회귀분석론 HW2

212STG18 예지혜

2021년 3월 6일

1.

```
commercial <- read.table("Commercial properties.txt")
names(commercial) <- c("Y","X1","X2","X3","X4")
head(commercial)
```

```
##      Y X1    X2   X3     X4
## 1 13.5  1  5.02 0.14 123000
## 2 12.0 14  8.19 0.27 104079
## 3 10.5 16  3.00 0.00  39998
## 4 15.0  4 10.70 0.05  57112
## 5 14.0 11  8.97 0.07  60000
## 6 10.5 15  9.45 0.24 101385
```

(a)

```
stem(commercial$X1)
```

```
## 
##   The decimal point is at the |
## 
##    0 | 0000000000000000
##    2 | 000000000000000000000000
##    4 | 00000
##    6 | 0
##    8 | 0
##   10 | 00
##   12 | 00000
##   14 | 0000000000000
##   16 | 0000000000
##   18 | 000
##   20 | 00
```

```
stem(commercial$X2)
```

```
## 
##   The decimal point is at the |
## 
##    2 | 0
##    4 | 080003358
##    6 | 012613
##    8 | 00001223456001555689
##   10 | 0133445666777781233446666668
##   12 | 00011115777889002
##   14 | 6
```

```
stem(commercial$X3)
```

```
## 
##   The decimal point is 1 digit(s) to the left of the |
## 
##   0 | 00000000000000000000000000000023333333333344444455555566678889
##   1 | 023444469
##   2 | 1223477
##   3 | 3
##   4 | 
##   5 | 7
##   6 | 0
##   7 | 3
```

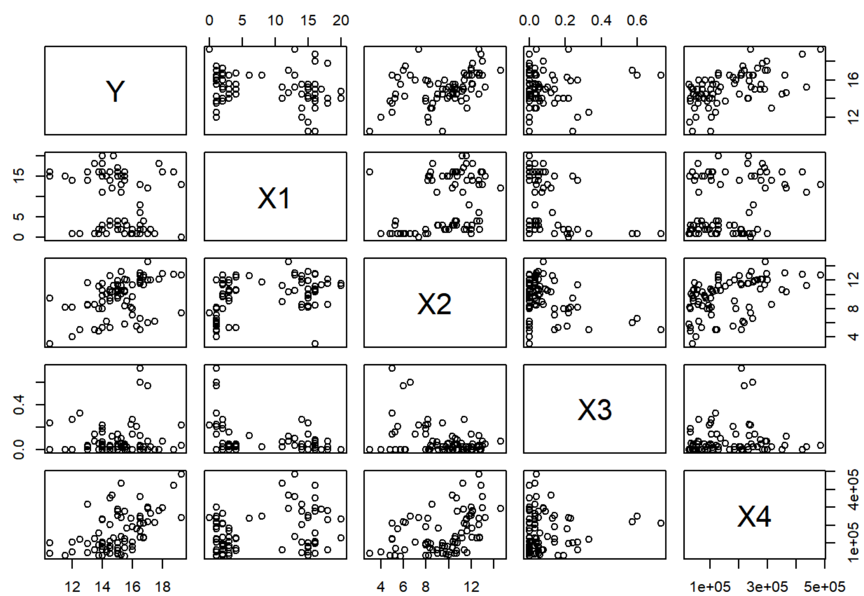
```
stem(commercial$X4)
```

```
##
##   The decimal point is 5 digit(s) to the right of the |
##
##   0 | 333333444444
##   0 | 555666667778899
##   1 | 000001111222333334
##   1 | 578889
##   2 | 011122334444
##   2 | 555788899
##   3 | 002
##   3 | 567
##   4 | 23
##   4 | 8
```

각 설명변수들의 분포를 알 수 있다.

(b)

```
pairs(commercial)
```




```
cor(commercial)
```

```
##              Y          X1         X2          X3         X4
## Y   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## X1 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## X2  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## X3  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## X4  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

X2와 X4가 Y와 상대적으로 뚜렷한 양의 선형관계를 가지고 있다. 이는 correlation에서도 확인할 수 있다.

(c)

```
lm.1 <- lm(Y~., data = commercial)
summary(lm.1)
```

```
##
## Call:
## lm(formula = Y ~ ., data = commercial)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

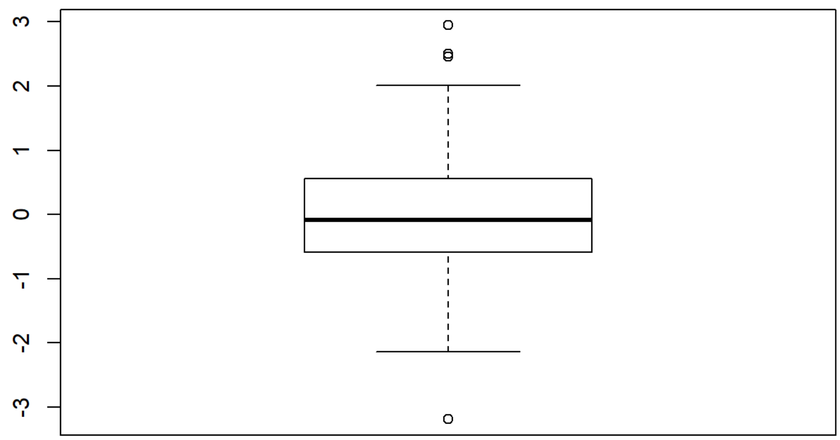
$Y\hat{} = 12.2006 - 0.1420X1 + 0.2820X2 + 0.6193X3 + 0.0000079X4$

(d)

```
lm.1$residuals
```

```
##            1            2            3            4            5
## -1.035672440 -1.513806414 -0.591053402 -0.133568082  0.313283765
##            6            7            8            9           10
## -3.187185224 -0.538356749  0.236302386  1.989220372  0.105829603
##           11           12           13           14           15
##  0.023124830 -0.337070751  0.717869468 -0.392411015 -0.201019573
##           16           17           18           19           20
## -0.814937024  0.101690072 -1.759131637 -1.210114916 -0.634341765
##           21           22           23           24           25
## -0.366004170  0.288596123 -0.093200248  0.233884284 -0.853339941
##           26           27           28           29           30
## -2.123934469  0.466014057 -0.573974675 -1.068826727 -0.197717691
##           31           32           33           34           35
## -1.121737177 -0.173906919 -1.030125636 -0.090953654  0.215053952
##           36           37           38           39           40
##  0.784804746  1.083920373 -2.132451269 -0.185470952 -1.120385453
##           41           42           43           44           45
## -0.012771680  2.500938643 -1.582833452  0.929599530  0.394236721
##           46           47           48           49           50
##  0.117200255  0.815339787  1.605896564  0.557941960  0.494737472
##           51           52           53           54           55
##  0.207611404 -0.032045798  1.155796537  0.234272601 -1.073489739
##           56           57           58           59           60
##  1.059646672 -0.261711555  1.031651273 -0.345957207  0.203372872
##           61           62           63           64           65
##  0.917961126  2.944144932  2.459696482  1.859088749  1.451807658
##           66           67           68           69           70
## -0.483857748 -0.756250356  2.011402309  0.078550427  0.009892809
##           71           72           73           74           75
##  1.766898426 -0.463930876 -0.510410866 -0.106354746  1.209427169
##           76           77           78           79           80
## -0.261085606 -0.627547725  0.910085787 -0.550846871 -2.030180944
##           81
## -0.906819056
```

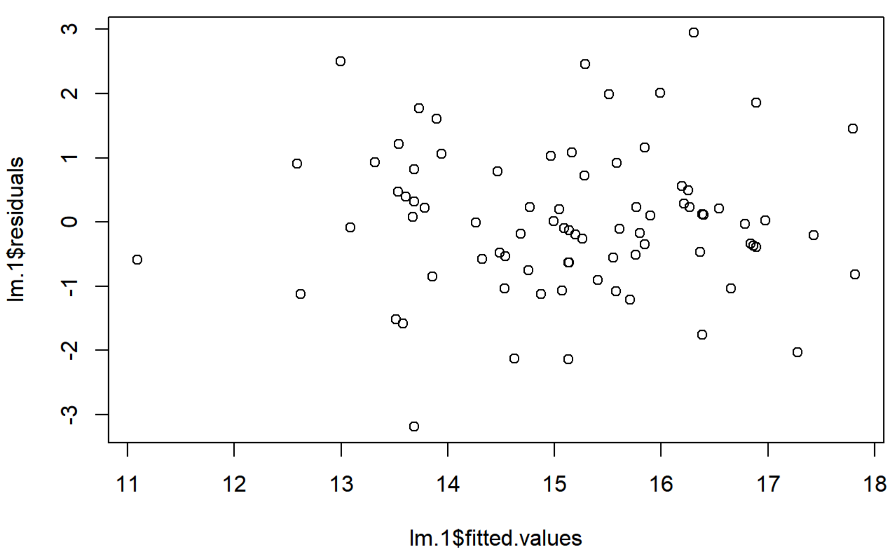
```
boxplot(lm.1$residuals)
```



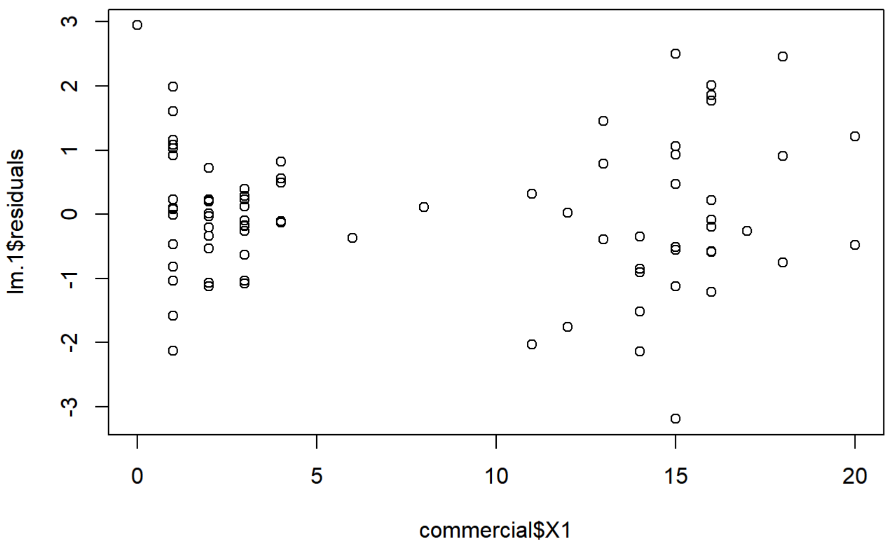
오차항은 0을 기준으로 거의 대칭적으로 분포해있다.

(e)

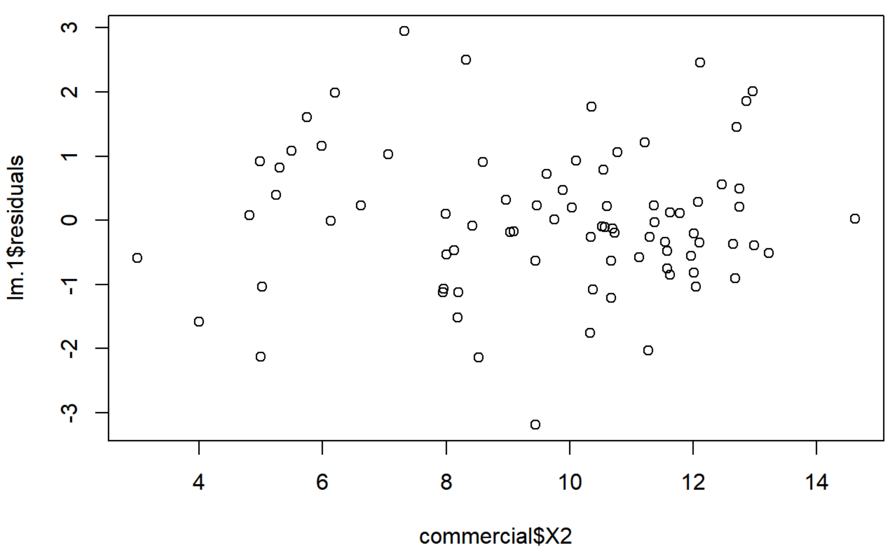
```
plot(lm.1$fitted.values, lm.1$residuals)
```



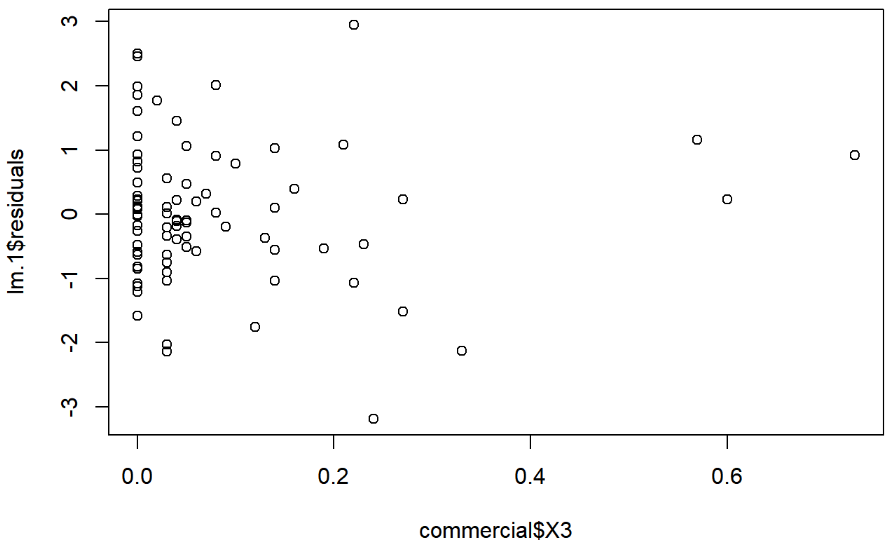

```
plot(commercial$X1, lm.1$residuals)
```



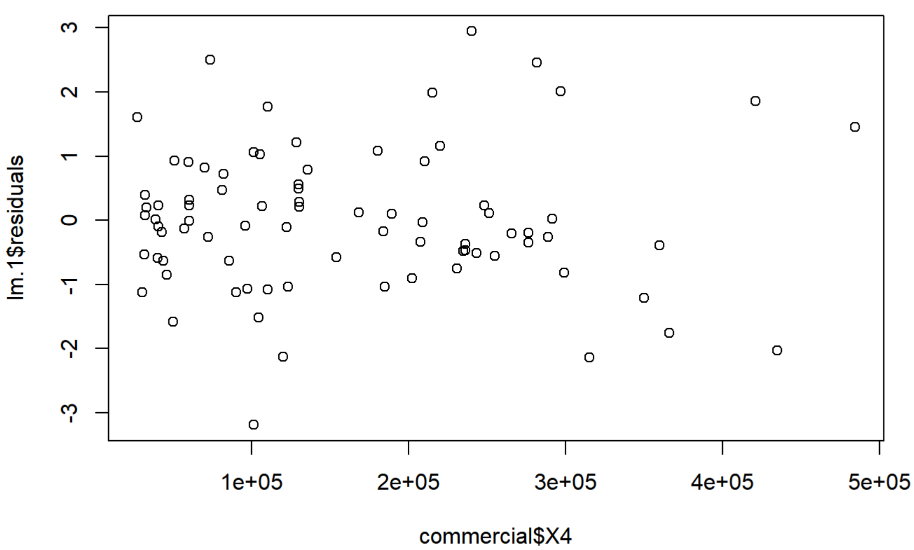

```
plot(commercial$X2, lm.1$residuals)
```



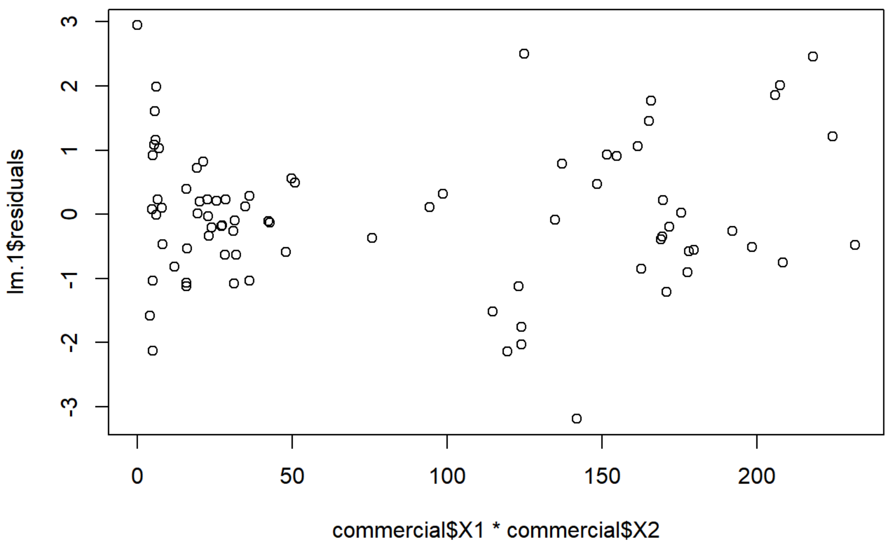

```
plot(commercial$X3, lm.1$residuals)
```



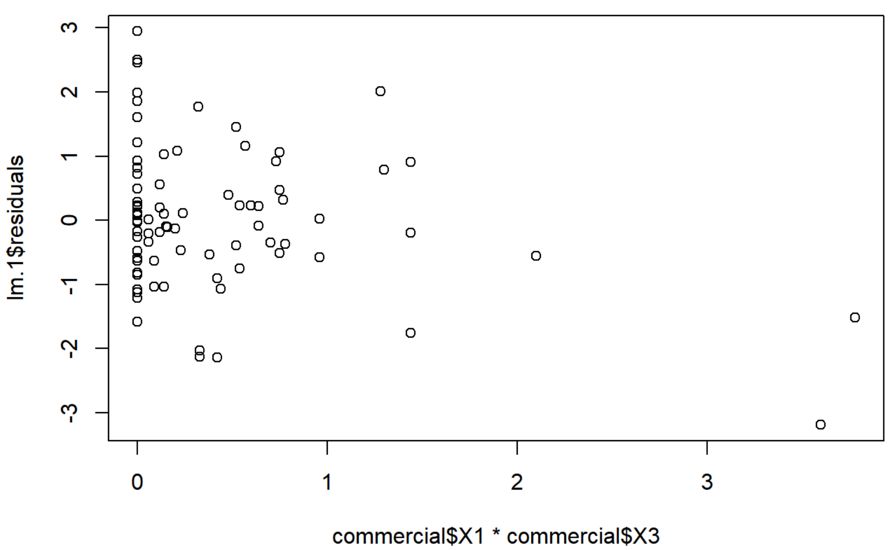

```
plot(commercial$X4, lm.1$residuals)
```



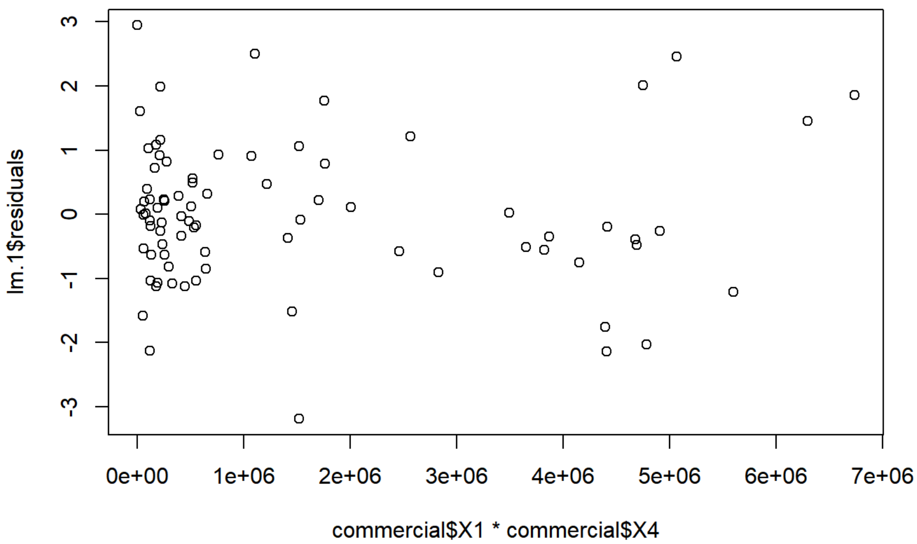

```
plot(commercial$X1*commercial$X2, lm.1$residuals)
```



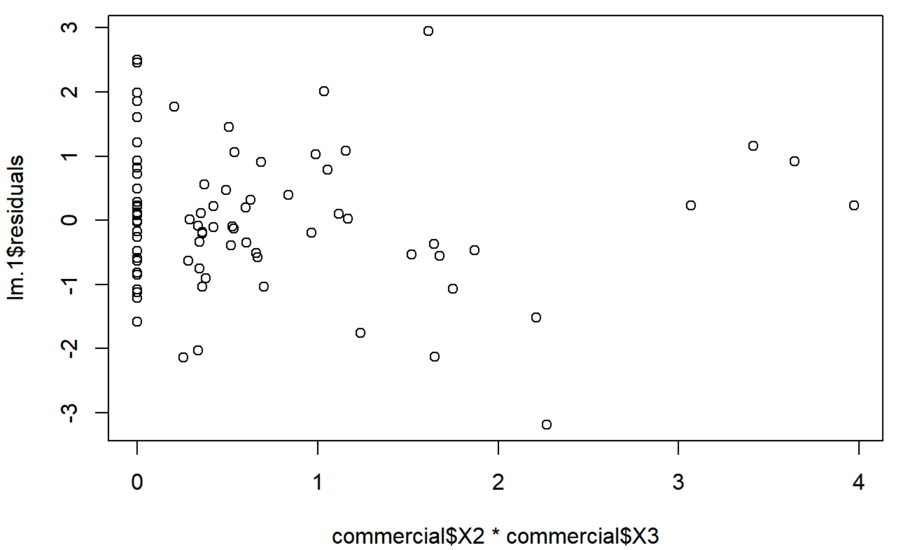

```
plot(commercial$X1*commercial$X3, lm.1$residuals)
```



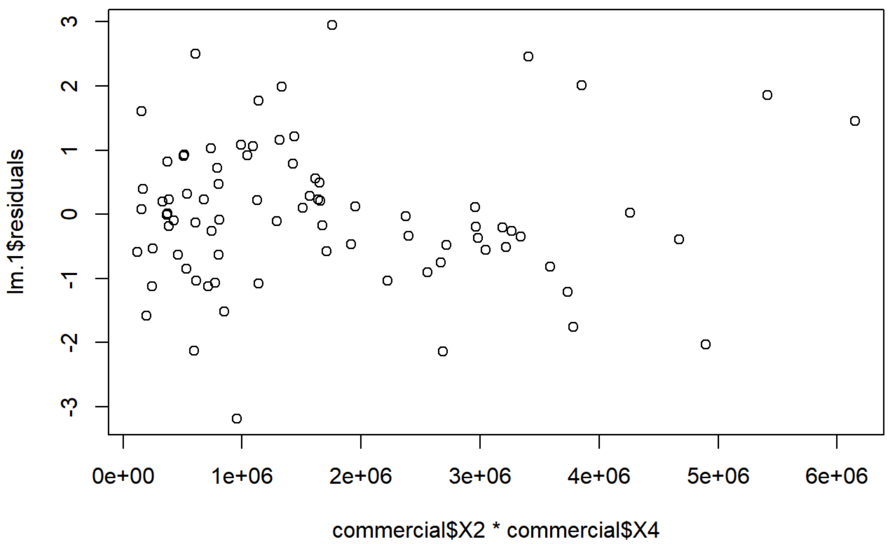

```
plot(commercial$X1*commercial$X4, lm.1$residuals)
```



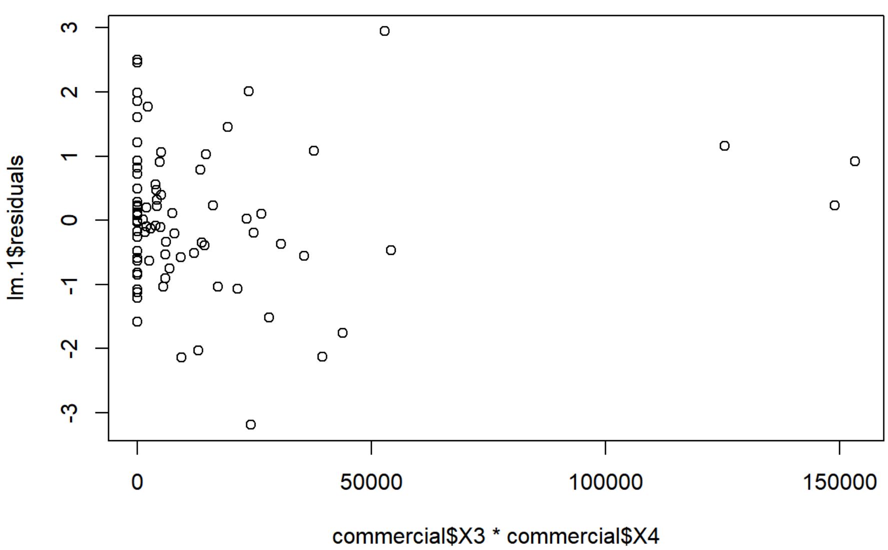

```
plot(commercial$X2*commercial$X3, lm.1$residuals)
```



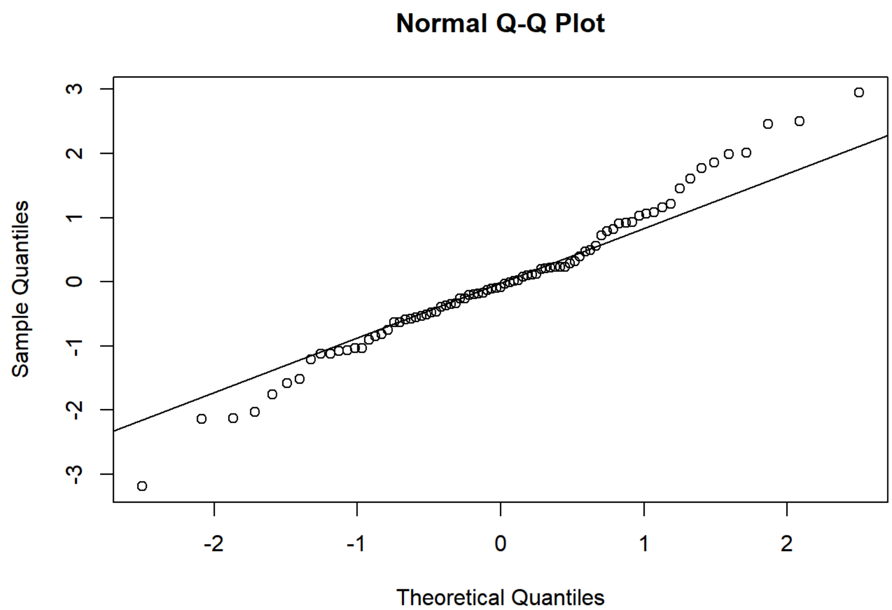

```
plot(commercial$X2*commercial$X4, lm.1$residuals)
```




```
plot(commercial$X3*commercial$X4, lm.1$residuals)
```




```
qqnorm(lm.1$residuals)
qqline(lm.1$residuals)
```



오차항이 대부분 랜덤하게 분포되어 있으며, 꼬리를 제외하곤 정규성을 잘 따른다.

(f) 각 설명변수들이 반복되지 않기 때문에 lack of fit test를 진행할 수 없다.

(g)

```
resid <- lm.1$residuals
r0_index <- lm.1$fitted.values<median(lm.1$fitted.values)
r1_index <- lm.1$fitted.values>=median(lm.1$fitted.values)
abs.r0 <- abs(resid[r0_index] - median(resid[r0_index]))
abs.r1 <- abs(resid[r1_index] - median(resid[r1_index]))
abs.r <- c(abs.r0, abs.r1)
t.test(abs.r0, abs.r1, var.eqaul=TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  abs.r0 and abs.r1
## t = 0.55235, df = 78.988, p-value = 0.5823
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2350075  0.4155329
## sample estimates:
## mean of x mean of y
## 0.8695662 0.7793035
```

```
qt(0.975, 79)
```

```
## [1] 1.99045
```

H0 : error variance constant vs. H1 : not H0

t = 0.55235 < 1.99045 이므로 귀무가설을 기각하지 못하며, 등분산이라 할 수 있다.

2.

(a)

```
lm.2 <- lm(Y~X1+X4, data=commercial)
summary(lm.2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4, data = commercial)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2032 -0.4593  0.0641  0.7730  2.5083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.436e+01  2.771e-01  51.831  < 2e-16 ***
## X1          -1.145e-01  2.242e-02  -5.105 2.27e-06 ***
## X4           1.045e-05  1.363e-06   7.663 4.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 78 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4515
## F-statistic: 33.93 on 2 and 78 DF,  p-value: 2.506e-11
```

$Y\hat{} = 14.3613 - 0.1145X1 + 0.0000104X4$

(b)

```
lm.1$coefficients
```

```
##   (Intercept)            X1            X2            X3            X4
##  1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
```

```
lm.2$coefficients
```

```
##   (Intercept)            X1            X4
##  1.436128e+01 -1.144670e-01  1.044493e-05
```

X1과 X4의 계수 및 상수항은 두 식에서 각각 비슷한 값을 가진다. X1은 그 절댓값이 살짝 감소, X4는 살짝 증가하였다.

(c)

```
lm.c.34 <- lm(Y~X3+X4, data = commercial)
anova(lm.c.34)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X3         1   1.047   1.047  0.4842    0.4886
## X4         1  66.858  66.858 30.9213 3.626e-07 ***
## Residuals 78 168.652   2.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm.c.4 <- lm(Y~X4, data = commercial)
anova(lm.c.4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR(X4) = 67.775, SSR(X4|X3) = 66.858로 서로 다르다.

```
lm.c.31 <- lm(Y~X3+X1, data = commercial)
anova(lm.c.31)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## X3         1   1.047  1.0469  0.3683 0.54570
## X1         1  13.774 13.7743  4.8454 0.03068 *
## Residuals 78 221.736  2.8428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm.c.1 <- lm(Y~X1, data = commercial)
anova(lm.c.1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## X1         1  14.819 14.8185  5.2795 0.02422 *
## Residuals 79 221.739  2.8068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR(X1) = 14.819, SSR(X1|X3) = 13.774로 서로 다르다.

(d) 각 변수 간에 correltaion이 존재하기 때문에 다른 변수의 유무에 계수의 크기나 SSR이 영향받음을 알 수 있다.

3.

```
senic <- read.table("SENIC.txt")
head(senic)
```

```
##   V1    V2   V3  V4   V5    V6  V7 V8 V9 V10 V11 V12
## 1  1  7.13 55.7 4.1  9.0  39.6 279  2  4 207 241  60
## 2  2  8.82 58.2 1.6  3.8  51.7  80  2  2  51  52  40
## 3  3  8.34 56.9 2.7  8.1  74.0 107  2  3  82  54  20
## 4  4  8.95 53.7 5.6 18.9 122.8 147  2  4  53 148  40
## 5  5 11.20 56.5 5.7 34.5  88.9 180  2  1 134 151  40
## 6  6  9.76 50.9 5.1 21.9  97.0 150  2  2 147 106  40
```

```
names(senic) <- c("V1","X1","X2","Y","V5","X3","V7","X4","V9","V10","V11","V12")
```

(a)

```
senic$X4[senic$X4==2] <- 0
table(senic$X4)
```

```
##
##  0  1
## 96 17
```

```
lm.senic <- lm(Y~X1+X2+X3+X4, data = senic)
summary(lm.senic)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = senic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74669 -0.76646 -0.00283  0.77267  2.59703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.85738    1.32434   0.647  0.51874
## X1           0.28882    0.06291   4.591  1.2e-05 ***
## X2          -0.01805    0.02411  -0.749  0.45569
## X3           0.01995    0.00577   3.458  0.00078 ***
## X4           0.28782    0.30668   0.938  0.35009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 108 degrees of freedom
## Multiple R-squared:  0.3681, Adjusted R-squared:  0.3447
## F-statistic: 15.73 on 4 and 108 DF,  p-value: 3.574e-10
```

(b)

```
c(0.28782-qt(1-0.02/2, 108)*0.30668, 0.28782+qt(1-0.02/2, 108)*0.30668)
```

```
## [1] -0.4363656  1.0120056
```

Y에 대한 X4의 98% 신뢰구간은 [-0.4364, 1.012]이다. 이 구간은 0을 포함하기 때문에 X4가 어떤 영향을 끼친다고 확실히 말하기는 어렵다.

(c)

```
lm.senic.c <- lm(Y~X1+X2+X3+X4+X2*X4+X3*X4, data=senic)
summary(lm.senic.c)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X2 * X4 + X3 * X4, data = senic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75072 -0.70321 -0.07468  0.76468  2.60903
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.994129   1.394456   0.713 0.477465
## X1           0.264139   0.063369   4.168 6.29e-05 ***
## X2          -0.022829   0.024699  -0.924 0.357435
## X3           0.024289   0.006478   3.749 0.000289 ***
## X4          -5.695202   4.600959  -1.238 0.218514
## X2:X4        0.155756   0.092677   1.681 0.095778 .
## X3:X4       -0.024059   0.013893  -1.732 0.086234 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 106 degrees of freedom
## Multiple R-squared:  0.3939, Adjusted R-squared:  0.3596
## F-statistic: 11.48 on 6 and 106 DF,  p-value: 7.251e-10
```

```
anova(lm.senic, lm.senic.c)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X3 + X4
## Model 2: Y ~ X1 + X2 + X3 + X4 + X2 * X4 + X3 * X4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    108 127.24
## 2    106 122.05  2    5.1964 2.2566 0.1097
```

```
qf(1-0.1,2,nrow(senic)-7)
```

```
## [1] 2.353335
```

$H0 : \beta5 = \beta6 = 0 vs. H1 : not H0$

F = 2.2566 < 2.3533 으로 귀무가설을 기각하지 못한다. 즉, 교호작용들의 영향은 유의하지 않다. p값 또한 0.1 이상이다.

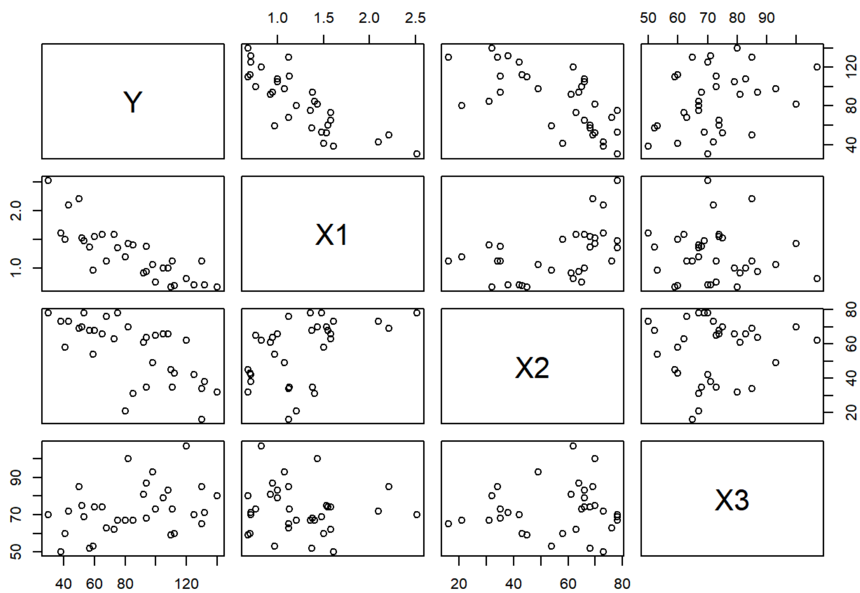
4.

```
kidney <- read.table("Kidney function.txt")
names(kidney) <- c("Y","X1","X2","X3")
head(kidney)
```

```
##     Y   X1 X2  X3
## 1 132 0.71 38  71
## 2  53 1.48 78  69
## 3  50 2.21 69  85
## 4  82 1.43 70 100
## 5 110 0.68 45  59
## 6 100 0.76 65  73
```

(a)

```
pairs(kidney)
```




```
cor(kidney)
```

```
##             Y          X1          X2          X3
## Y   1.0000000 -0.80181086 -0.66787239  0.34591487
## X1 -0.8018109  1.00000000  0.46773179 -0.08898262
## X2 -0.6678724  0.46773179  1.00000000  0.06848147
## X3  0.3459149 -0.08898262  0.06848147  1.00000000
```

Y와 설명변수들 간의 관계를 살펴보면, X1과 X2는 음의 상관관계, X3와는 상대적으로 약한 양의 상관관계를 보인다. X1과 X2가 어느정도 양의 상관관계를 가지고 있어, 다중 공선성이 의심된다.

(b)

```
lm.4 <- lm(Y~., data=kidney)
summary(lm.4)
```

```
##
## Call:
## lm(formula = Y ~ ., data = kidney)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.668  -7.002   1.518   9.905  16.006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
## X1          -39.9393     5.6000  -7.132 7.55e-08 ***
## X2           -0.7368     0.1414  -5.211 1.41e-05 ***
## X3            0.7764     0.1719   4.517 9.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.46 on 29 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8398
## F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12
```

모든 변수가 0.001 수준에서 유의하게 나타나므로 의미가 있다.

(c)

```
kidney$newX1 <- kidney$X1 - mean(kidney$X1)
kidney$newX2 <- kidney$X2 - mean(kidney$X2)
kidney$newX3 <- kidney$X3 - mean(kidney$X3)

library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

```
reg.all <- regsubsets(Y~newX1+newX2+newX3+I(newX1^2)+I(newX2^2)+I(newX3^2)
                      +newX1*newX2+newX1*newX3+newX2*newX3, data=kidney)
summary.reg.all <- summary(reg.all)
best3 <- order(summary.reg.all$cp)[1:3]
cbind(summary.reg.all$which[best3,], cp = summary.reg.all$cp[best3])
```

```
##   (Intercept) newX1 newX2 newX3 I(newX1^2) I(newX2^2) I(newX3^2)
## 4           1     1     1     1          0          0          0
## 5           1     1     1     1          0          0          1
## 6           1     1     1     1          0          1          1
##   newX1:newX2 newX1:newX3 newX2:newX3       cp
## 4           1           0           0 3.302215
## 5           1           0           0 3.384990
## 6           1           0           0 4.766392
```

(d) 첫번째 모델과 두번째 모델의 Cp는 거의 차이가 없다. 두번째와 세번째도 크게 차이나지는 않지만 첫번째와 두번째의 차이에 비해 크다.

5.

(a)

Hard hat : $E(Y) = \beta 0 + \beta 1X1 + \beta 3$

Bump cap : $E(Y) = \beta 0 + \beta 1X1 + \beta 2$

None : $E(Y) = \beta 0 + \beta 1X1$

(b)

(1) $H0 : β2<=0 vs. H1 : β2>0 $

(2) $H0 : β3>=0 vs. H1 : β3<0 $

6.

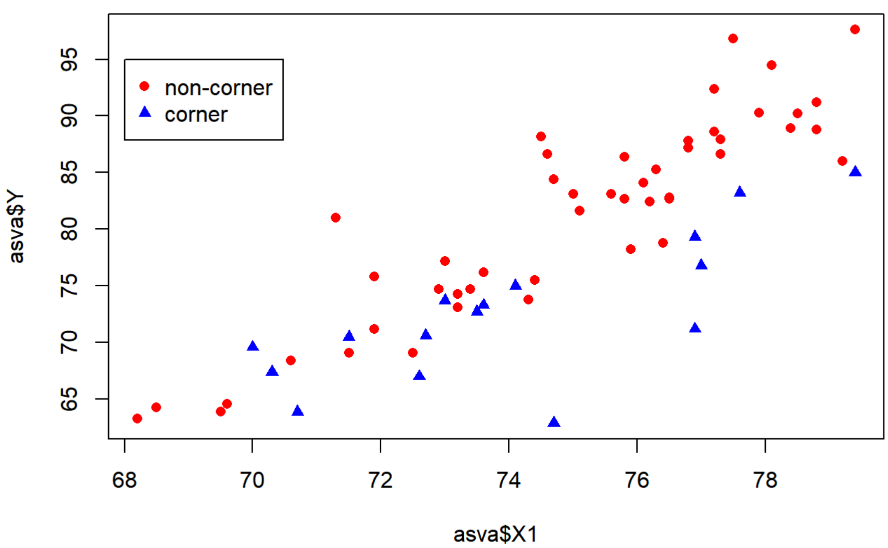
```
asva <- read.table("Assessed valuations.txt")
head(asva)
```

```
##     V1   V2 V3
## 1 78.8 76.4  0
## 2 73.8 74.3  0
## 3 64.6 69.6  0
## 4 76.2 73.6  0
## 5 87.2 76.8  0
## 6 70.6 72.7  1
```

```
names(asva) <- c("Y","X1","X2")
```

(a)

```
plot(asva$X1, asva$Y, col = ifelse(asva$X2==0,'red','blue'), pch = c(16, 17)[as.factor(asva$X2)])
legend(68, 95,
       legend = c("non-corner", "corner"),
       col = c("red", "blue"),
       pch = c(16, 17))
```



non-corner 그룹의 경우 더 가파른 회귀 직선이 적합될 것으로 보인다.

(b)

```
asva$X2 <- as.factor(asva$X2)
lm.6 <- lm(Y~X1+X2+X1*X2, data=asva)
summary(lm.6)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2, data = asva)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8470  -2.1639   0.0913   1.9348   9.9836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.9052    14.7225  -8.620 4.33e-12 ***
## X1             2.7759     0.1963  14.142  < 2e-16 ***
## X21           76.0215    30.1314   2.523  0.01430 *
## X1:X21        -1.1075     0.4055  -2.731  0.00828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.893 on 60 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8145
## F-statistic: 93.21 on 3 and 60 DF,  p-value: < 2.2e-16
```

```
lm.6.2 <- lm(Y~X1, data=asva)
anova(lm.6.2, lm.6)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + X1 * X2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     62 1475.2
## 2     60  909.1  2    566.15 18.683 4.925e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.95,2,60)
```

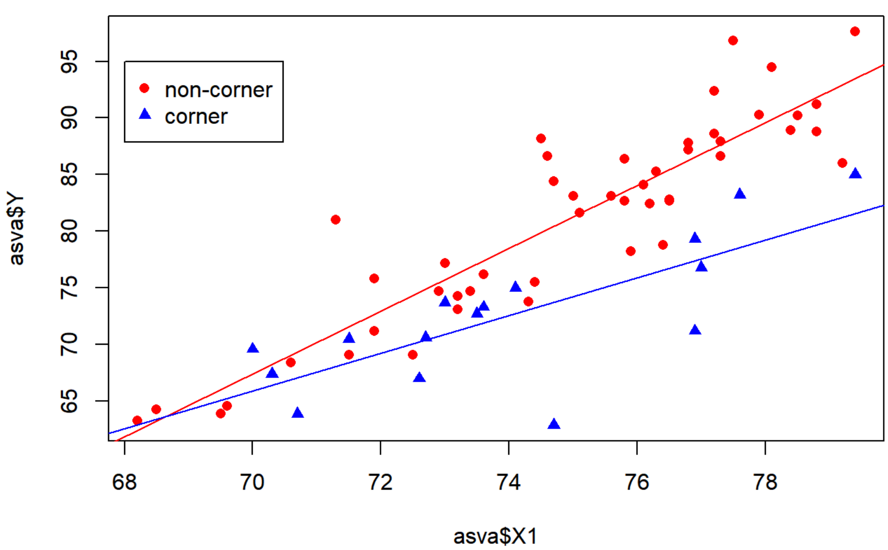
```
## [1] 3.150411
```

$H0 : \beta 2 = \beta 3 = 0 vs. H1 : not H0$

F로 = 18.683 > 3.15 이므로 귀무가설을 기각한다. 즉, 두 회귀식은 같지 않다.

(c)

```
plot(asva$X1, asva$Y, col = ifelse(asva$X2==0,'red','blue'), pch = c(16, 17)[as.factor(asva$X2)])
legend(68, 95,
       legend = c("non-corner", "corner"),
       col = c("red", "blue"),
       pch = c(16, 17))
abline(lm.6$coefficients[1], lm.6$coefficients[2], col = c('red'))
abline(lm.6$coefficients[1]+lm.6$coefficients[3],
       lm.6$coefficients[2]+lm.6$coefficients[4], col = c('blue'))
```



non-corner가 corner보다 더 가파르게 적합되었다.