

# Multicollinearity

20182101019 김도혜

## 목차

1. 다중공선성이란?
2. 다중공선성의 영향과 문제들
3. 다중공선성 진단
4. 다중공선성의 해결방안

## 1. 다중공선성이란?

- 다중공선성 : 다중선형회귀분석에서 독립변수들 간에 강한 선형적 관계가 있는 경우
- 완전공선성 : 계획행렬(design matrix)의 열벡터들이 선형종속인 경우

\*선형종속 : 벡터 집합  $x_1, x_2, \dots, x_k$ 을 이루는 벡터의 선형조합이 영벡터가 되도록 하는 스칼라 계수  $c_1, c_2, \dots, c_k$ 이 존재하면, 선형종속이라 함.  $c_1x_1 + c_2x_2 + \dots + c_kx_k = 0$  (계수가 모두 0인 경우 제외)

→ 이 경우에는 최소제곱추정량을 구하는데 문제가 발생함.

\*최소제곱추정량을 구할 때,  $(X^TX)^{-1}$ 의 역행렬이 존재할 조건  
 $\exists (X^TX)^{-1} \Leftrightarrow X^TX$ 는 정방행렬이면서 full rank  
 $\Leftrightarrow \det(X^TX)^{-1} \neq 0$

→ 즉, 독립변수들이 선형종속관계에 있으면 full rank가 아니고,  $\det(X^TX)^{-1} = 0$ 이므로 역행렬이 존재하지 않고, 최소제곱추정량을 구할 수 없게 된다.

\*rank : 행렬의 열(행)벡터 중 서로 독립인 열(행)벡터의 최대 개수  
 $\text{rank } X \leq \min(n, p), X \in R^{n \times p}$  여기서 등호가 성립하면 full rank

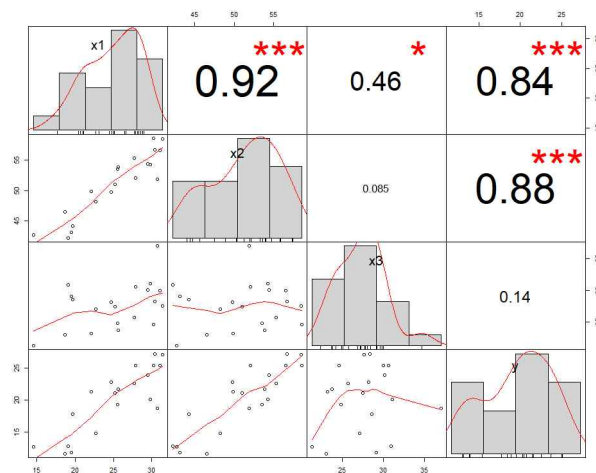
- 완전공선성은 아니더라도 독립변수들 사이에 강한 선형적 관계가 있는 경우 모형의 회귀계수 추정과 예측에 문제가 발생할 수 있음. →  $\det(X^TX)^{-1} \approx 0$ 이고 최소제곱추정량의 추정오차  $\sigma^2\{b\} = \sigma^2(X^TX)^{-1}$ 가 커짐

## 2. 다중공선성의 영향 및 문제점

- 일반적으로 회귀분석에서는 어떤 독립변수의 영향력을 파악할 때 다른 독립변수들을 모두 일정하다고 생각함. 즉, 독립변수들끼리 서로 독립이라는 것을 가정. 그래야 알아보고자 하는 변수의 영향력만을 알 수 있음
- 그러나 어느 두 독립변수가 서로에게 영향을 주고 있다면 둘 중 하나의 영향력을 검증할 때 다른 하나의 영향력을 완벽히 통제할 수 없음
- 영향 및 문제점
  - 1) 독립변수의 포함여부가 회귀변수들을 변화시킴.
  - 2) 이미 모형에 포함된 독립변수가 어떤 것들인지에 따라 한 독립변수의 추가제곱합이 달라짐.
  - 3) 회귀모형의 독립변수들이 서로 높은 상관성일 때 회귀계수의 추정표준편차는 커짐.
  - 4) 종속변수와 독립변수들 집단이 통계적으로 관계가 확실히 있음에도 불구하고, 추정된 회귀계수들 각각은 통계적으로 유의하지 않을 수 있음.

- 예시 데이터 (x1: 삼두근 피하지방 두께, x2: 허벅지 둘레, x3: 팔 둘레, y: 체지방)

<pre>&gt; head(ex)</pre>					<pre>&gt; round(cor(ex),3)</pre>				
x1	x2	x3	y		x1	x2	x3	y	
1	19.5	43.1	29.1	11.9	x1	1.000	0.924	0.458	0.843
2	24.7	49.8	28.2	22.8	x2	<b>0.924</b>	1.000	0.085	0.878
3	30.7	51.9	37.0	18.7	x3	0.458	0.085	1.000	0.142
4	29.8	54.3	31.1	20.1	y	0.843	0.878	0.142	1.000
5	19.1	42.2	30.9	12.9					
6	25.6	53.9	23.7	21.7					



## 다중공선성의 영향

### 1) 회귀계수와 회귀계수의 추정오차( $s\{b_k\}$ )의 영향

- 다른 변수의 포함 여부에 의해 회귀계수가 크게 달라짐.
- 추정된 회귀계수는 더 많은 독립변수들이 회귀모형에 추가될 때 점점 더 부정밀해짐.
- $s\{b_k\}$ 는 점점 커짐. 또한, 다중공선성이 있다면 예측된 회귀계수에 대해 분산을 증가시킴.

변수	b1	b2	b3	변수	$s\{b_1\}$	$s\{b_2\}$	$s\{b_3\}$
x1	0.857	-	-	x1	0.129	-	-
x2	-	0.857	-	x2	-	0.110	-
x3	-	-	0.199	x3	-	-	0.327
x1,x2	0.222	0.659	-	x1,x2	0.303	0.291	-
x1,x3	0.984	-	-0.308	x1,x3	0.128	-	0.176
x2,x3	-	0.872	0.068	x2,x3	-	0.112	0.161
x1,x2,x3	<b>-4.334</b>	<b>-2.857</b>	<b>-2.186</b>	x1,x2,x3	<b>3.015</b>	<b>2.582</b>	<b>1.595</b>

## 2) 추가제곱합의 영향

- 추가제곱합(extra sum of squares) : 원래 X변수(들)가 회귀모형에 있을 때의 오차제곱합과 새로운 X변수(들)가 회귀모형에 추가되었을 때 오차제곱합의 차이.

ex)  $SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17$

이 오차제곱합의 감소는  $X_1$ 이 이미 모형에 있을 때  $X_2$ 를 추가하는 것의 한계효과를 측정하는 것.

```
> anova(lm(y~x1,data=ex)) # x1만 포함된 모형
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	44.305	3.024e-06 ***
Residuals	18	143.12	7.95		

```
> anova(lm(y~x2,data=ex)) # x2만 포함된 모형
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	381.97	381.97	60.617	3.6e-07 ***
Residuals	18	113.42	6.30		

```
> anova(lm(y~x1+x2,data=ex)) # x1,x2가 포함된 모형
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	54.4661	1.075e-06 ***
x2	1	33.17	33.17	5.1284	0.0369 *
Residuals	17	109.95	6.47		

- $SSR(X_2)=381.97$  와 비교하여  $SSR(X_2|X_1)=33.17$  가 매우 작은 것은 모형에  $X_2$ 가 포함되는 것이  $X_1$ 과 거의 같은 정보를 주기 때문임 ( $X_2$ 은  $X_1$ 가 제공하고 있는 정보를 넘어서서 추가 정보를 제공하는 양이 적다는 것)

## 2) $\beta_k$ 검정예의 영향

- 다중회귀모형을 분석할 때 각각의 회귀계수를 차례차례  $\beta_k = 0$  ( $k = 1, \dots, p-1$ )인지를 결정하기 위해  $t^*$ 통계량을 살펴보게 됨.
- 그런데 다중공선성 문제가 있는 경우에는 분산이 크니까  $t^* = \frac{b_k}{s\{b_k\}}$ 의 분모가 커져서 0처럼 보이게 되고, t-test 결과 유의하지 않다는 결과가 나옴.
- 앞의 예시로 보면,

```
> summary(lm(y~x1+x2,data=ex))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-19.1742	8.3606	-2.293	0.0348 *
x1	0.2224	0.3034	0.733	0.4737
x2	0.6594	0.2912	2.265	0.0369 *

**F검정**  $H_0: \beta_1 = \beta_2 = 0$  결과는  $F^* = \frac{MSR}{MSE} = \frac{197.72}{6.47} = 29.8$  이고,  $F_{0.95;2,17} = 3.59$  이므로 기각.

→ 두 회귀계수는 0이 아니다. 유의하다.

**t검정**  $H_0: \beta_1 = 0$  결과는  $t^* = \frac{b_k}{s\{b_k\}} = \frac{0.2224}{0.3034} = 0.733$  이고  $t_{0.9875,17} = 2.46$  이므로 기각할 수 없음.

→ 회귀계수  $\beta_1$ 은 유의하지 않다.

$H_0: \beta_2 = 0$  결과는  $t^* = \frac{b_k}{s\{b_k\}} = \frac{0.6594}{0.2912} = 2.265$  이고  $t_{0.9875,17} = 2.46$  이므로 기각할 수 없음.

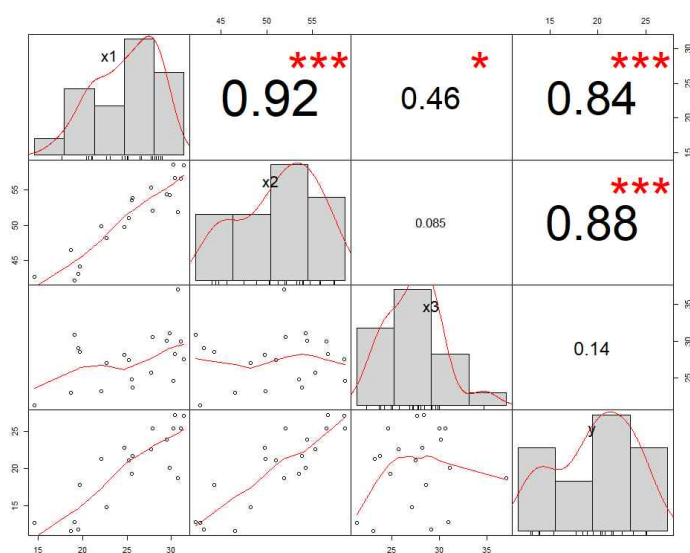
→ 회귀계수  $\beta_2$ 은 유의하다.

### 3. 진단

#### 1) 비정형적 진단

- 독립변수나 관측값의 포함 여부에 의한 추정회귀계수의 큰 변화
- 중요한 독립변수들에 대해 각 회귀계수 검정이 유의하지 않은 결과(t-value가 0에 가까움)
- 추정회귀계수가 사전 경험이나 이론적 고찰과 반대 방향의 부호를 가짐
- 두 독립변수의 상관행렬  $r_{XX}$ 에서 상관계수의 큰 값
- 독립변수들끼리 산점도를 그렸을 때 점들이 직선 주위에 밀집
- 하나의 독립변수를 종속변수로 두고, 나머지 독립변수들을 독립변수로 하여 선형회귀분석을 실시하였을 때 결정계수의 값이 1에 가까움

- 산점도, 상관계수로 판단



- x1과 x2의 상관관계가 높게 나타난다(0.92)
- x1과 x2의 산점도를 보면 직선 주위에 밀집되어 있다

- 회귀분석 결과로 판단 (x1,x2,x3 변수 모두 모형에 포함)

```
> fit<-lm(y~.,data=ex)
> summary(fit)

Call:
lm(formula = y ~ ., data = ex)

Residuals:
    Min       1Q   Median       3Q      Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173   0.258
x1           4.334       3.016   1.437   0.170
x2          -2.857       2.582  -1.106   0.285
x3          -2.186       1.595  -1.370   0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared:  0.8014,    Adjusted R-squared:  0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06

> vif(fit)
      x1      x2      x3
708.8429 564.3434 104.6060
```

- 모든 회귀계수들의 t-value는 0에 가까운데 F-value는 값이 큼
- F-test 결과는 유의한데, t-test 결과는 유의하지 않음

- 'x1' 변수 제거 후 회귀분석

```
> fit2<-lm(y~.-x1,data=ex)
> summary(fit2)

Call:
lm(formula = y ~ . - x1, data = ex)

Residuals:
    Min       1Q   Median       3Q      Max
-4.0777 -1.8296  0.1893  1.3545  4.1275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.99695     6.99732  -3.715  0.00172 **
x2           0.85088     0.11245   7.567 7.72e-07 ***
x3           0.09603     0.16139   0.595  0.55968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.557 on 17 degrees of freedom
Multiple R-squared:  0.7757,    Adjusted R-squared:  0.7493
F-statistic: 29.4 on 2 and 17 DF,  p-value: 3.033e-06

> vif(fit2)
      x2      x3
1.00722 1.00722
```

#### - 'x2' 변수 제거 후 회귀분석

```
> fit3<-lm(y~.-x2,data=ex)
> summary(fit3)

Call:
lm(formula = y ~ . - x2, data = ex)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8794 -1.9627  0.3811  1.2688  3.8942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7916     4.4883   1.513  0.1486
x1             1.0006     0.1282   7.803 5.12e-07 ***
x3            -0.4314     0.1766  -2.443  0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.496 on 17 degrees of freedom
Multiple R-squared:  0.7862,    Adjusted R-squared:  0.761
F-statistic: 31.25 on 2 and 17 DF,  p-value: 2.022e-06

> vif(fit3)
           x1           x3
1.265118 1.265118
```

#### - 'x3' 변수 제거 후 회귀분석

```
> summary(lm(y~.-x3,data=ex))

Call:
lm(formula = y ~ . - x3, data = ex)

Residuals:
    Min       1Q   Median       3Q      Max
-3.9469 -1.8807  0.1678  1.3367  4.0147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293  0.0348 *
x1             0.2224     0.3034   0.733  0.4737
x2             0.6594     0.2912   2.265  0.0369 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared:  0.7781,    Adjusted R-squared:  0.7519
F-statistic: 29.8 on 2 and 17 DF,  p-value: 2.774e-06

> vif(lm(y~.-x3,data=ex))
           x1           x2
6.825239 6.825239
```

- 'x3'을 종속변수로, 'x1','x2'를 독립변수로 하고 회귀분석

```
> fit5<-lm(x3~.-y,data=ex)
> summary(fit5)

Call:
lm(formula = x3 ~ . - y, data = ex)

Residuals:
    Min       1Q   Median       3Q      Max
-0.58200 -0.30625  0.02592  0.29526  0.56102

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.33083    1.23934   50.29  <2e-16 ***
x1           1.88089     0.04498   41.82  <2e-16 ***
x2          -1.60850     0.04316  -37.26  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.377 on 17 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9893
F-statistic: 880.7 on 2 and 17 DF,  p-value: < 2.2e-16
```

- 결정계수의 값이 1에 가까움

#### ▪ 비정형적 진단의 한계점

- 다중공선성의 영향에 대한 양적인 측도 제시 불가
- 다중공선성이 없는 때에도 위와 같은 모습들이 발견될 수도 있음

## 2) 정형적 진단

- 분산팽창인수(VIF) : 독립변수들이 선형관계가 없는 경우와 비교했을 때 추정회귀계수들의 분산이 얼마나 크게 (팽창) 되었는지를 측정하는 지수

\*다중회귀모형 :  $Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i$ ,  $\epsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, n$

\*표준화회귀모형(표준화한 변수로부터 얻어진 회귀모형)

$$Y_i^* = \beta_1^* X_{i,1}^* + \dots + \beta_{p-1}^* X_{i,p-1}^* + \epsilon_i^*$$

\*여기서  $r_{XX}(X \text{ 변수들의 상관행렬}) = X^T X$

$$r_{XX}^{(p-1) \times (p-1)} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{21} & 1 & \dots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \dots & 1 \end{bmatrix}, \quad X = \begin{bmatrix} X_{11}^* & \dots & X_{1,p-1}^* \\ X_{21}^* & \dots & X_{2,p-1}^* \\ \vdots & & \vdots \\ X_{n1}^* & \dots & X_{n,p-1}^* \end{bmatrix} \Rightarrow \begin{matrix} X^T X \\ (p-1) \times (p-1) \end{matrix} = r_{XX}$$



$$\sum (X_{i1}^*)^2 = \sum \left( \frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}s_1} \right)^2 = \frac{\sum (X_{i1} - \bar{X}_1)^2}{n-1} \div s_1^2 = 1 \text{ 이고}$$

$$\sum X_{i1}^* X_{i2}^* = \sum \left( \frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}s_1} \right) \left( \frac{X_{i2} - \bar{X}_2}{\sqrt{n-1}s_2} \right)$$

$$= \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{s_1 s_2} = \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2}} = r_{12}$$

\* VIF는 최소제곱법 추정회귀계수의 정밀도(precision), 즉 추정회귀계수의 분산으로부터 측정됨

$$\sigma^2\{b\} = \sigma^2 (X^T X)^{-1}$$

$\sigma^2\{b^*\} = (\sigma^*)^2 r_{XX}^{-1}$  여기서  $r_{XX}^{-1}$  행렬의  $k$  번째 대각원소를  $(VIF)_k$ 라 표기하면,

$b_k^* (k=1, \dots, p-1)$ 는 다음과 같다.  $\sigma^2\{b_k^*\} = (\sigma^*)^2 (VIF)_k$

→ 대각원소  $(VIF)_k$ 는  $b_k^*$ 의 분산팽창인수(variance inflation factor, VIF)라고 함.

$$* (VIF)_k = \frac{1}{1 - R_k^2}, \quad k = 1, 2, \dots, p-1$$

여기서  $R_k^2$ 은  $X_k$ 를 다른  $p-2$ 개의  $X$ 변수들로 회귀했을 때의 다중결정계수. 따라서  $\sigma^2\{b^*\} = \frac{(\sigma^*)^2}{1 - R_k^2}$  임.

\*  $R_k^2=0$ ( $X_k$ 가 다른  $X$ 변수들과 선형의 관계가 아닐 때) 이면  $(VIF)_k=1$

$R_k^2=1$ ( $X_k$ 가 다른  $X$ 변수들과 완벽한 선형관계) 이면  $(VIF)_k \approx \infty$

$R_k^2$ 이 1에 가깝다는 것은  $X_k$ 가 다른 독립변수로 표현될 수 있다는 것이고,  $(VIF)_k$ 가 크다는 것은,  $X_k$ 가 다른 독립변수들에 의해 선형 함수로 표현될 수 있다는 것. 그래서 결정계수들 중에 적어도 하나의 값은 1에 가까울 때 다중공선성 존재한다고 함.

\* VIF의 최대값이 10을 초과하는 경우 다중공선성이 존재한다고 함.

\* **VIF의 평균값** 역시 추정표준화회귀계수  $b_k^*$ 들이 참값인  $\beta_k^*$ 들로부터 얼마나 멀리 떨어져 있는지에 대한 다중공선성의 강도의 정보를 제공함. 즉, 큰 VIF의 값들은 평균적으로 표준화회귀계수의 추정된 값과 참값 사이의 큰 차이를 만듦.

$$E\left\{\sum_{k=1}^{p-1} (b_k^* - \beta_k^*)^2\right\} = (\sigma^*)^2 \sum_{k=1}^{p-1} (VIF)_k \quad \dots\dots\dots (1)$$

여기서  $R_k^2=0$  일 때, 제곱오차합의 평균은 아래와 같음.

$$E\left\{\sum_{k=1}^{p-1} (b_k^* - \beta_k^*)^2\right\} = (\sigma^*)^2 (p-1) \quad (VIF)_k \equiv 1 \text{ 일때} \quad \dots\dots\dots (2)$$

→ (1)과 (2)의 비율은 제곱오차의 합에 미치는 다중공선성의 영향에 대한 유용한 정보를 제공함.

$$\frac{(\sigma^*)^2 \sum_{k=1}^{p-1} (VIF)_k}{(\sigma^*)^2 (p-1)} = \frac{\sum_{k=1}^{p-1} (VIF)_k}{(p-1)} = (\overline{VIF})$$

즉, 이 비율은 VIF의 평균이 되고, 이 평균이 1보다 크다는 것은 다중공선성 문제로 생각.

```
> fit<-lm(y~,data=ex)
> vif(fit)
      x1      x2      x3
708.8429 564.3434 104.6060

> mean(vif(fit))
[1] 459.2641
```

- 제곱오차합의 기댓값은 X변수들이 무상관일 경우에 비해 거의 460배.
- 또한, 변수 x3의 VIF가 104.60으로 큰 값을 보임. ( $r_{13}^2, r_{23}^2$ 은 크지 않음)
- x3이 x1과 x2의 조합과 강하게 연관되어 있음을 알 수 있음.

- 상태지수(condition number) :  $X^T X$ 의 최대고유값과 최소고유값의 차이

\* 다중공선성 문제가 있으면 상태지수(condition number)의 값이 큼.

\*  $k = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$ , 보통  $k \geq 30$ 이면 다중공선성이 있다고 판단

```
> X = cbind(rep(1,length(ex$x1)),ex$x1,ex$x2,ex$x3)
> eigen.X = eigen(t(X)%*%X)
> eigen.X
eigen() decomposition
$values
[1] 8.129024e+04 2.942499e+02 1.198158e+02 6.165867e-04

$vectors
      [,1]      [,2]      [,3]      [,4]
[1,] -0.01560414 0.00965285 -0.03861731 0.99908560
[2,] -0.40266387 -0.18165656 0.89663553 0.03012348
[3,] -0.80607618 -0.39389683 -0.44093021 -0.02582705
[4,] -0.43342763 0.90097337 -0.01157473 -0.01592177

> # condition number
> sqrt(max(eigen.X$values)/min(eigen.X$values))
[1] 11482.12
```

- 상태지수가 매우 크기 때문에 다중공선성 문제가 있다고 할 수 있음.

#### 4. 다중공선성 해결방안

##### 1) 변수제거

: 다중공선성 문제를 일으키는 변수를 제외함. 일반적으로 다중공선성 문제를 일으키는 변수 중 종속변수와의 상관관계가 높은 것을 남겨두고, 상관관계의 차이가 거의 없다면 해석이 용이한 독립변수를 남겨둠.

##### 2) 능형회귀(ridge regression), 주성분회귀

: 다중공선성 문제일 때, 최소고유값  $\lambda_p \approx 0$  이고, 선형독립일 때는  $\lambda_p = 0$ 임. 이때,  $\frac{1}{\lambda_p}$  역수 취하면 엄청 커

지고 분산과  $MSE(MSE(\hat{\beta}^{LSE}) = \sum_{i=1}^p \frac{\sigma^2}{\lambda_i})$ 도 커짐. **결국 분산이 커서 문제가 발생**

$$\text{*능형회귀 : } MSE(\hat{\beta}^{\text{Ridge}}) = \sum_{i=1}^p \frac{\sigma^2}{\lambda_i + \lambda}$$

→  $\lambda_i + \lambda$  'λ(lambda)'라는 양수를 더해줘서 0에 가깝지 않게 해주면 분산이 엄청 크지는 않게됨. 그렇게 분산을 줄이는 아이디어

$$\text{*주성분회귀 : } MSE(\hat{\beta}^{\text{PCA}}) = \sum_{i=1}^q \frac{\sigma^2}{\lambda_i}$$

→ 0에 가까운 고유값들은 버리고 적당한 값을 가지는 것만 쓰는 것