error

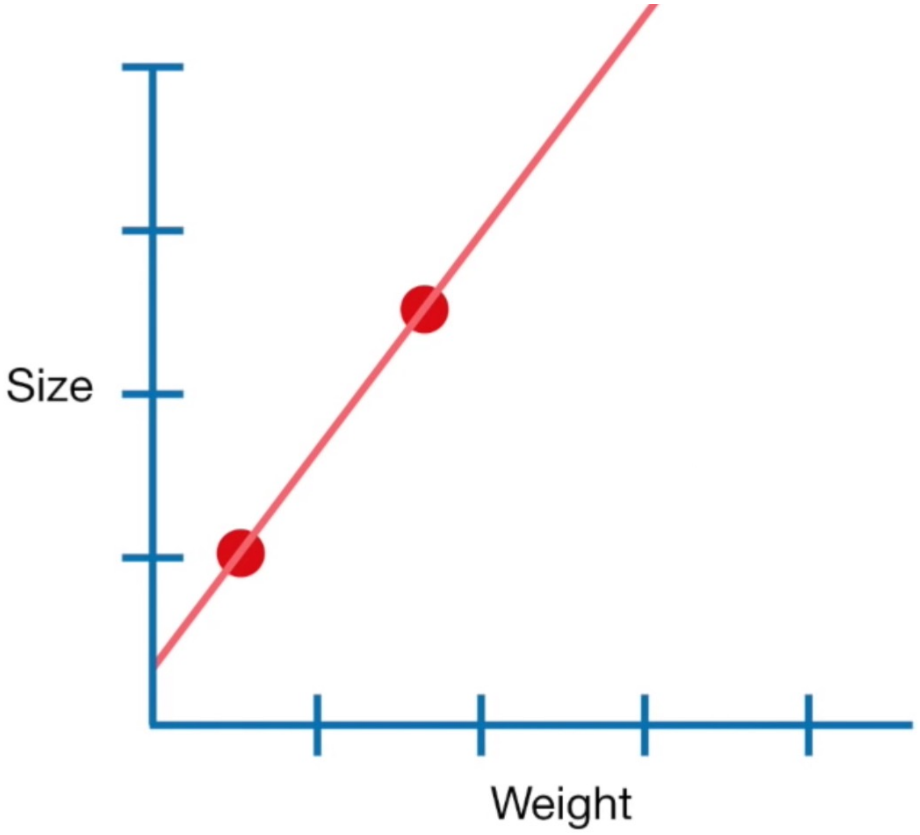
$$\begin{aligned}\text{Expected MSE} &= E\left[\left(Y - \hat{Y}\right)^2 | X\right] \\ &= \sigma^2 + (E[\hat{Y}] - \hat{Y})^2 + E[\hat{Y} - \mathrm{E}[\hat{Y}]]^2 \\ &= \sigma^2 + \mathrm{B}ias^2(\hat{Y}) + \mathrm{Var}(\hat{Y}) \\ &= \text{Irreducible Error} + \mathrm{B}ias^2 + \text{Variance}\end{aligned}$$

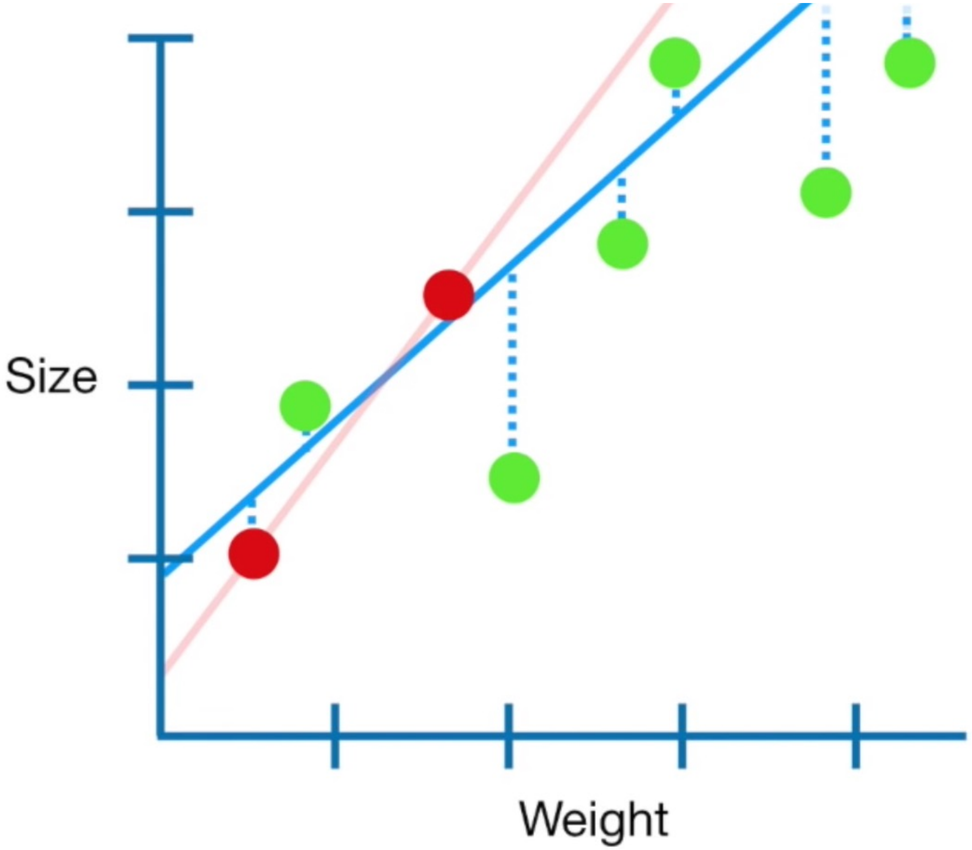
Review

What is a good model?



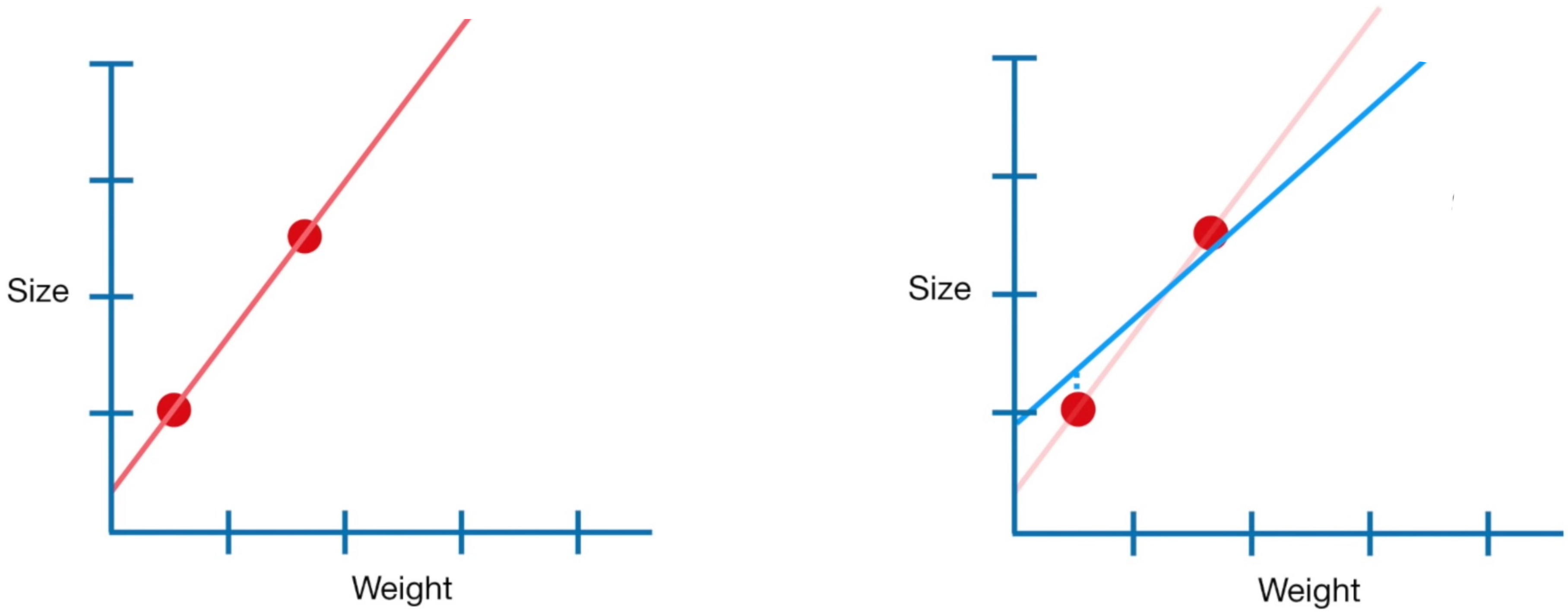
Regularization





Regularization

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



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

Ridge

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

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

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

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

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
```

$$
\begin{aligned}
MSE(\beta_1, \beta_2) &= \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \\
&= \sum_{i=1}^{n} {y_i}^2 - 2\sum_{i=1}^{n} y_i(\beta_1 x_{i1} + \beta_2 x_{i2}) + \sum_{i=1}^{n} (\beta_1 x_{i1} + \beta_2 x_{i2})^2 \\
&= \sum_{i=1}^{n} {y_i}^2 - 2\left(\sum_{i=1}^{n} y_i x_{i1}\right)\beta_1 - 2\left(\sum_{i=1}^{n} y_i x_{i2}\right)\beta_2 + \sum_{i=1}^{n} (\beta_1^2 x_{i1}^2 + \beta_2^2 x_{i2}^2 + 2\beta_1\beta_2 x_{i1} x_{i2}) \\
&= \left(\sum_{i=1}^{n} x_{i1}^2\right)\beta_1^2 + \left(\sum_{i=1}^{n} x_{i2}^2\right)\beta_2^2 + \left(2\sum_{i=1}^{n} x_{i1} x_{i2}\right)\beta_1\beta_2 \\
&\qquad -2\left(\sum_{i=1}^{n} y_i x_{i1}\right)\beta_1 - 2\left(\sum_{i=1}^{n} y_i x_{i2}\right)\beta_2 + \sum_{i=1}^{n} {y_i}^2 \\
&= A\beta_1^2 + B\beta_1\beta_2 + C\beta_2^2 + D\beta_1 + E\beta_2 + F
\end{aligned}
$$

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])

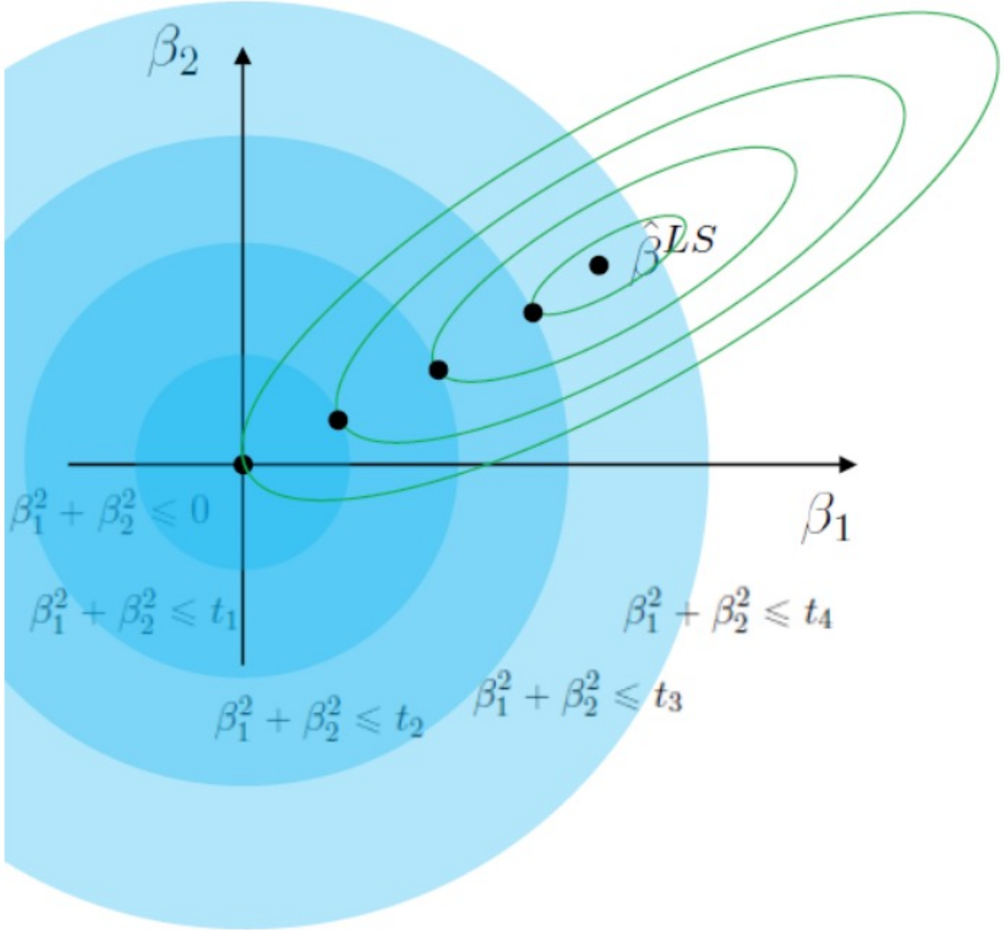
where $\langle \cdot, \cdot \rangle$ is the inner product. Examples of inner products include the real and complex dot product; see the examples in inner product. Every inner product gives rise to a norm, called the *canonical* or *induced norm*, where the norm of a vector $\mathbf{u}$ is denoted and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

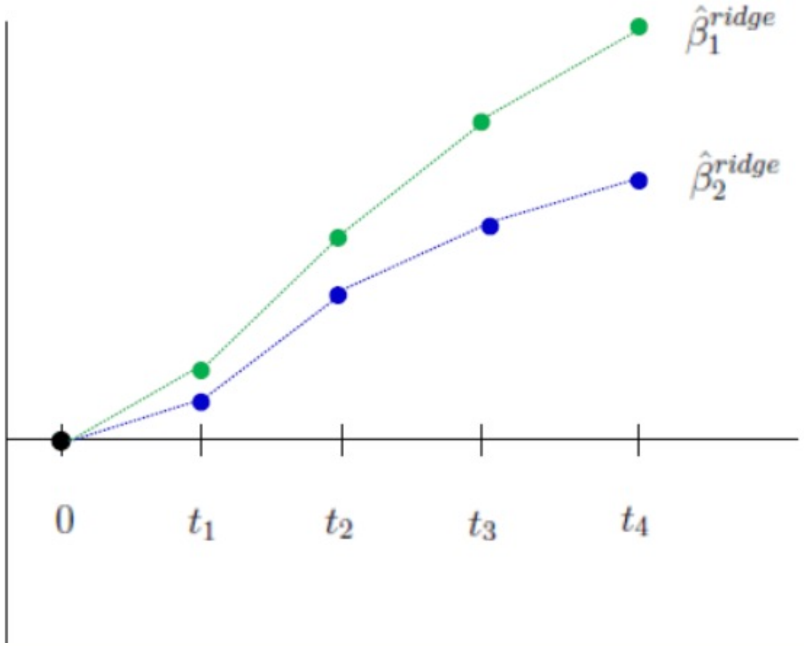
$$\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$$

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]



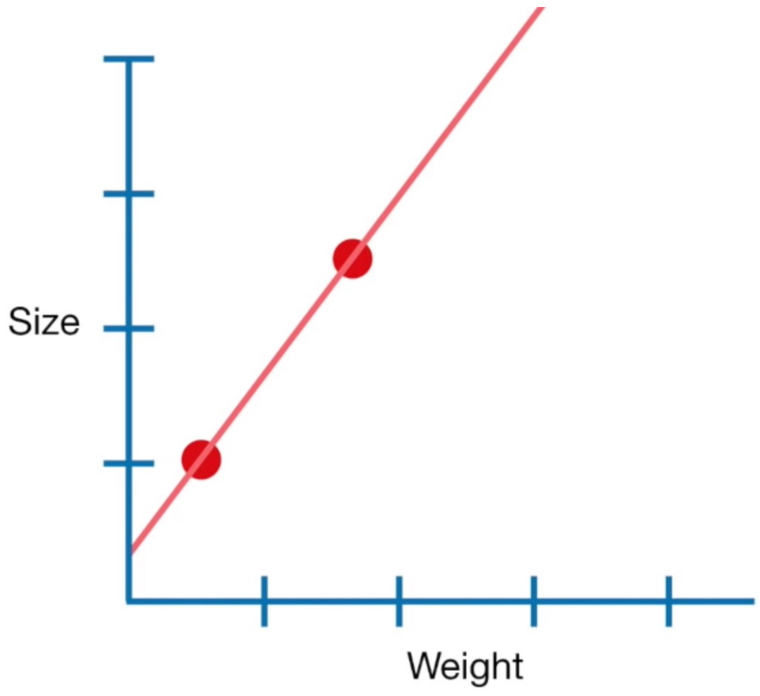


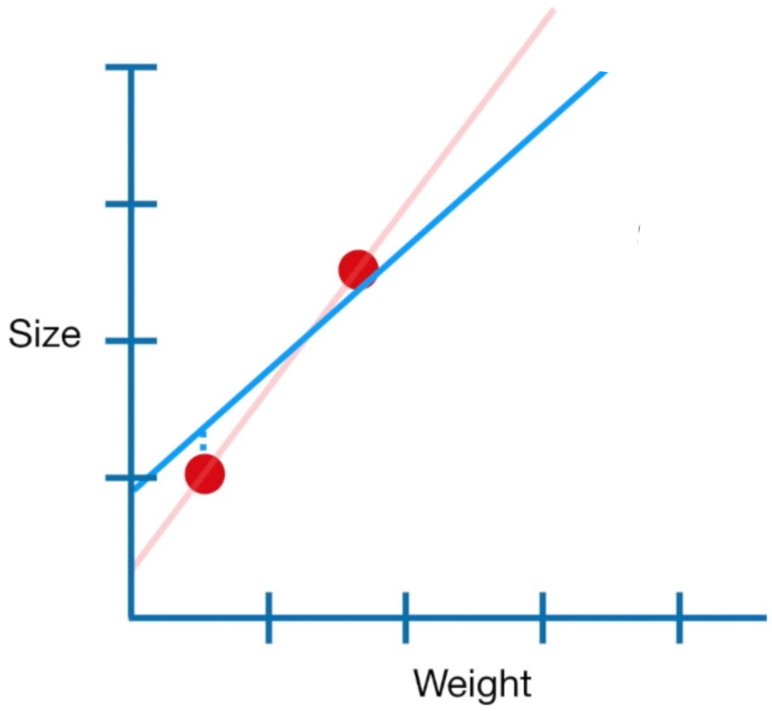
$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

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

Regularization

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





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

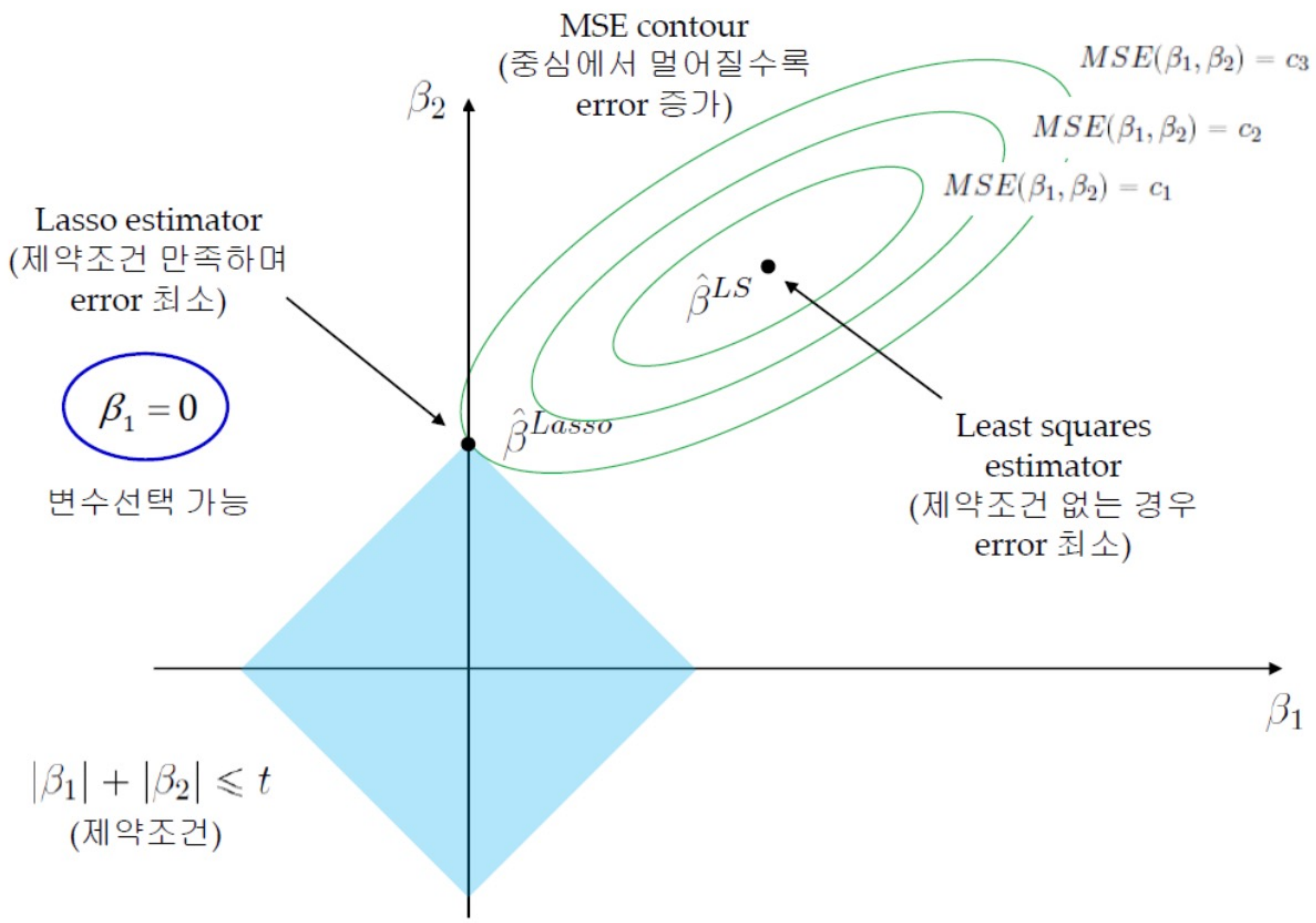
Lasso

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

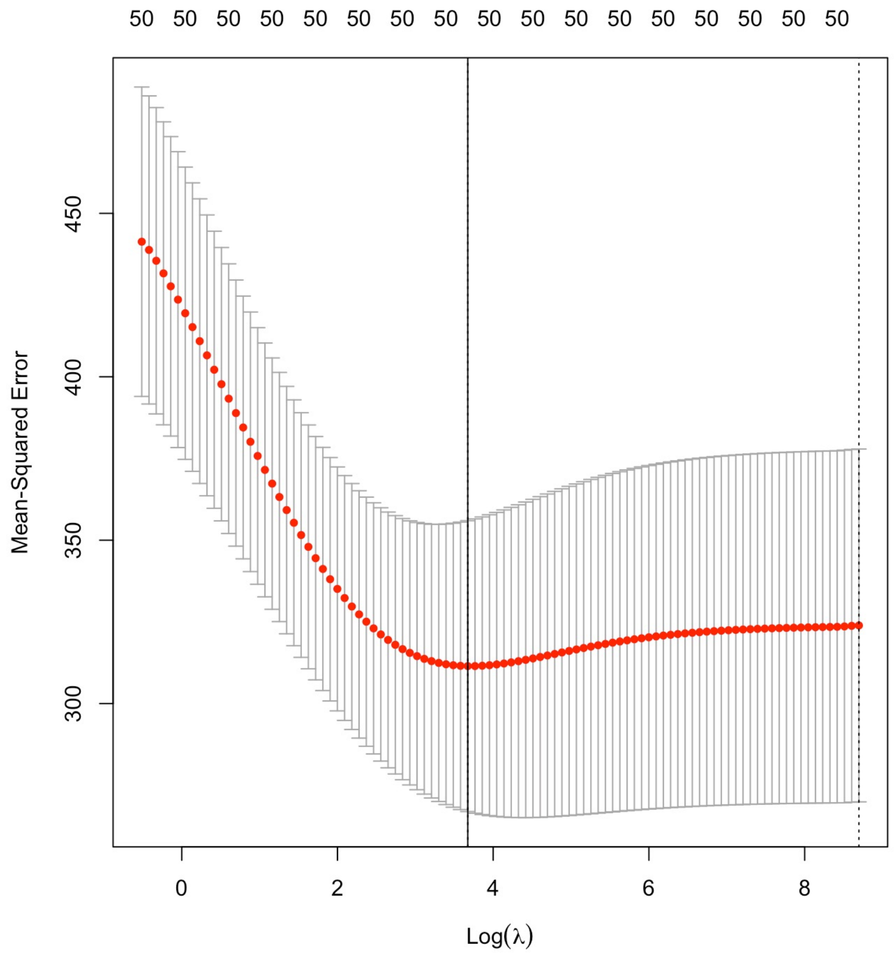
$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

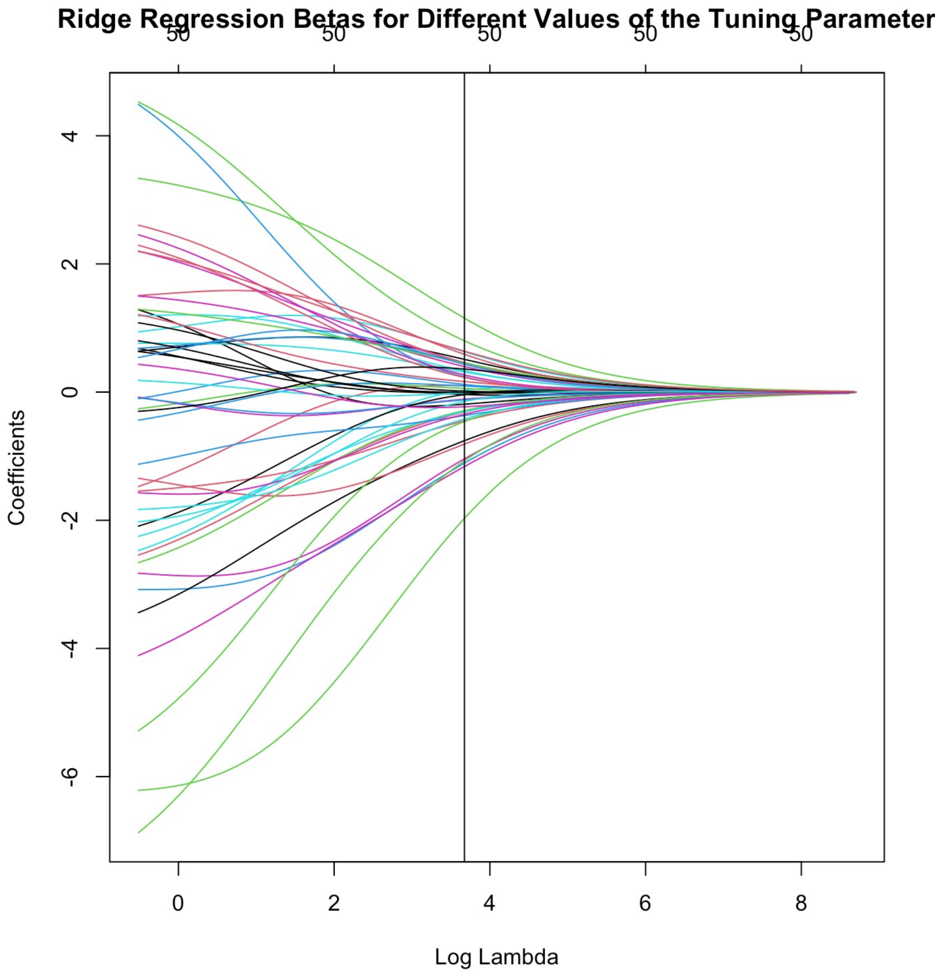
$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

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

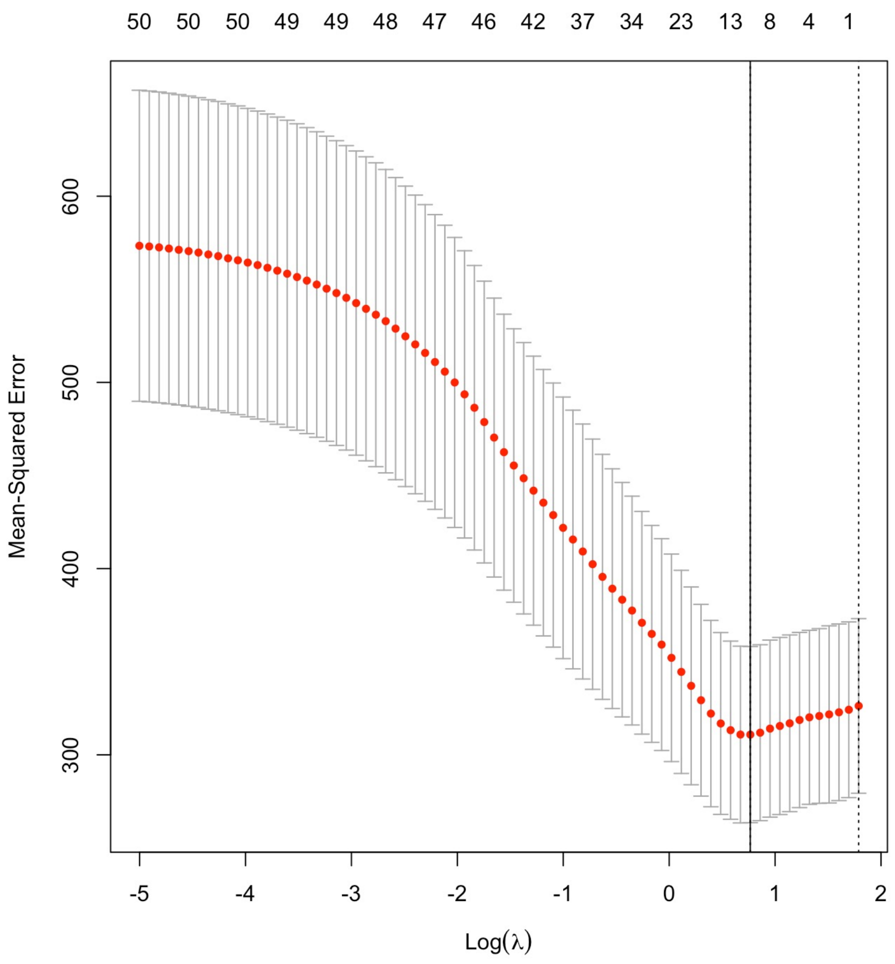

MSE contour
(중심에서 멀어질수록
error 증가)
β_2
$MSE(\beta_1, \beta_2) = c_3$
$MSE(\beta_1, \beta_2) = c_2$
$MSE(\beta_1, \beta_2) = c_1$
Lasso estimator
(제약조건 만족하며
error 최소)
$\hat{\beta}^{LS}$
$\beta_1 = 0$
$\hat{\beta}^{Lasso}$
Least squares
estimator
(제약조건 없는 경우
error 최소)
변수선택 가능
β_1
$|\beta_1| + |\beta_2|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

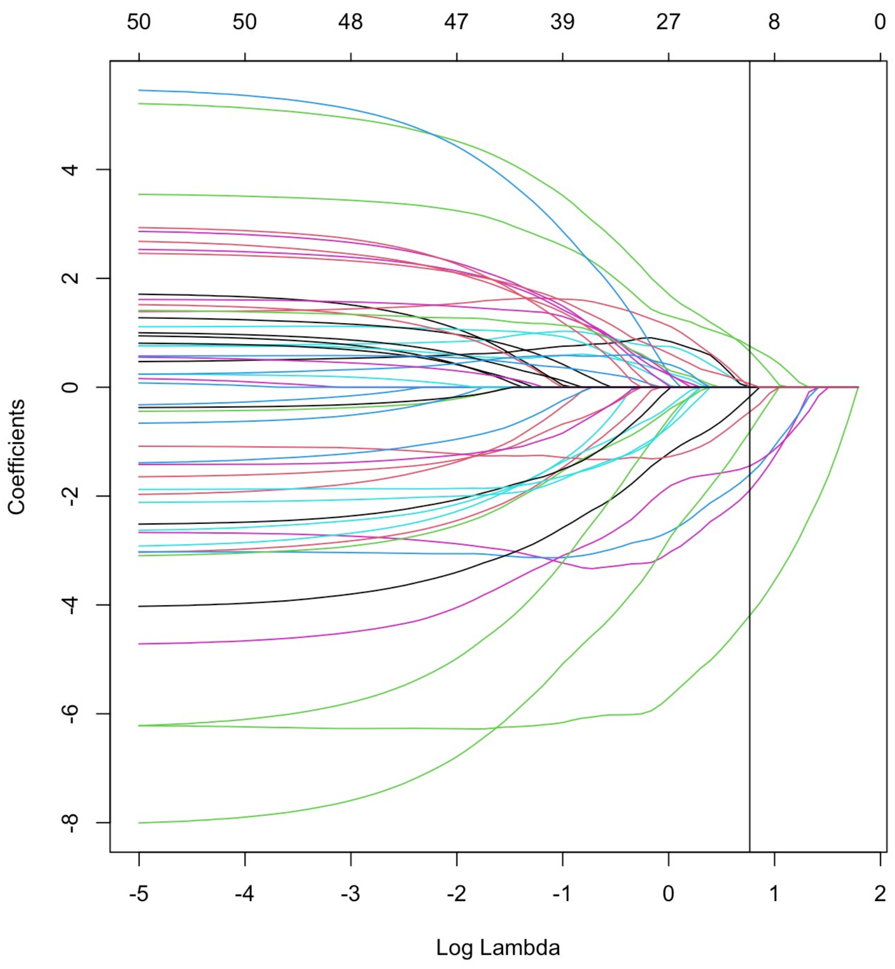
Cross-Validation





Cross-Validation





HW ??