DACON 스터디
3주차

Regression

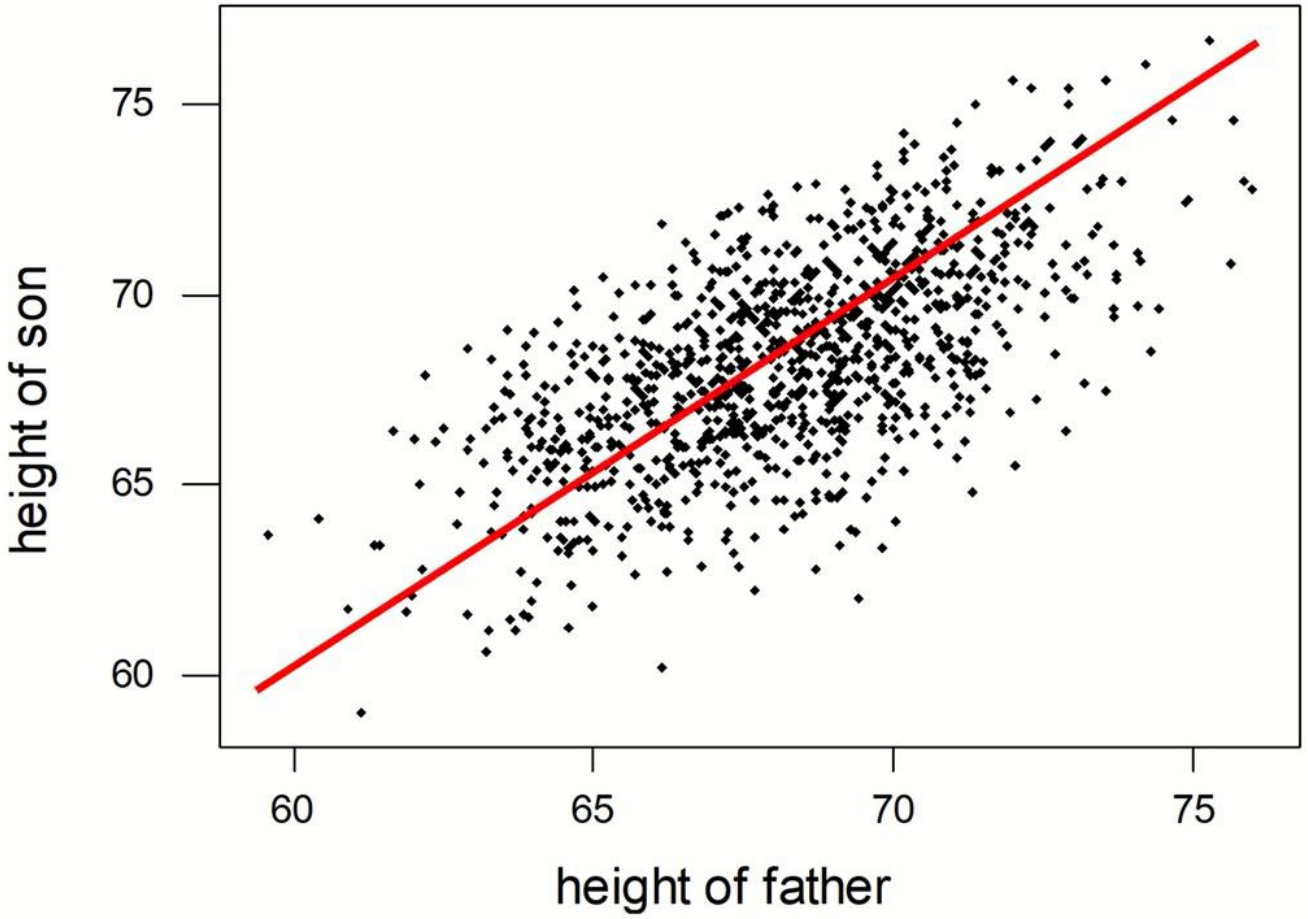
이제윤

20220802

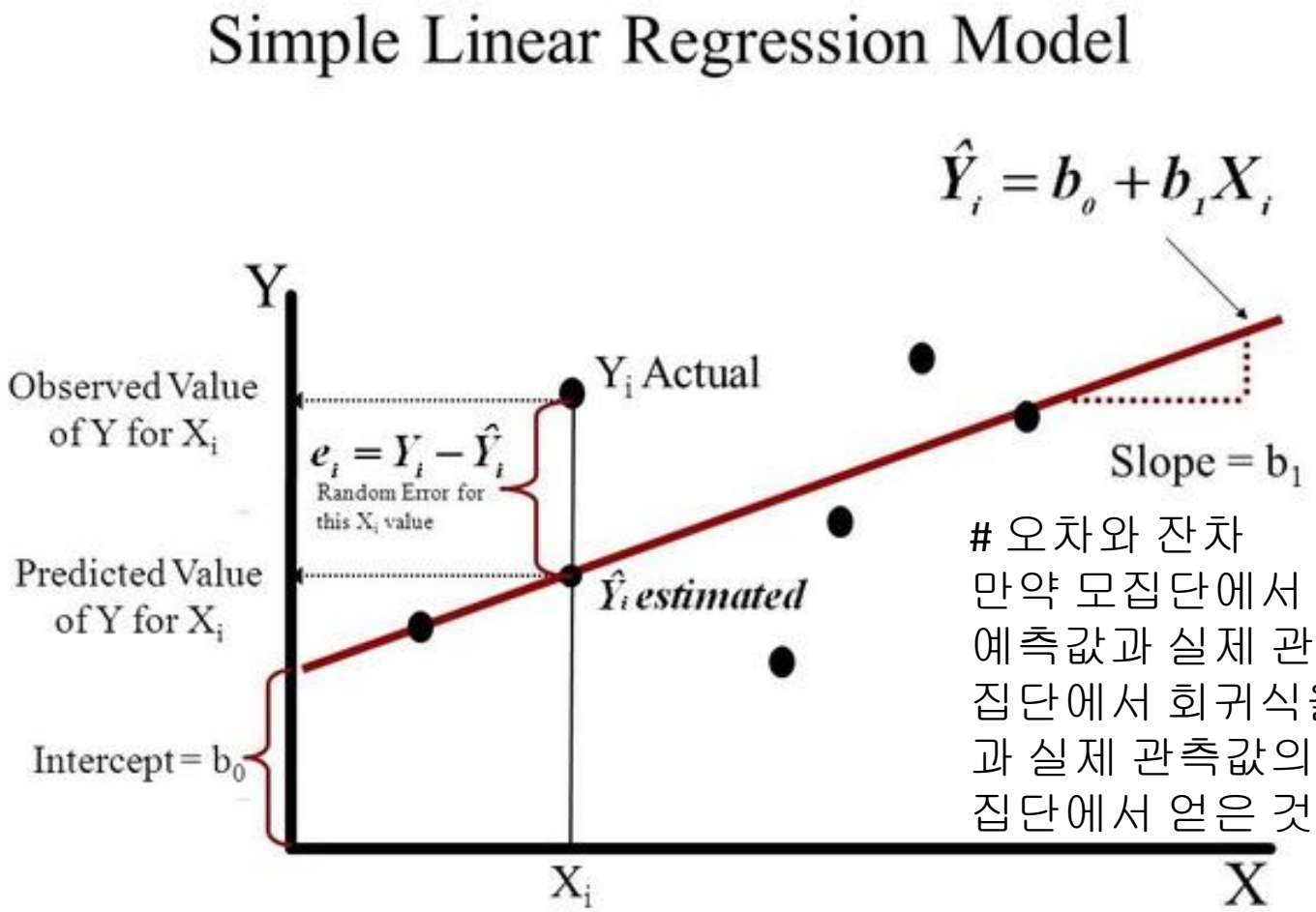
CONTENTS

Linear Regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



Linear Regression

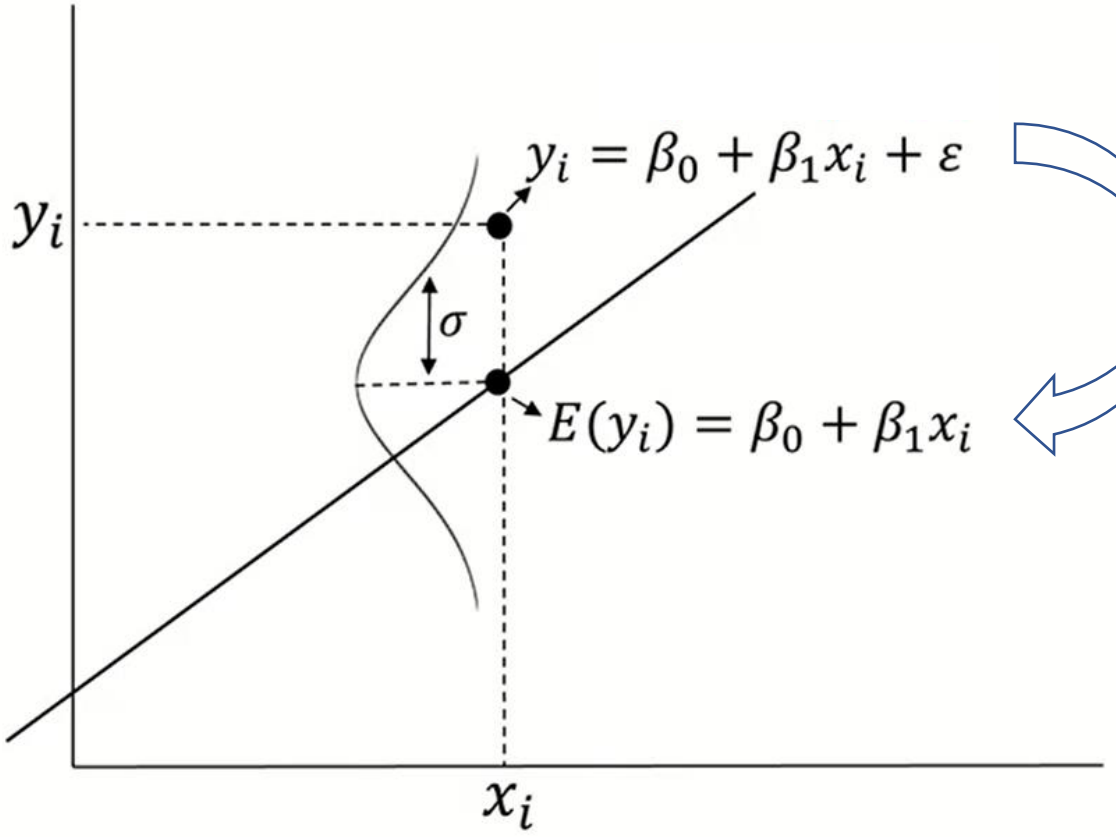


오차와 잔차

만약 모집단에서 회귀식을 얻었다면, 그 회귀식을 통해 얻은 예측값과 실제 관측값의 차이가 **오차(error)**이다. 반면 표본집단에서 회귀식을 얻었다면, 그 회귀식을 통해 얻은 예측값과 실제 관측값의 차이가 **잔차(residual)**이다. 둘의 차이는 모집단에서 얻은 것이냐 표본집단에서 얻은 것이냐 뿐이다.

Linear Regression

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2),\ i = 1, 2, \ldots, n$$



평균으로
회귀

Linear Regression

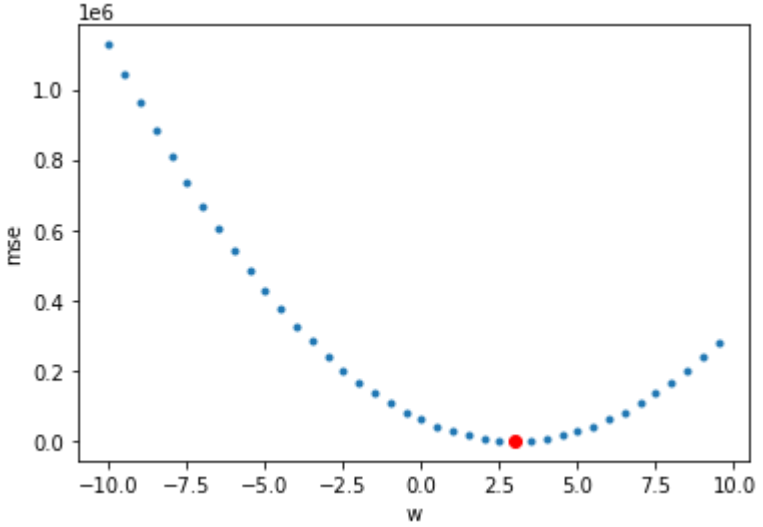
Cost function (비용함수)

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) = 0$$

$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i)X_i = 0$$

미분이 가능하므로 gradient descent 가능



Linear Regression

<오차에 대한 가정>

보통은 표본을 가지고 회귀식을 추정하기 때문에, 모집단으로부터 추정한 회귀식으로부터 얻은 예측값과 실제값의 차이인, 오차는 관측할수가 없겠죠. 따라서, 회귀분석에서는 관측할수 없는 오차에 대한 몇가지 가정을 전제로 회귀식의 모수들을 추정합니다.

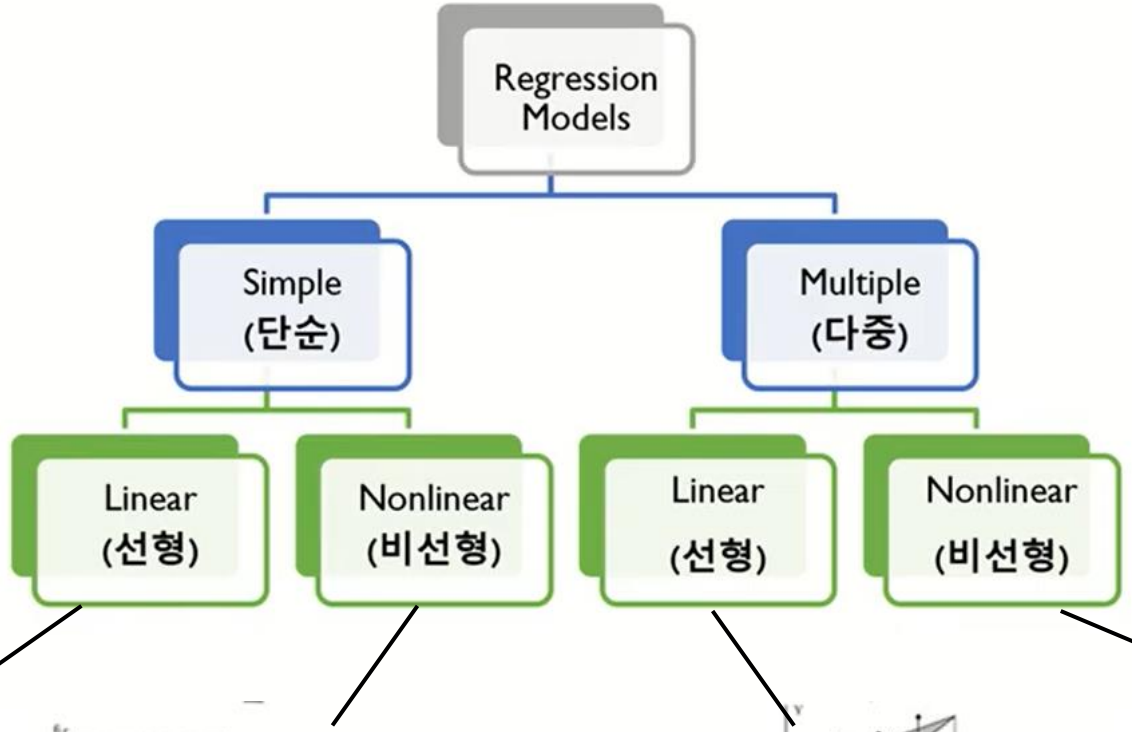
1. E(εi) = 0

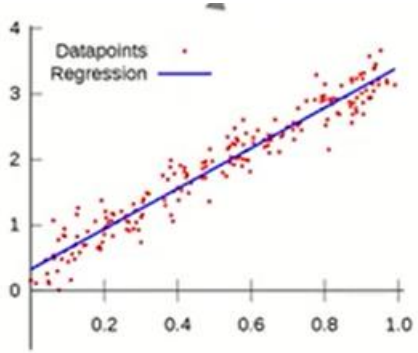
2. Var(εi) = σ^2 * I "등분산성"

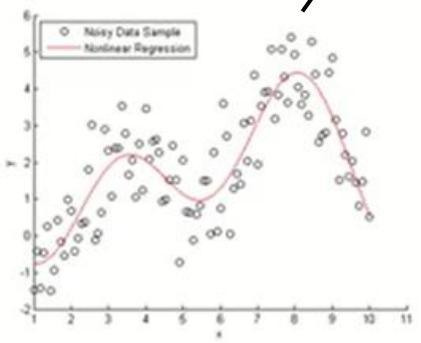
3. Cov(εi, εj) = 0 "독립성"

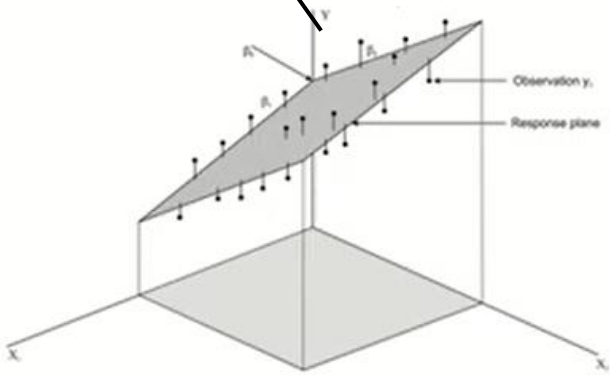
첫번째 가정은, 고정오차가 없다는 가정입니다. 즉, '모집단으로부터 추정한 회귀식은 모집단의 관측값을 설명하기에 적합하다' 라는 전제를 가지고 모수를 추정하는 거죠. 고정오차가 없으면, 오차의 평균은 0입니다.
두번째 가정은, '모든 오차는 동일한 분산을 가진다.' 입니다. 이 가정은,
세번째 가정은, 오차들이 서로에게 영향을 주지 않는다. 즉, εi, εj 는 서로에게 상관관계를 가지지 않는다는 가정입니다.

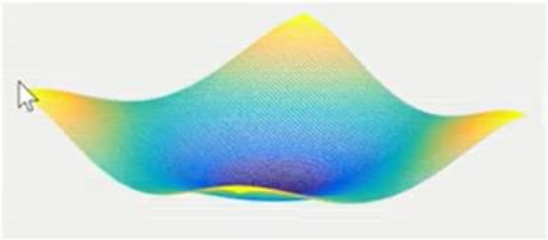
Linear Regression











Linear Regression

- Linearity?

엄밀한 선형회귀모델

일반적인 선형회귀모델

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \epsilon_i$$

함수 $f(x)$가 다음을 만족하면 선형$(linear)$이라 한다.

$$f(x_1+x_2) = f(x_1) + f(x_2)$$
$$f(kx) = kf(x)$$

or

벡터공간 V에 속하는 벡터 $v_1, \cdots, v_n$ 와 어떤 스칼라 $a_1, \cdots, a_n$ 에 대한 선형결합$(linear\ combination)$은 다음의 꼴로 나타낸다.

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$$

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \Rightarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- $y = \beta_0 x^{\beta_1} \Rightarrow log(y) = log(\beta_0 x^{\beta_1}) \Rightarrow log\beta_0 + \beta_1 \log(x) \Rightarrow y^* = \beta_0^* + \beta_1 x^*$
- $y = \frac{e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}} \Rightarrow \frac{y}{1-y} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \Rightarrow \log\left(\frac{y}{1-y}\right) = y^* = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

Linear Regression

선형 회귀 모델은 파라미터 계수에 대한 해석이 단순하지만 비선형 모델은 모델의 형태가 복잡할 경우 해석이 매우 어렵습니다. 그래서 보통 모델의 해석을 중시하는 통계 모델링에서는 비선형 회귀 모델을 잘 사용하지 않습니다.

그런데 만약 회귀 모델의 목적이 해석이 아니라 예측에 있다면 비선형 모델은 대단히 유연하기 때문에 복잡한 패턴을 갖는 데이터에 대해서도 모델링이 가능합니다. 그래서 충분히 많은 데이터를 갖고 있어서 variance error를 충분히 줄일 수 있고 예측 자체가 목적인 경우라면 비선형 모델은 사용할만한 도구입니다. 기계 학습 분야에서는 실제 이런 비선형 모델을 대단히 많이 사용하고 있는데 가장 대표적인 것이 소위 딥 러닝이라고 부르는 뉴럴 네트워크입니다.

3. 결론

정리하자면, 선형 회귀 모델은 파라미터가 선형식으로 표현되는 회귀 모델을 의미합니다. 그리고 이런 선형 회귀 모델은 파라미터를 추정하거나 모델을 해석하기가 비선형 모델에 비해 비교적 쉽기 때문에, 데이터를 적절히 변환하거나 도움이 되는 feature들을 추가하여 선형 모델을 만들 수 있다면 이렇게 하는 것이 적은 개수의 feature로 복잡한 비선형 모델을 만드는 것보다 여러 면에서 유리합니다.

반면 선형 모델은 표현 가능한 모델의 가짓수(파라미터의 개수가 아니라 파라미터의 결합 형태)가 한정되어 있기 때문에 유연성이 떨어집니다. 따라서 복잡한 패턴을 갖고 있는 데이터에 대해서는 정확한 모델링이 불가능한 경우가 있습니다. 그래서 최근에는 모델의 해석보다는 정교한 예측이 중요한 분야의 경우 뉴럴 네트워크와 같은 비선형 모델이 널리 사용되고 있습니다.

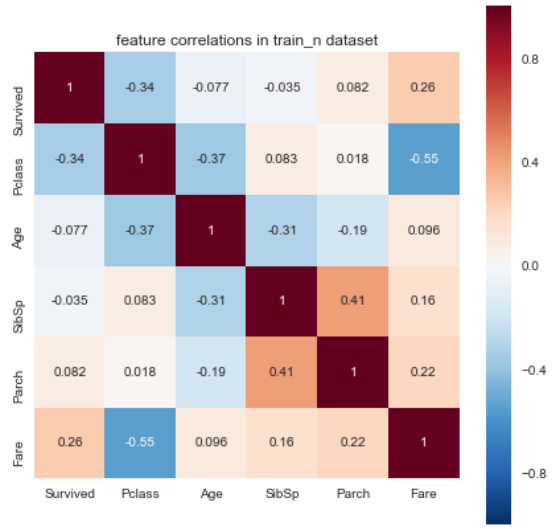
다중공선성

선형회귀의 엄격한 통계적 가정을 만족하기란 실제 데이터에서는 거의 불가능...

다중공선성 문제가 발생했는지 어떻게 알 수 있는가?

1) correlation matrix를 그려보자

수많은 독립변수 중 어떤 것을 선택해야될 지 모르겠을 때는, 종속변수도 correlation matrix에 포함시켜서, 종속변수와 가장 높은 상관관계를 가지는 독립변수를 선택하라.



2) VIF를 확인하자

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

각각의 독립변수에 대한 VIF를 구해보자. VIF값이 클 수록, 해당 독립변수와 나머지 독립변수간의 상관관계가 높아진다.

- Feature Selection

  - Subset selection, Stepwise method, LASSO, Least Angle Regression etc..

- Feature Extraction (Dimension Reduction)

  - Principal Component Analysis, Partial Least Square, Discriminant Analysis, Factor Analysis, Latent Class Analysis, etc..

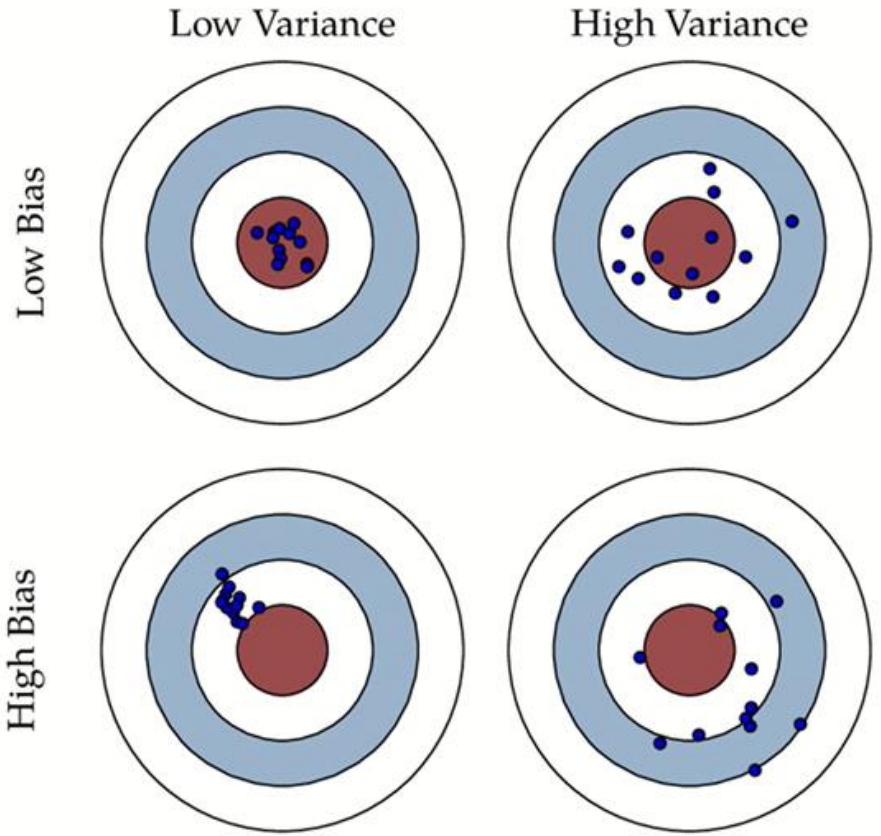
Regularization Model

$$\text{Expected MSE} = \text{Irreducible Error} + \text{B}ias^2 + \text{Variance}$$

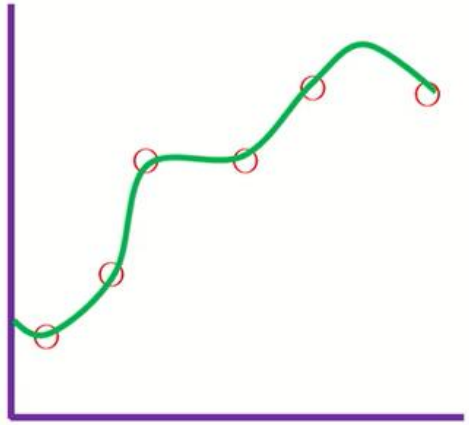
MSE로 부터 구한 베타는

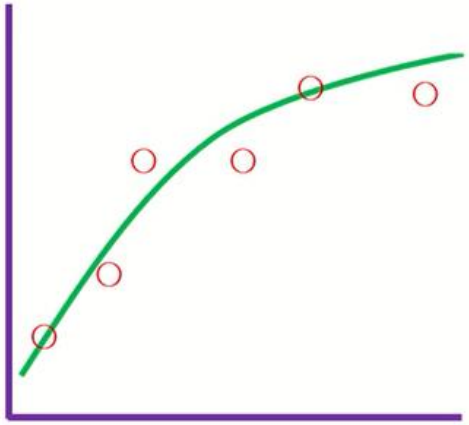
Unbiased Estimators중에서
가장 Variance가 작은 베타

…그러면 B를 키우고 V를
낮출 수는 없나~?



Regularization Model





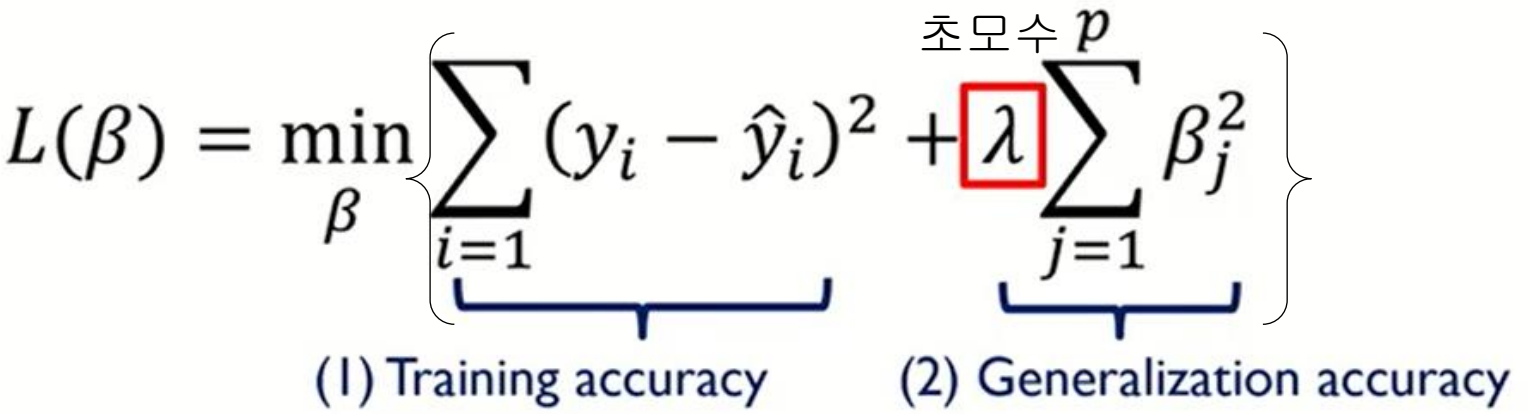
$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

$\beta_3 \approx 0$ $\beta_4 \approx 0$

$$\min_{\beta} \left\{ \sum_{i=1} (y_i - \hat{y}_i)^2 + 5000{\beta_3}^2 + 5000{\beta_4}^2 \right\}$$

Regularization Model



λ : regularization parameter that controls the tradeoff between (1) and (2)

Regularization Model

$\beta_1^2+\beta_2^2 \leq 30$

(β_1, β_2)	$\beta_1^2+\beta_2^2$	MSE
(4,5)	41	20
(3,5)	34	23
(4,4)	32	25
(2,5)	27	27
(2,4)	18	25
(2,3)	13	29

MSE는 낮은데 overfitting!

MSE를 희생하고 overfitting해결

1. Ridge Regression

L_2-norm regularization:
제곱 오차를 최소화하면서 회귀 계수 β의 L_2-norm을 제한

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

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

⇕ Equivalent (Lagrangian multiplier)

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

1. Ridge Regression

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

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

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

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

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

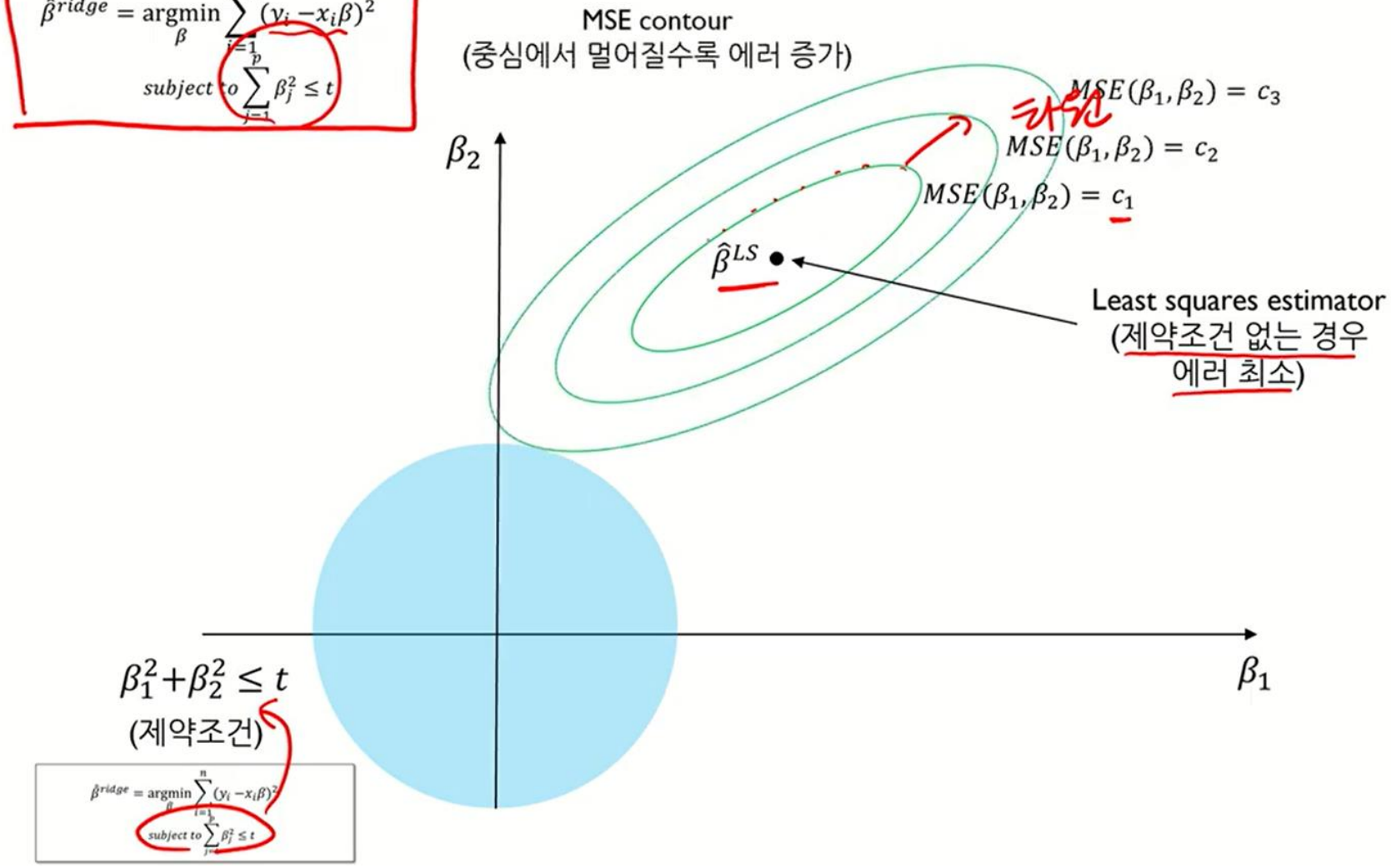
$$= 4\left\{\left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2 - \sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2\right\} < 0$$

By Cauchy-Schwartz inequality

1. Ridge Regression

[핵심 머신러닝] 정규화모델 1(Regularization 개념, Ridge Regression) - YouTube

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$
$$subject\ to \sum_{j=1}^{p} \beta_j^2 \le t$$



2. Lasso Regression

Least **A**bsolute **S**hrinkage and **S**election **O**perator

변수 선택 가능

L_1-norm regularization: 회귀 계수 β의 L_1-norm을 제한

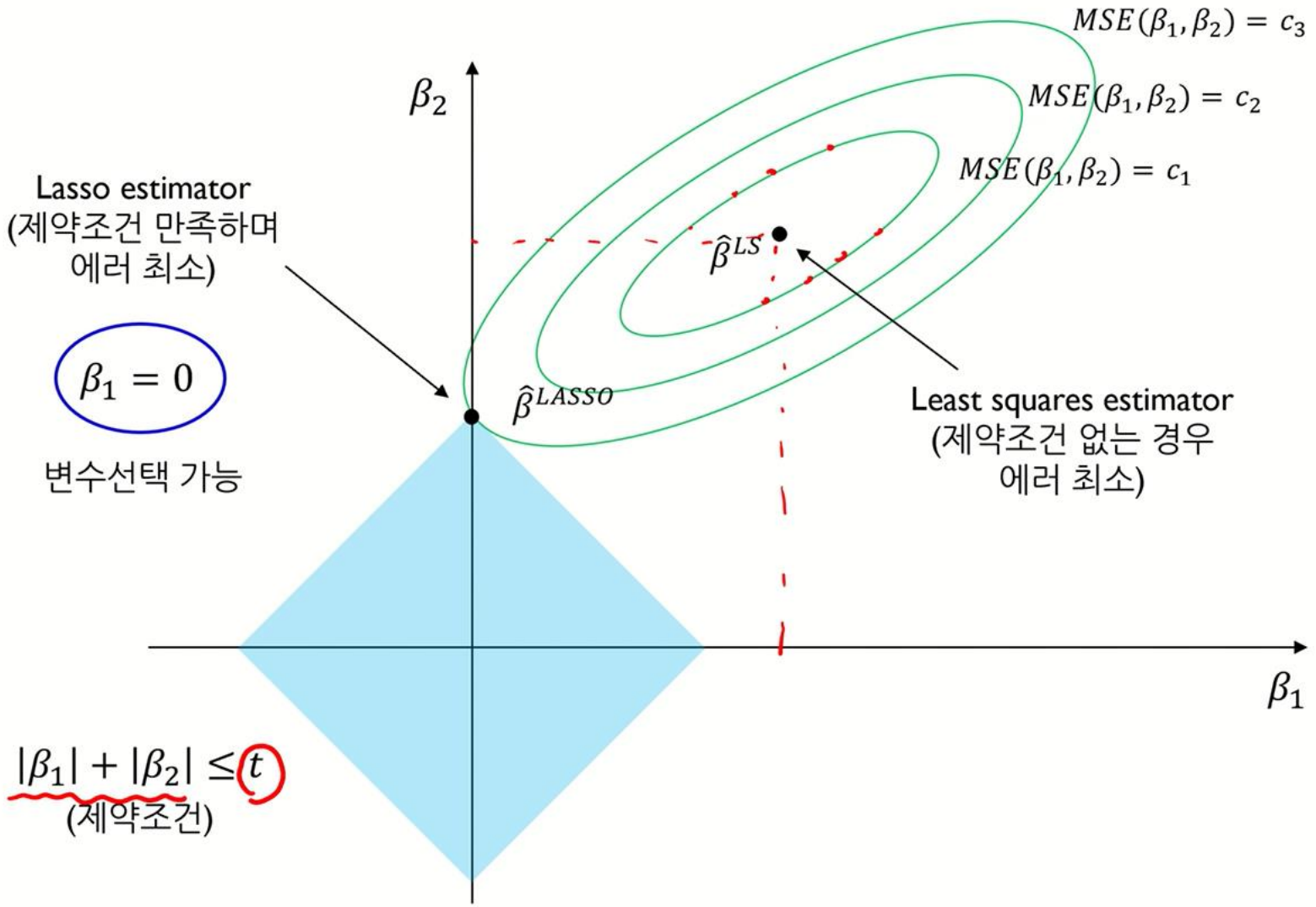
$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

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

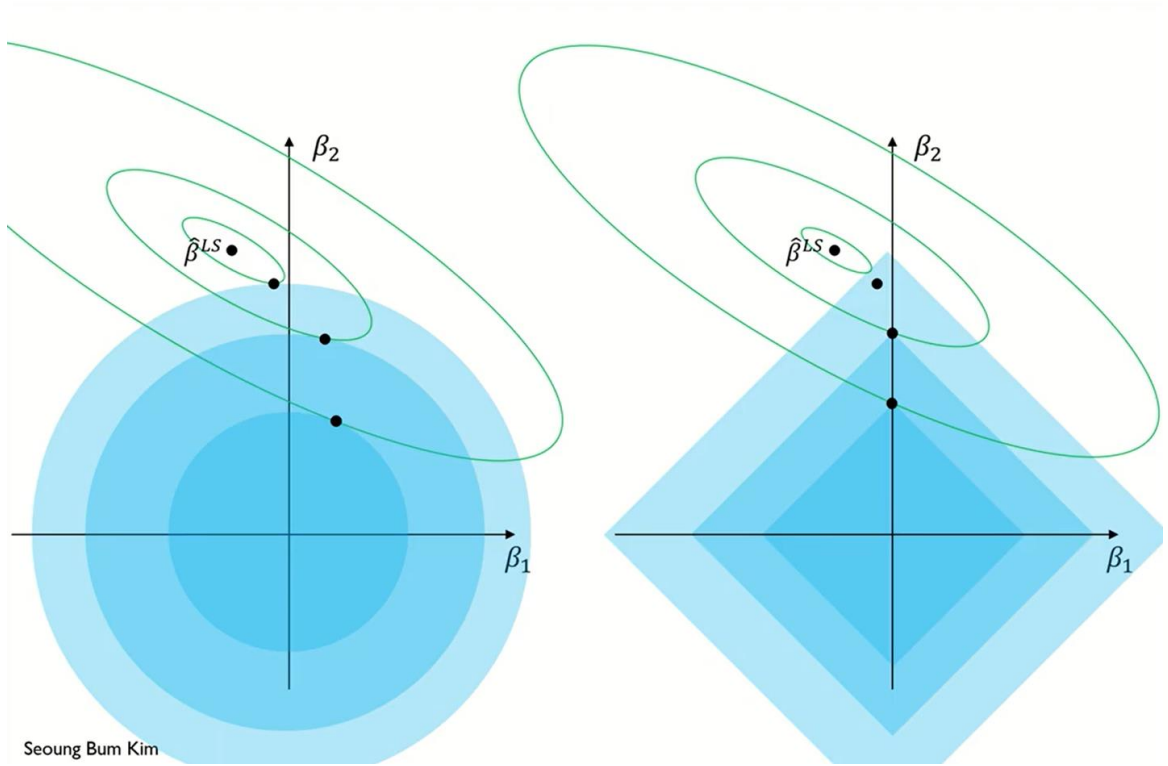
⇕ Equivalent

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

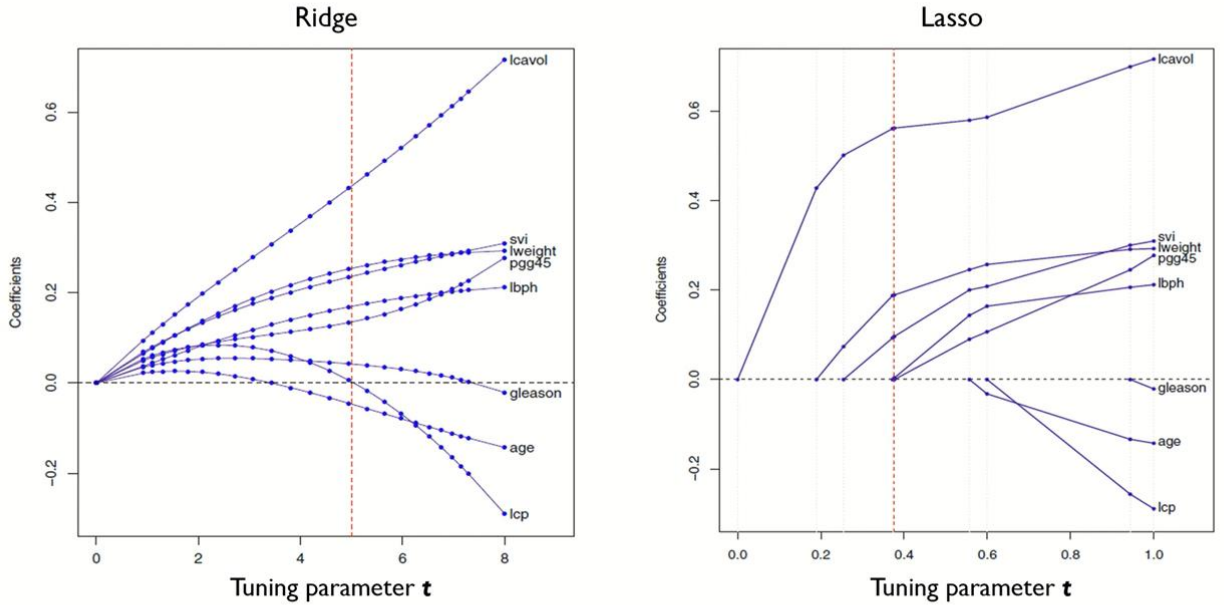
2. Lasso Regression



Ridge & Lasso



Prostate cancer data (Y: 전립선 암 항체, X: 환자 의료 데이터)



- Ridge와 Lasso 모두 t가 작아짐에 따라 모든 계수의 크기가 감소
- Lasso: 예측에 중요하지 않은 변수가 더 빠르게 감소, t가 작아짐에 따라 예측에 중요하지 않은 변수가 0이 됨

3. Elastic Net Regression

- Elastic net = Ridge + Lasso (L_1- and L_2-regularization)
- Elastic net은 상관관계 큰 변수를 동시에 선택/배제하는 특성

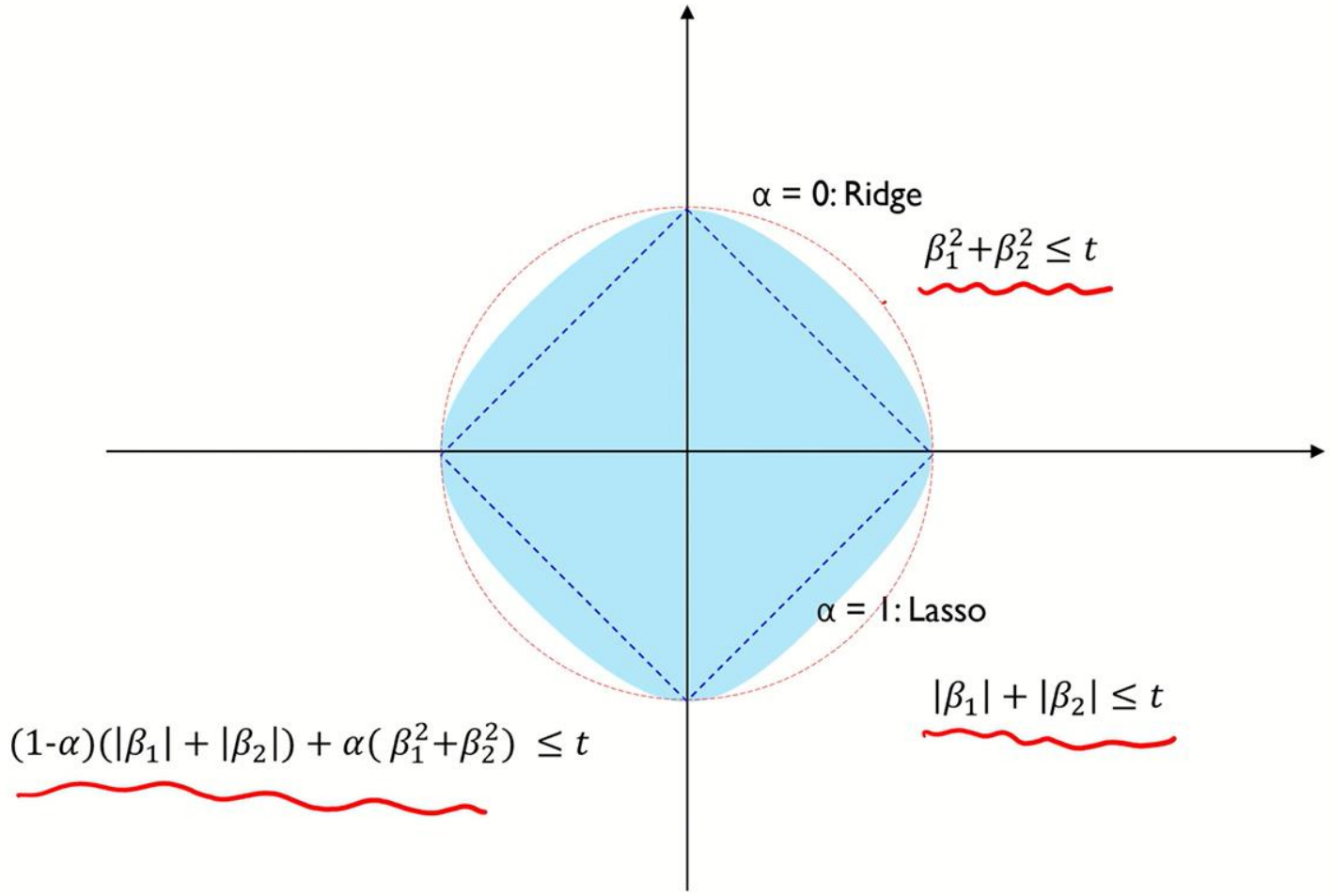
$$\hat{\beta}^{enet} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

$$subject\ to\ s_1 \sum_{j=1}^{p} |\beta_j| + s_1 \sum_{j=1}^{p} {\beta_j}^2 \leq t$$

⇓ Equivalent

$$\hat{\beta}^{enet} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 \right\}$$

3. Elastic Net Regression



Week4 과제

원래 팀 과제를 내려고 했으나 머신러닝 과정 자체가 직선적이고 실시간으로 수정해야 하는 부분이 있어서 당장 팀플이 어려울 수 있다고 생각해서 팀플이라는 형식 때문에 스트레스를 주는 것보다 개인 과제를 제공해서 복습 및 실습을 더 튼튼히 하고자 함.

따라서 week6에 진행될 예정이었던 팀 발표 수업을 취소하고, week5에서 스터디를 마무리함.

선형 모델 Linear Models – YouTube

위 링크를 들어가면 강의가 나오는데 유튜브 내 설명에도 있는 _3 선형 모델(Linear Models).ipynb – Colaboratory ipynb파일을 드라이브에 복사한 다음, 영상을 보면서 빈칸 채워 실행을 성공시키기.

(더 공부하고 싶다면…House Price Prediction 🏠 | Kaggle 따라하기를 권장)

Reference

[핵심 머신러닝] 선형회귀모델 1 (개요, 모델가정) - YouTube

선형 회귀분석의 4가지 기본가정

5.1 선형 모델 | Forecasting: Principles and Practice

선형 회귀 모델에서 '선형'이 의미하는 것은 무엇인가?

회귀 모델의 종류와 특징

[핵심 머신러닝] 정규화모델 1(Regularization 개념, Ridge Regression) – YouTube

설명이 부족하다면

설명이 부족하다면

KUBIG 2021-Fall ML STUDY