Homework Assignment2

김경민(20210344)

Question 1

Ozone Data

```
library( ElemStatLearn )
data( ozone )
n = nrow( ozone ) ; p = 1

head( ozone )
```

```
##   ozone radiation temperature wind
## 1    41       190          67  7.4
## 2    36       118          72  8.0
## 3    12       149          74 12.6
## 4    18       313          62 11.5
## 5    23       299          65  8.6
## 6    19        99          59 13.8
```

회귀식 $\text{ozone}_i = \beta_0 + \beta_1 \text{wind}_i + \epsilon_i$ 의 설명변수와 반응변수는 다음과 같이 만들 수 있다.

- y: Ozone 벡터 $\in \mathbb{R}^n$
- X: Wind 벡터 $\in \mathbb{R}^n$

```
y = ozone$ozone
x = ozone$wind
```

뒤에서 confidence interval과 predict interval 을 구할 때 필요한 $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ 을 계산하자.

- Sxx: S_{xx}
- Sxy: S_{xy}
- beta1_hat: $\hat{\beta}_1$
- beta0_hat: $\hat{\beta}_0$
- sse : SSE
- sigma2_hat: $\hat{\sigma}^2$

```
Sxx = sum( (x - mean(x))^2 )
Sxy = sum( (x - mean(x))*(y - mean(y)) )

beta1_hat = Sxy / Sxx
beta0_hat = mean(y) - beta1_hat * mean(x)

sse = sum( (y - ( beta0_hat + beta1_hat * x ) )^2 )
sigma2_hat = sse / (n - ( p + 1 ))


data.frame( beta0_hat, beta1_hat, sigma2_hat )
```

```
##   beta0_hat beta1_hat sigma2_hat
## 1  99.05425 -5.730622   697.6136
```

1.1 Combine the results and make a dataframe `res` with columns.

- `wind`: the sequence of x

Sequence of predictor

```
xval = data.frame( wind = seq( min( ozone$wind ), max( ozone$wind), length.out = 20 ) )
c(head( xval ) , tail( xval ) )
```

```
## $wind
## [1] 2.300000 3.268421 4.236842 5.205263 6.173684 7.142105
##
## $wind
## [1] 15.85789 16.82632 17.79474 18.76316 19.73158 20.70000
```

Confidence interval

Mean response $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 의 표준오차 $\mathrm{SE}\left(\hat{\mu}_0\right)$는 $\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\left(x_0-\bar{x}\right)^2}{S_{xx}}\right)}$ 이다. 따라서 confidence interval의 하한은 $\hat{\mu}_0 - t_{n-2,1-\alpha/2} \times \mathrm{SE}\left(\hat{\mu}_0\right)$ 이고 상한은 $\hat{\mu}_0 + t_{n-2,1-\alpha/2} \times \mathrm{SE}\left(\hat{\mu}_0\right)$ 이다.

- `qt0.025`: $t_{n-2,\,0.025}$
- `se.fit`: $\mathrm{SE}\left(\hat{\mu}_0\right)$
- `conf.lwr`: the lower limit of the confidence interval
- `conf.upr`: the upper limit of the confidence interval

```
qt0.025 = qt(p = 0.025, df = n-2, lower.tail = FALSE)

conf.lwr = rep( 0, 20 ) ; conf.upr = rep( 0, 20 )
for( i in 1:20 ){
  se.fit = sqrt( sigma2_hat * ( (1/n) + (xval[i,1] - mean(x))^2 / Sxx ) )
  conf.lwr[i] = beta0_hat + beta1_hat * xval[i,1] - qt0.025 * se.fit
  conf.upr[i] = beta0_hat + beta1_hat * xval[i,1] + qt0.025 * se.fit
}

cbind(head( conf.lwr ), head( conf.upr ))
```

```
##          [,1]     [,2]
## [1,] 74.06547 97.68217
## [2,] 69.73237 90.91596
## [3,] 65.36049 84.18854
## [4,] 60.93328 77.51643
## [5,] 56.42502 70.92538
## [6,] 51.79556 64.45553
```

Predict interval

Prediction value of the response $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon_0$의 표준오차 SE $\left(\hat{Y}_0\right)$는 $\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$이다. 따라서 confidence interval의 하한은 $\hat{Y}_0 - t_{n-2,1-\alpha/2} \times \mathrm{SE}\left(\hat{Y}_0\right)$이고 상한은 $\hat{Y}_0 + t_{n-2,1-\alpha/2} \times \mathrm{SE}\left(\hat{Y}_0\right)$이다.

- `qt0.025`: $t_{n-2,\,0.025}$
- `se.pred`: SE $\left(\hat{Y}_0\right)$
- `pred.lwr`: the lower limit of the prediction interval
- `pred.upr`: the upper limit of the prediction interval

```r
qt0.025 = qt(p = 0.025, df = n-2, lower.tail = FALSE)

pred.lwr = rep( 0, 20 ) ; pred.upr = rep( 0, 20 )
for( i in 1:20 ){
  se.pred = sqrt( sigma2_hat * ( 1 + (1/n) + (xval[i,1] - mean(x))^2 / Sxx ) )
  pred.lwr[i] = beta0_hat + beta1_hat * xval[i,1] - qt0.025 * se.pred
  pred.upr[i] = beta0_hat + beta1_hat * xval[i,1] + qt0.025 * se.pred
}
```

Make a dataframe `res`

```r
res = data.frame( "wind" = xval,
                  "conf.lwr" = conf.lwr,
                  "conf.upr" = conf.upr,
                  "pred.lwr" = pred.lwr,
                  "pred.upr" = pred.upr )

res
```

```
##          wind   conf.lwr  conf.upr     pred.lwr  pred.upr
## 1   2.300000  74.065471 97.682169  32.21005715 139.53758
## 2   3.268421  69.732368 90.915963  26.91491055 133.73342
## 3   4.236842  65.360486 84.188535  21.58629604 127.96273
## 4   5.205263  60.933284 77.516429  16.22379488 122.22592
## 5   6.173684  56.425021 70.925382  10.82704705 116.52336
## 6   7.142105  51.795564 64.455531   5.39575464 110.85534
## 7   8.110526  46.984749 58.167036  -0.07031537 105.22210
## 8   9.078947  41.913342 52.139134  -5.57132876  99.62380
## 9  10.047368  36.505552 46.447615 -11.10738251  94.06055
## 10 11.015789  30.733738 41.120120 -16.67850380  88.53236
## 11 11.984211  24.640034 36.114515 -22.28464974  83.03920
## 12 12.952632  18.304488 31.350752 -27.92570781  77.58095
## 13 13.921053  11.802993 26.752937 -33.60149697  72.15743
## 14 14.889474   5.190877 22.265745 -39.31176944  66.76839
## 15 15.857895  -1.495484 17.852796 -45.05621310  61.41353
## 16 16.826316  -8.232804 13.490807 -50.83445446  56.09246
## 17 17.794737 -15.006068  9.164763 -56.64606210  50.80476
## 18 18.763158 -21.805387  4.864772 -62.49055064  45.54994
## 19 19.731579 -28.624073  0.584149 -68.36738495  40.32746
## 20 20.700000 -35.457486 -3.681747 -74.27598468  35.13675
```

Question 2

Ozone Data

```
rm(list = ls())
library(ElemStatLearn)
data(ozone)
n = nrow( ozone )
p = 3

y = matrix( ozone$ozone )
X = cbind( "(intercept)" = rep(1,n),
           "wind" = ozone$wind,
           "temperature" = ozone$temperature,
           "radiation" = ozone$radiation )

beta_hat = solve( t(X) %*% X ) %*% t(X) %*% y
data.frame( beta_hat )
```

```
##                  beta_hat
## (intercept) -64.23208116
## wind         -3.33759763
## temperature   1.65120780
## radiation     0.05979717
```

```
y_hat = X %*% beta_hat
resid = y - y_hat
SSE = sum( resid^2 )
SST = sum( (y - mean(y))^2 )
SSR = SST - SSE



# Anova Table
Source = c( "Regression", "Error", "Total" )
df = c( p, n - ( p + 1 ), n - 1 )
SS = c( SSR, SSE, SST )
MS = c( SSR / p, SSE / ( n - ( p + 1 ) ), NA )
Fstat = c(MS[1] / MS[2], NA, NA)

data.frame( Source,df, SS, MS, Fstat )
```

```
##       Source  df        SS         MS    Fstat
## 1 Regression   3  73837.79 24612.5972 54.90663
## 2      Error 107  47964.12   448.2628       NA
## 3      Total 110 121801.91         NA       NA
```

```
fstat1 = Fstat[1]
```