STAT409 Takehome Midterm

2021320322 윤민서

1.

(a)

주어진 분포의 likelihood function을 l이라고 하면

$$l(\mu_i \mid y_{i1}, y_{i2}, \cdots, y_{iJ}) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

y_{ij}'s are independent

$$L := \log l$$

$$L(\mu_i \mid y_{i1}, y_{i2}, \cdots, y_{iJ})$$

$$= \log \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

$$= \sum_{j=1}^{J} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

$$= J \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{j=1}^{J} \frac{(y_{ij}-\mu_i)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu_i} = \frac{1}{\sigma^2} \sum_{j=1}^{J} (y_{ij}-\mu_i) \overset{\text{set}}{=\!=} 0$$

$$\Rightarrow \sum_{j=1}^{J} y_{ij} = J\mu_i \quad \cdots ①$$

$$\frac{\partial^2 L}{\partial \mu_i^2} = -\frac{J}{\sigma^2} < 0 \quad \cdots ②$$

By ① & ②, $\hat{\mu}_i^{MLE} = \frac{1}{J}\sum_{j=1}^{J} y_{ij} = \bar{y}_i$ for $i = 1, \cdots, I$

(b)

$y = X\mu + \varepsilon$, $y \in \mathbb{R}^n$, $\varepsilon \in \mathbb{R}^n$, $\mu \in \mathbb{R}^I$ 에서

식이 성립하기 위한 X의 차원은 $X \in \mathbb{R}^{n \times I}$

이때 one-way anova에서의 design matrix X는 다음과 같다.

$$X = \begin{array}{c} \\ J\left\{ \begin{array}{c} \\ \\ \\ \\ \end{array} \right. \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \left. \begin{array}{c} \\ \\ \\ \\ \end{array} \right\} J \end{array}$$

여기에서 X는 각각의 observation이 어떤 group에 속해 있는지 나타낸다.

μ_i의 OLS estimator는 아래와 같이 계산된다.

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} (y - X\mu)^T (y - X\mu) \quad (\mu = (\mu_1, \mu_2, \cdots, \mu_I))$$

$$S(\mu) := (y - X\mu)^T (y - X\mu)$$

$$\frac{\partial S(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(y^T y - \mu^T X^T y - y^T X \mu + \mu^T X^T X \mu\right)$$

$$= \frac{\partial}{\partial \mu}\left(-2\mu^T X^T y + \mu^T X^T X \mu\right)$$

$$= -2X^T y + 2X^T X \mu \overset{set}{=} 0$$

$$2X^T X \mu = 2X^T y \quad \Rightarrow (X^T X)\mu = X^T y$$

$$\Rightarrow (X^T X)^{-1}(X^T X)\mu = (X^T X)^{-1} X^T y$$

$\therefore \hat{\mu} = (X^T X)^{-1} X^T y$ ($\because$ 위에서 X의 column들이 independent 하다는 것을 알 수 있다. $\Rightarrow X^T X$ is invertible.)

$$X^T X = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} J & 0 & \cdots & 0 \\ 0 & J & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & J \end{bmatrix} \qquad (X^T X)^{-1} = \begin{bmatrix} 1/J & 0 & \cdots & 0 \\ 0 & 1/J & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 1/J \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}^T \begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = \begin{bmatrix} y_{11} + \cdots + y_{1J} \\ y_{21} + \cdots + y_{2J} \\ \vdots \\ y_{I1} + \cdots + y_{IJ} \end{bmatrix}$$

$$(X^T X)^{-1} X^T y = \begin{bmatrix} 1/J & 0 & \cdots & 0 \\ 0 & 1/J & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 1/J \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{J} y_{1j} \\ \sum_{j=1}^{J} y_{2j} \\ \vdots \\ \sum_{j=1}^{J} y_{Ij} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_I \end{bmatrix}$$

(c)

Fitted value of y denoted by $\hat{y}$ is

$$\hat{y} = X\hat{\mu} = X(X^TX)^{-1}X^Ty$$

$$\therefore P_X = X(X^TX)^{-1}X^T$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1/J & 0 & \cdots & 0 \\ 0 & 1/J & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 1/J \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}^T = \begin{bmatrix} 1/J & 0 & \cdots & 0 \\ 1/J & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1/J & 0 & \cdots & 0 \\ 0 & 1/J & \cdots & 0 \\ 0 & 1/J & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1/J & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/J \\ 0 & 0 & \cdots & 1/J \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/J \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}^T$$

$$= \begin{bmatrix} 1/J & \cdots & 1/J & 0 & \cdots & 0 & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 1/J & \cdots & 1/J & 0 & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/J & \cdots & 1/J & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1/J & \cdots & 1/J & & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & & \ddots & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 1/J & \cdots & 1/J \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 1/J & \cdots & 1/J \end{bmatrix}$$

(d)

Fitted value of y

: $\hat{y} = P_X y = X(X^TX)^{-1}X^Ty$

$$= \begin{bmatrix} 1/J & \cdots & 1/J & 0 & \cdots & 0 & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 1/J & \cdots & 1/J & 0 & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/J & \cdots & 1/J & & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1/J & \cdots & 1/J & & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & \ddots & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 1/J & \cdots & 1/J \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 1/J & \cdots & 1/J \end{bmatrix} \begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_2 \\ \vdots \\ \bar{y}_I \\ \vdots \\ \bar{y}_I \end{bmatrix}$$

$\hat{e} = y - \hat{y} = y - P_X y = (I - P_X)y$

$$= \begin{bmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1J} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2J} - \bar{y}_2 \\ \vdots \\ y_{I1} - \bar{y}_I \\ \vdots \\ y_{IJ} - \bar{y}_I \end{bmatrix}$$

(e)

문제에서 언급되었듯이 MS_{trt}와 MSE가 independent함을 보이는 것은

SS_{trt}와 SSE가 independent함을 보이는 것과 동일하다.

ANOVA Decomposition이 성립하는지 확인함으로써, 즉

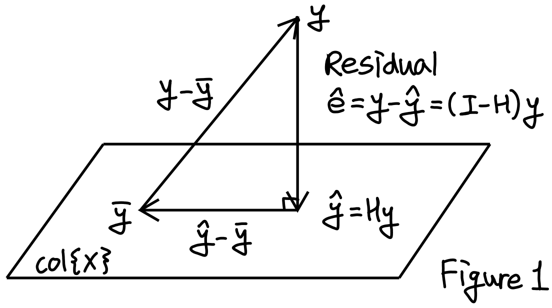


Figure 1

$$SST = \|y - \bar{y}\|_2^2 = \|y - \hat{y} + \hat{y} - \bar{y}\|_2^2$$
$$= \|y - \hat{y}\|_2^2 + \|\hat{y} - \bar{y}\|_2^2 + 2(y - \hat{y})^T(\hat{y} - \bar{y})$$
$$= SSE + SS_{trt} + 2(y - \hat{y})^T(\hat{y} - \bar{y})$$ 에서

$(y-\hat{y})^T(\hat{y}-\bar{y}) = 0 \Rightarrow$ $\underbrace{(y-\hat{y})}_{\text{residual}}$과 $\underbrace{(\hat{y}-\bar{y})}_{\text{treatment}}$이 orthogonal함을 보임으로써

SS_{trt}와 SSE가 independent함을 알 수 있다. (∵ 이 경우에서는 $\hat{y}$가 $\bar{y}_i$들의 vector이므로 주어진 상황에 적용 가능)

$$(y-\hat{y})^T(\hat{y}-\bar{y})$$
$$= (y - P_X y)^T(P_X y - \bar{y})$$
$$= y^T P_X y - y^T P_X^T P_X y - y^T \bar{y} + y^T P_X^T \bar{y}$$
(∵ P_X is idempotent and symmetric, 증명은 하단에)
$$= y^T P_X y - y^T P_X y - y^T \bar{y} + y^T P_X^T \bar{y}$$
(∵ P_X is symmetric)
$$= y^T P_X \bar{y} - y^T \bar{y}$$
(∵ $\bar{y} = P_X \bar{y}$, by Figure 1)
$$= y^T P_X \bar{y} - y^T P_X \bar{y}$$
$$= 0$$

cf) $$P_X P_X = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}X^T$$
$$= P_X$$

$$P_X^T = \{X(X^TX)^{-1}X^T\}^T$$
$$= X\{(X^TX)^{-1}\}^T X^T$$
$$= X(X^TX)^{-1}X^T$$
$$= P_X$$

(f)

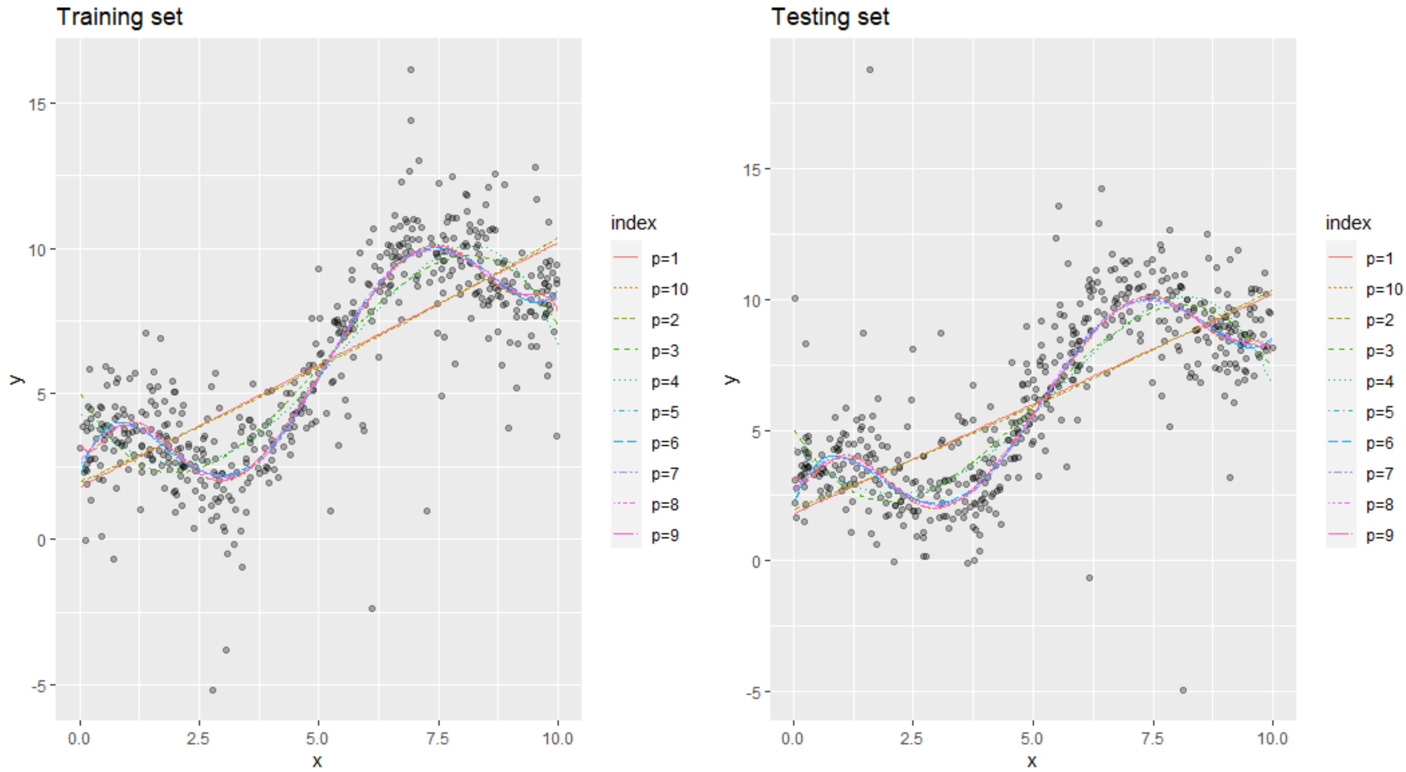
STAT409_Midterm_2021320322.R에 과정 첨부

결과
MStrt = 11.701
MSE = 1.338
F = 8.746

2.

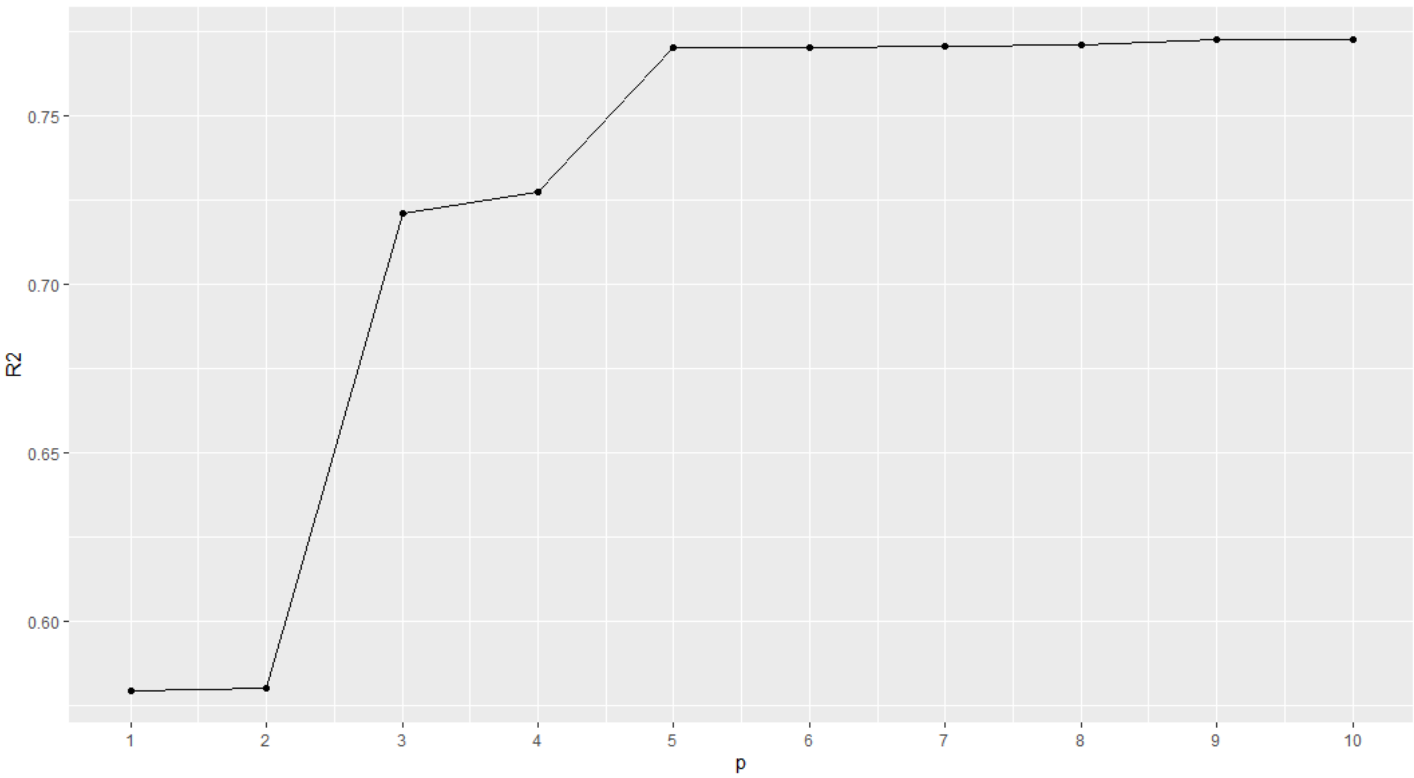
(a)

STAT409_Midterm_2021320322.R에 과정 첨부



(b)

STAT409_Midterm_2021320322.R에 과정 첨부



(c)

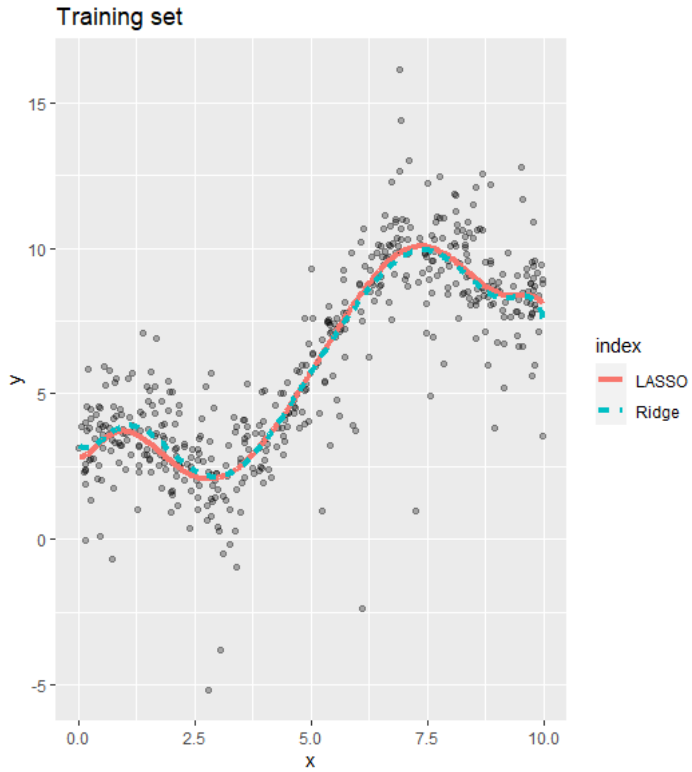
2-(b)의 결과에서 확인할 수 있듯이, polynomial의 차수가 높아질수록 R-squared가 단순히 커지고 있는 경향을 볼 수 있다. 즉, 모형이 복잡할수록 R-squared는 계속해서 커지기만 할 것이다. 이러한 경우 polynomial의 차수가 증가함에 따라 모형이 training set에 더욱 친화적이게 적합이 되기 때문에 training-error는 계속해서 감소하게 될 것이다.
그러나 R-squared를 model selection의 지표로 결정한다면 단순히 복잡한 모형을 선호하게 된다는 것과 같다. 모형이 복잡할수록 bias는 감소하겠지만 variance는 증가할 것이고 모형의 일반화 성능이 감소하게 될 것이다. 종합적으로 R-squared를 model selection의 지표로 사용한다면 training set을 지나치게 많이 고려하여 traning-error가 낮지만 일반화 성능이 떨어지는 모형을 선택하게 될 것이고, 이것은 overfitting이 될 가능성이 높은 모형을 선택하게 되는 것과 같으므로 R-squared는 model selection의 지표로 사용되기에 적절하지 않다.

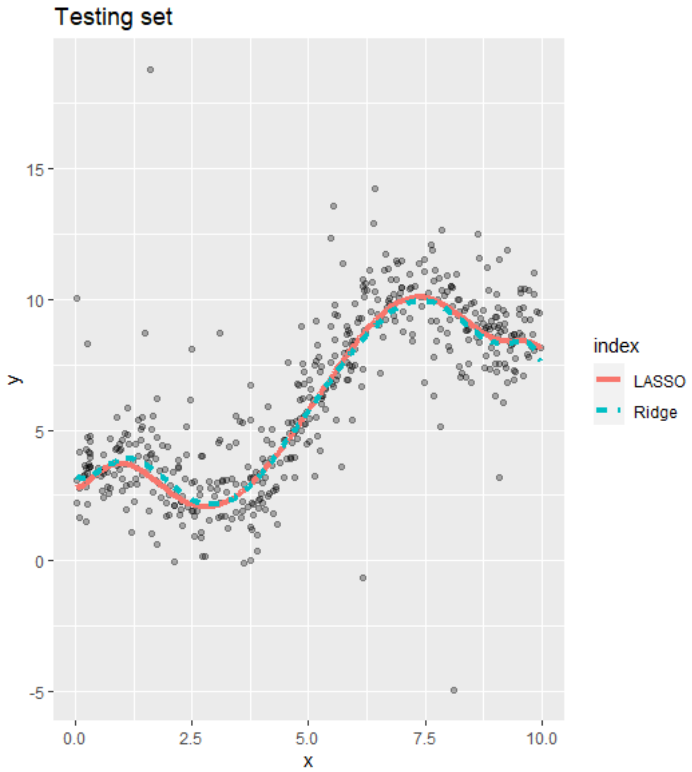
(d)

STAT409_Midterm_2021320322.R에 과정 첨부
BIC에 근거했을 때 p = 5인 모델을 선택한다.

(e)

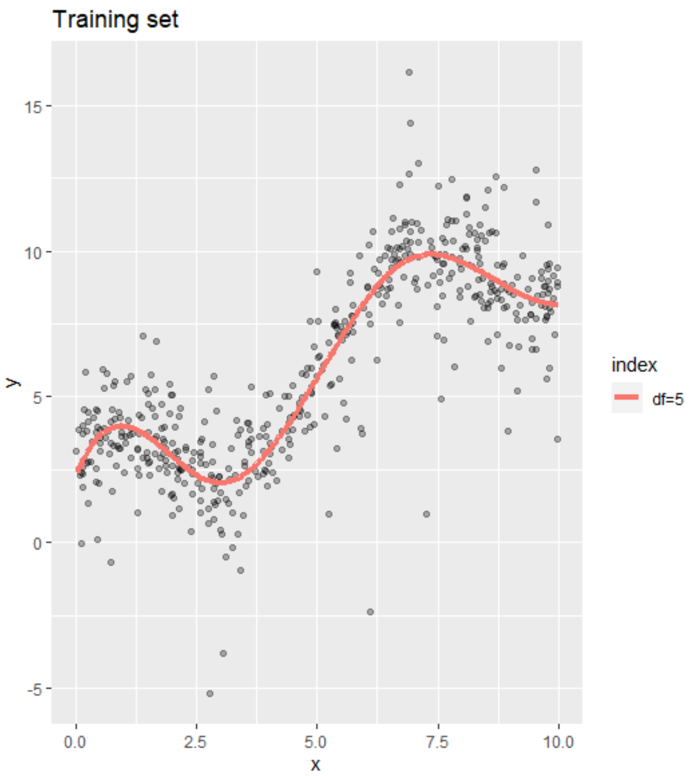
STAT409_Midterm_2021320322.R에 과정 첨부

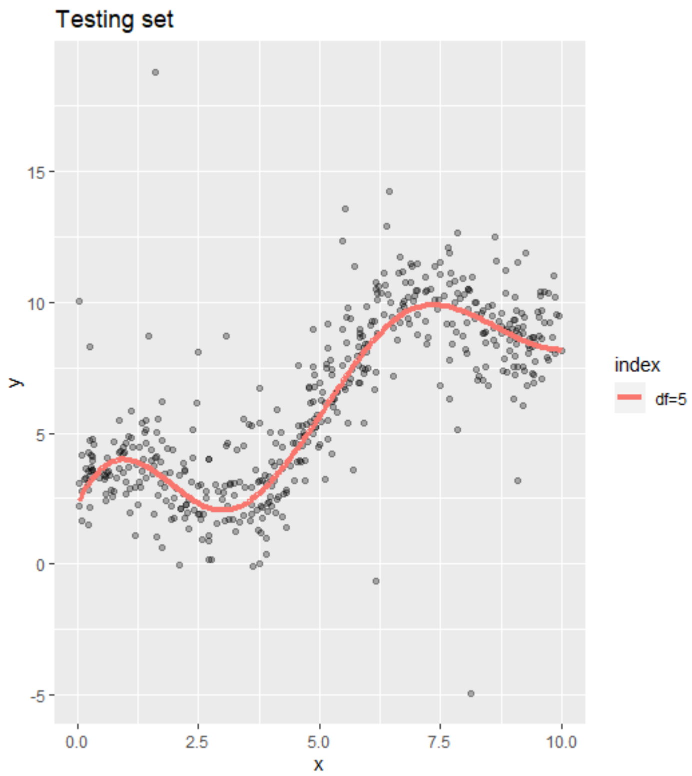




(f)

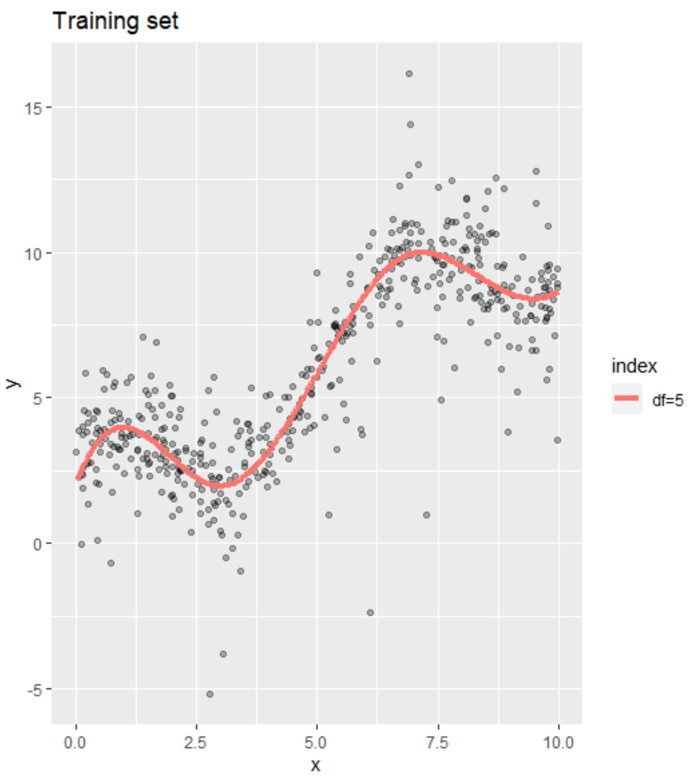
STAT409_Midterm_2021320322.R에 과정 첨부
Tuning의 기준은 BIC로 결정하였다.
BIC에 근거했을 때 df = 5인 모델을 선택한다.





(g)

STAT409_Midterm_2021320322.R에 과정 첨부
Tuning의 기준은 BIC로 결정하였다.
BIC에 근거했을 때 df = 5인 모델을 선택한다.



(h)

2-(g)에서 fitting한 B-spline median regression model을 가장 좋은 모형이라고 생각할 수 있다. B-spline regression 특성 상 자유도를 직접 조정할 수 있기 때문에 overfitting을 더욱 효과적으로 방지할 수 있다. 또한 B-spline regression은 계산적인 면에서 안정적이다. 그리고 scatter plot을 보았을 때 곳곳에 outlier로 의심되는 observation들이 있다. median regression은 mean regression에 비해서 outlier에 robust하다는 장점이 있기 때문에 주어진 데이터셋과 같은 데이터를 적합할 때 유용하게 사용할 수 있다.
이러한 면에서 B-spline regression의 장점과 median regression의 장점을 모두 가진 B-spline median regression model을 가장 좋은 모형으로 선택할 것을 고려할 수 있다.

정량적인 지표를 알아보기 위하여 testing error를 확인해 보았다.
(STAT409_Midterm_2021320322.R에 과정 첨부)
B-spline median regression에서의 testing error가 가장 낮게 측정된 것을 확인할 수 있었고, 이것을 근거로 B-spline median regression을 가장 좋은 모형으로 선택한다.