Homework 9

Luo Bingjun 2017013573 Software 71

2019/5/30

KNNL 8.16

```
n=120
p=2
data0 = read.table(file='CH01PR19.txt', header=F)
data1 = read.table(file='CH08PR16.txt', header=F)
data = cbind(data0, data1)
colnames(data) <- c('Y', 'X1', 'X2')
attach(data)
```

a.

β_0 is the common part of intercept of the regression lines.

β_1 is the common slope of the regression lines.

β_2 shows how much higher (or lower) the mean regression line is for the student who had chosen a major field of concentration than those who hadn't.

b.

```
reg <- lm(Y~X1+X2)
reg
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Coefficients:
## (Intercept)           X1           X2
##     2.19842      0.03789     -0.09430
```

$$Y = 2.19842 + 0.03789X_1 - 0.09430X_2$$

c.

$$H_0 : \beta_2 = 0 \leftrightarrow H_1 : \beta_2 \neq 0$$

Under H_0

$$t = \frac{|\hat{\beta}_2|}{s_2} \sim t_{n-3}$$

```r
alpha=0.01
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70304 -0.35574  0.02541  0.45747  1.25037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19842    0.33886   6.488 2.18e-09 ***
## X1           0.03789    0.01285   2.949  0.00385 **
## X2          -0.09430    0.11997  -0.786  0.43341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6241 on 117 degrees of freedom
## Multiple R-squared:  0.07749,    Adjusted R-squared:  0.06172
## F-statistic: 4.914 on 2 and 117 DF,  p-value: 0.008928
```

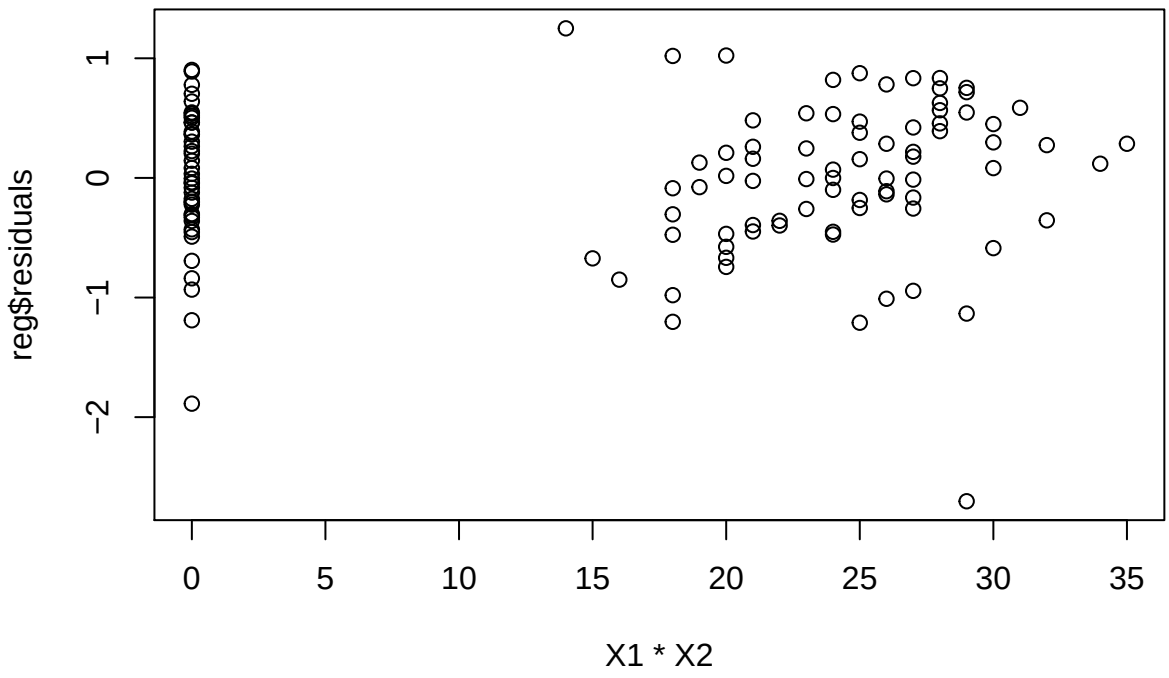
```r
qt(1-alpha/2,n-p-1)
```

```
## [1] 2.618504
```

We accept H_0 as $t = 0.786 < t_{117}(0.995)$, so X_2 cannot be dropped.

d.

```r
plot(X1*X2, reg$residuals)
```



There seems to be a linear tendency when $X_1 \times X_2 \neq 0$, but not obvious enough.

KNNL 8.20

a.

```
reg1 <- lm(Y~X1+X2+X1*X2)
reg1
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2)
##
## Coefficients:
## (Intercept)           X1           X2        X1:X2
##    3.226318    -0.002757    -1.649577     0.062245
```

$$Y = 3.226318 - 0.002757X_1 - 1.649577X_2 + 0.062245X_1X_2$$

b.

$$H_0 : \beta_3 = 0 \leftrightarrow H_1 : \beta_3 \neq 0$$

Under H_0:

$$t = \frac{|\hat{\beta}_3|}{s_3} \sim t_{n-4}$$

```
summary(reg1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80187 -0.31392  0.04451  0.44337  1.47544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.226318   0.549428   5.872 4.18e-08 ***
## X1          -0.002757   0.021405  -0.129   0.8977
## X2          -1.649577   0.672197  -2.454   0.0156 *
## X1:X2        0.062245   0.026487   2.350   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 116 degrees of freedom
## Multiple R-squared:  0.1194, Adjusted R-squared:  0.09664
## F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

```
alpha=0.05
qt(1-alpha/2, n-4)
```

```
## [1] 1.980626
```

```
detach(data)
```

We reject H_0 as $t = 2.350 > t_{116}(0.975)$, so the interaction term cannot be dropped.

KNNL 9.9

```
p=3
data = read.table(file='CH06PR15.txt', header=F)
n=nrow(data)
colnames(data) <- c('Y', 'X1', 'X2', 'X3')
attach(data)
```

a.

```
fullres <-  lm(Y ~ X1 + X2 + X3)$residuals
sigsqhat.big <- sum(fullres^2)/(n-5)
library(leaps)
predictors = data[,c("X1", "X2", "X3")]
response = Y
leapSet = leaps(x=predictors, y=response, nbest = 3)
models = leapSet$which
```

```
colnames(models) = c("X1", "X2", "X3")
m <- nrow(models)
cp <- aic <- bic <- press <- adjrsquared <- rep(0,m)

for( i in 1:m){
  selectVarsIndex = leapSet$which[i,]
  newData <- cbind(response, predictors[, selectVarsIndex])
  newData <- as.data.frame(newData)
  selectedMod <- lm(response ~ ., data=newData)  # build model
  summary(selectedMod)
  adjrsquared[i] <- summary(selectedMod)$adj.r.squared
  aic[i] <- extractAIC(selectedMod)[2] #n*log(sum(fit$res^2)/n)+2*p
  bic[i] <- extractAIC(selectedMod, k = log(n))[2]
  press[i] <- sum((selectedMod$residuals/(1 - hatvalues(selectedMod)))^2)
  cp[i] <- sum(selectedMod$residuals^2)/sigsqhat.big + 2 * selectedMod$rank - n
}

bestModels = cbind(adjrsquared, cp, aic, bic, press, models)
bestModels
```

```
##   adjrsquared        cp      aic      bic    press X1 X2 X3
## 1   0.6103248  7.154711 220.5294 224.1867 5569.562  1  0  0
## 1   0.4022134 33.406461 240.2137 243.8710 8451.432  0  0  1
## 1   0.3490737 40.109649 244.1312 247.7885 9254.489  0  1  0
## 2   0.6610206  1.787985 215.0607 220.5466 4902.751  1  0  1
## 2   0.6389073  4.514027 217.9676 223.4536 5235.192  1  1  0
## 2   0.4437314 28.574507 237.8450 243.3309 8115.912  0  1  1
## 3   0.6594939  3.000000 216.1850 223.4995 5057.886  1  1  1
```

I would recommend $\{X_1, X_3\}$ as best, as is shown in Row 4 in the table.

b.

Yes.

This doesn't always happen because different criterias evaluate the regression based on different factors.

c.

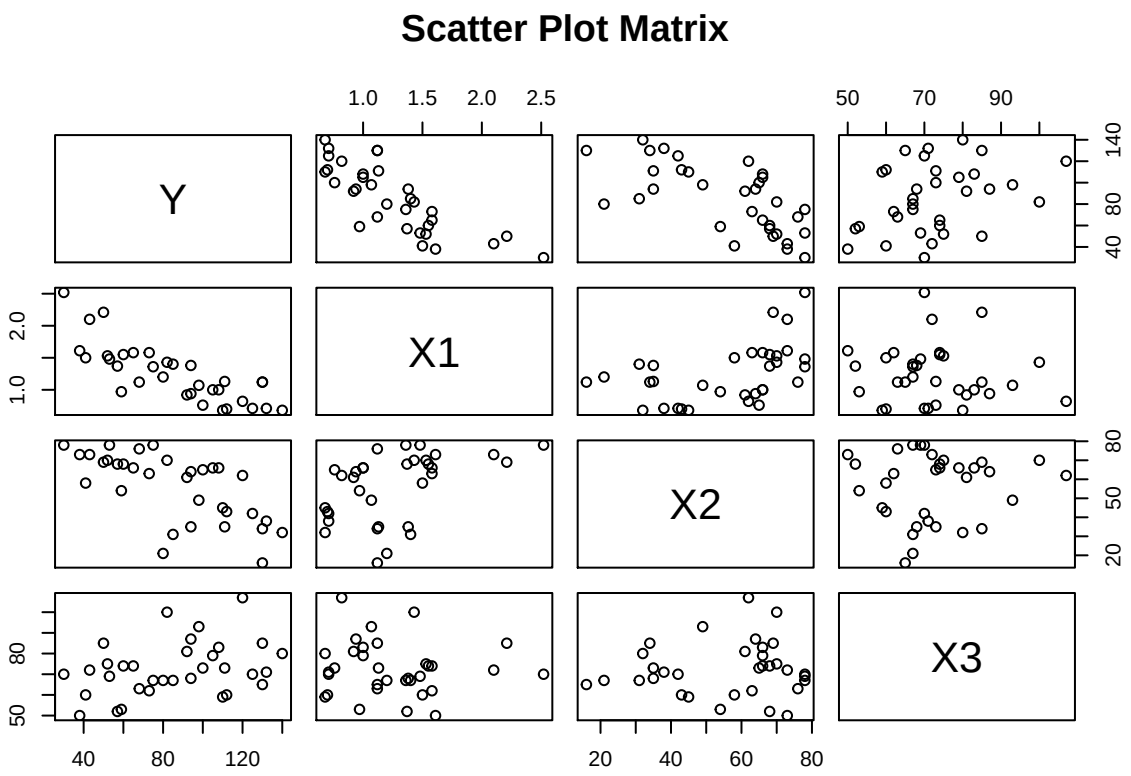
It saves a lot of computation resource, especially when the data is big.

KNNL 9.15

```
detach(data)
p=3
data = read.table(file='CH09PR15.txt', header=F)
n=nrow(data)
colnames(data) <- c('Y', 'X1', 'X2', 'X3')
attach(data)
```

b.

```
pairs(data, main = 'Scatter Plot Matrix')
```




```
cor(data)
```

```
##             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
```

The plots suggest the relationship between Y and each predictor variable seems to be linear.

Not any serious multicollinearity problems are evident.

c.

```
reg<-lm(Y~X1+X2+X3)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## 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
```

$$Y = 120.0473 - 39.9393X_1 - 0.7368X_2 + 0.7764X_3$$

Yes.

KNNL 9.16

a.

```
data1=cbind(data, X1^2, X2^2, X3^2,  X1*X2, X1*X3, X2*X3)
predictors = data1[, 2:10]
response = Y
lp=leaps(x=predictors, y=response, wt=rep(1, NROW(data)), int=TRUE, method=c("Cp","adjr2"), nbest=3)
lp
```

```
## $which
##       1     2     3     4     5     6     7     8     9
## 1 FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 1  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 2  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 3  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## 4  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
## 4  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 4  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE
## 5  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 5  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 5  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## 6  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 6  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
## 6  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 7  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
```

```
## 7  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 8  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 9  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
##  [1] "(Intercept)" "1"           "2"           "3"           "4"
##  [6] "5"           "6"           "7"           "8"           "9"
##
## $size
##  [1]  2  2  2  3  3  3  4  4  4  5  5  5  6  6  6  7  7  7  8  8  8  9  9
## [24]  9 10
##
## $Cp
##  [1] 34.478600 48.509656 65.615252 14.470471 16.646383 26.683707  5.753169
##  [8]  6.512138  8.013698  3.302215  3.674777  3.780059  3.384990  3.828242
## [15]  3.853000  4.587063  4.766392  5.015266  6.344160  6.356750  6.514940
## [22]  8.002543  8.332326  8.332355 10.000000
```

$$\{X_1, X_3, X_1^2, X_2^2, X_3^2, X_2X_3\}$$
$$\{X_1, X_2, X_3, X_2^2, X_3^2, X_1X_2\}$$
$$\{X_1, X_2, X_3, X_3^2, X_1X_2, X_1X_3\}$$

b.

```
lp$Cp[16:18]
```

```
## [1] 4.587063 4.766392 5.015266
```

There is not much difference in C_p.

KNNL 9.19

a.

```
subsets=regsubsets(x=predictors,y=response, nvmax = 4, really.big = T)
summary(subsets)
```

```
## Subset selection object
## 9 Variables  (and intercept)
##         Forced in Forced out
## X1          FALSE      FALSE
## X2          FALSE      FALSE
## X3          FALSE      FALSE
## X1^2        FALSE      FALSE
## X2^2        FALSE      FALSE
## X3^2        FALSE      FALSE
## X1 * X2     FALSE      FALSE
## X1 * X3     FALSE      FALSE
## X2 * X3     FALSE      FALSE
```

```
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          X1  X2  X3  X1^2 X2^2 X3^2 X1 * X2 X1 * X3 X2 * X3
## 1  ( 1 ) " " " " " " " " " "  " "  " "  "*"     " "     " "
## 2  ( 1 ) " " " " "*" " " "  " "  " "  "*"     " "     " "
## 3  ( 1 ) "*" " " "*" " " "  " "  " "  " "     " "     "*"
## 4  ( 1 ) "*" "*" "*" " " "  " "  " "  "*"     " "     " "
```

The best subset of predictor variables is $\{x_1, x_2, x_3, x_1x_2\}$.

b.

```
reg0 <- lm(Y~X1+X2+X3)
reg1 <- lm(Y~X1+X2+X3+(X1*X2))
summary(reg0)$adj.r.squared
```

```
## [1] 0.8397998
```

```
summary(reg1)$adj.r.squared
```

```
## [1] 0.8615103
```

$R^2_{a,p}$ of the best subset here is greater than the one in 9.11a, so we can see that it is better than the subset in 9.11a.

KNNL 10.11

```
detach(data)
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
p=3
data = read.table(file='CH06PR15.txt', header=F)
n=nrow(data)
colnames(data) <- c('Y', 'X1', 'X2', 'X3')
attach(data)
```

a.

```
reg <- lm(Y~X1+X2+X3)
sturesid = rstudent(reg)
round(sturesid,  digits = 4)
```

```
##       1       2       3       4       5       6       7       8       9
##  0.0116 -0.9332  0.4036  0.2147  0.5944 -0.3817 -1.3069  0.0435 -0.7290
##      10      11      12      13      14      15      16      17      18
##  0.2028  1.8358  1.3622 -1.4984 -1.5581  1.3576 -0.2815  1.8067  0.8745
##      19      20      21      22      23      24      25      26      27
## -1.1024  0.8676  0.5732  0.9011  0.3682 -0.4093 -0.4768 -0.4323 -1.9742
##      28      29      30      31      32      33      34      35      36
##  0.5899 -0.9887  0.3473 -1.7694  1.2199  0.0616 -1.5422  0.0959  1.1763
```

```
##      37      38      39      40      41      42      43      44      45
##  1.2278 -0.5494 -0.9870 -0.5898  1.1190 -0.0954 -1.4222  1.3454 -0.5671
##      46
##  1.0449
```

```
outlierTest(reg, cutoff=0.1)
```

```
## No Studentized residuals with Bonferonni p < 0.1
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 27 -1.974202           0.055121           NA
```

Samplpe 27 is most likely to be the outlier, but unadjusted p-value is less than 0.1, so we conclude that there is no outlier.

b.

```
round(hatvalues(reg), digits=4)
```

```
##      1      2      3      4      5      6      7      8      9     10
## 0.0782 0.0671 0.0372 0.1536 0.0967 0.1286 0.0345 0.0752 0.1843 0.0580
##     11     12     13     14     15     16     17     18     19     20
## 0.0876 0.0309 0.0903 0.0332 0.1429 0.0471 0.1195 0.0624 0.0335 0.1289
##     21     22     23     24     25     26     27     28     29     30
## 0.0777 0.1369 0.0329 0.1358 0.0434 0.1029 0.0868 0.1860 0.0594 0.0900
##     31     32     33     34     35     36     37     38     39     40
## 0.1171 0.1096 0.0450 0.0372 0.1030 0.0272 0.1212 0.0706 0.1810 0.0869
##     41     42     43     44     45     46
## 0.0380 0.1539 0.0610 0.0509 0.0726 0.0832
```

There seems to be no outlier.

d.

DFFITS:

```
dffits(reg)[c(11,17,27)]
```

```
##         11         17         27
##  0.5688200  0.6657370 -0.6087397
```

DFBETAS:

```
dfbetas(reg)[c(11,17,27),]
```

```
##    (Intercept)         X1         X2         X3
## 11  0.09910764 -0.3630892 -0.1899887 0.38998516
## 17 -0.44913479 -0.4711109  0.4432302 0.08926996
## 27 -0.01723432  0.4171827 -0.2498614 0.16136484
```

Cook's distance:

```
round(cooks.distance(reg), digits=4)[c(11,17,27)]
```

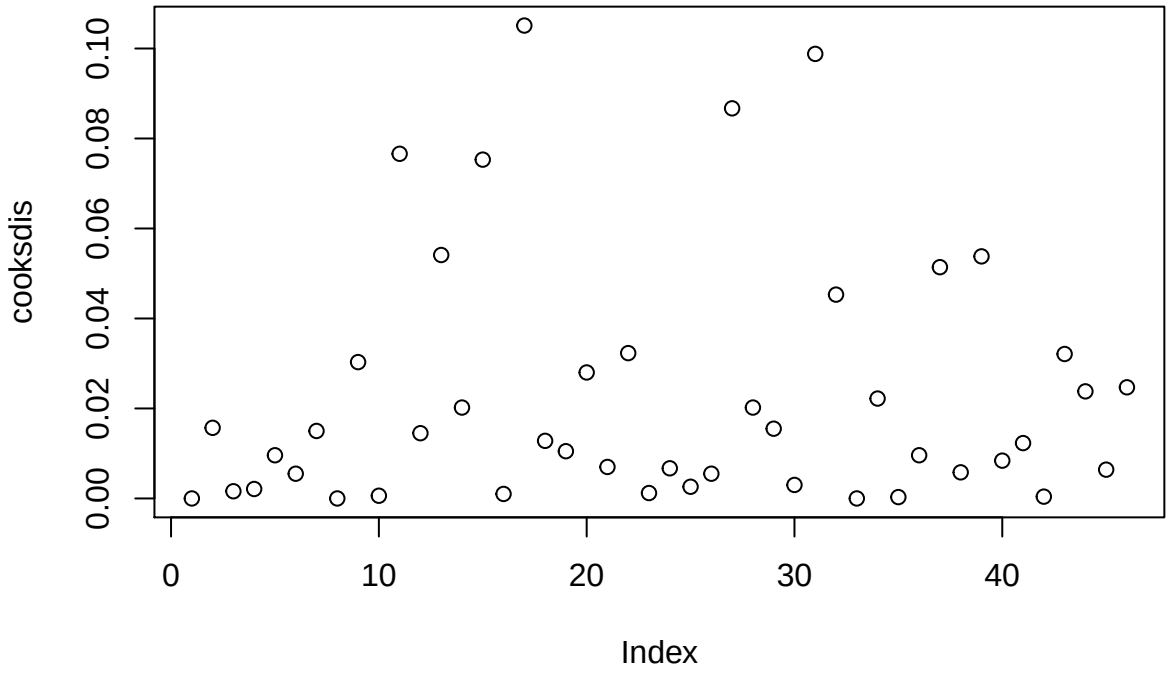
```
##     11     17     27
## 0.0766 0.1051 0.0867
```

Case 17 is most likely to be the outlier.

e.

f.

```
cooksdis=round(cooks.distance(reg), digits=4)
plot(cooksdis)
```

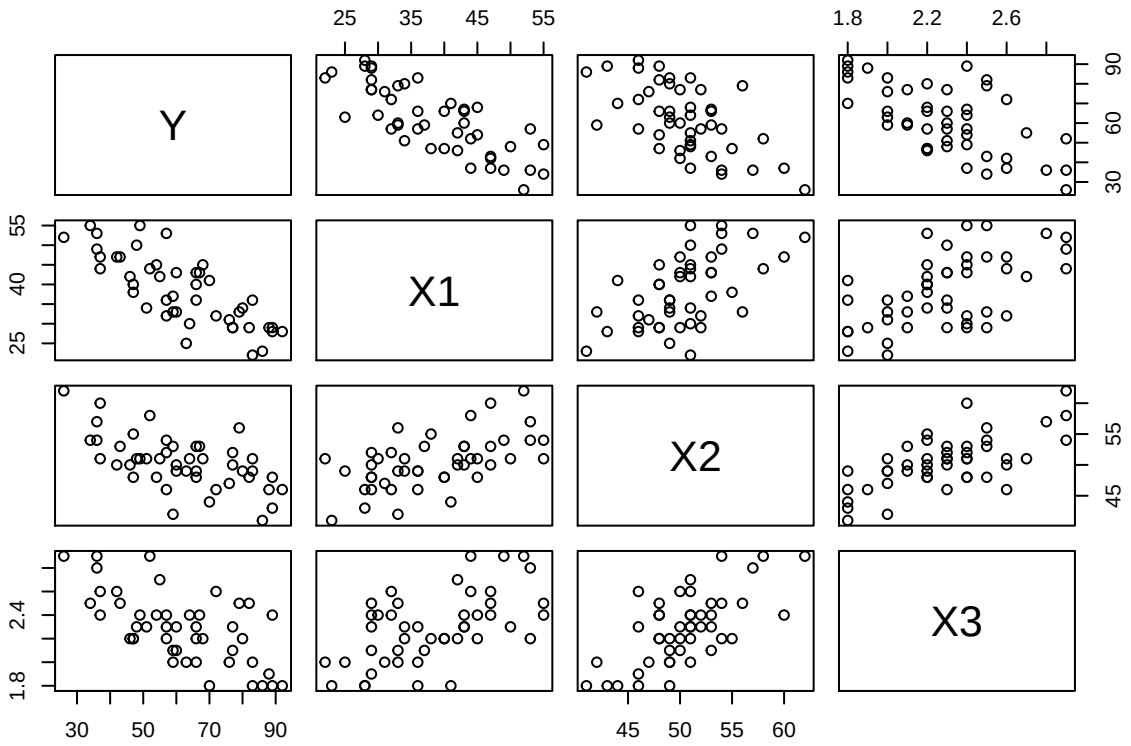


Case 17, 31 and 27 are most influential in this measure.

KNNL 10.17

a.

```
pairs(data)
```




```
cor(data)
```

```
##             Y         X1         X2         X3
## Y   1.0000000 -0.7867555 -0.6029417 -0.6445910
## X1 -0.7867555  1.0000000  0.5679505  0.5696775
## X2 -0.6029417  0.5679505  1.0000000  0.6705287
## X3 -0.6445910  0.5696775  0.6705287  1.0000000
```

There seem to be linear associations among the predictor variables, especially between X_2 and X_3.

b.

```
vif(reg)
```

```
##       X1       X2       X3
## 1.632296 2.003235 2.009062
```

There is excessive multicollinearity among X_1, X_2 and X_3.

Results are more quantitative and revealing.

KNNL 10.21

```
detach(data)
p=3
data = read.table(file='CH09PR15.txt', header=F)
```

```
n=nrow(data)
colnames(data) <- c('Y', 'X1', 'X2', 'X3')
attach(data)
```

a.

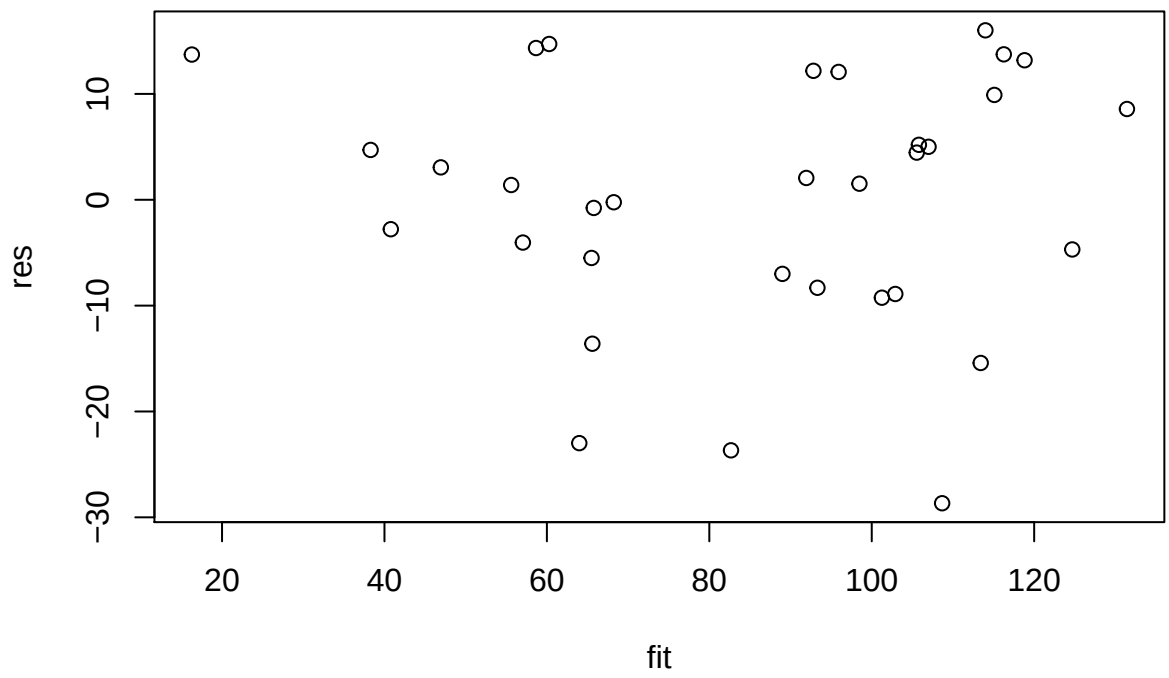
```
reg<-lm(Y~X1+X2+X3)
vif(reg)
```

```
##       X1       X2       X3
## 1.304608 1.300377 1.023997
```

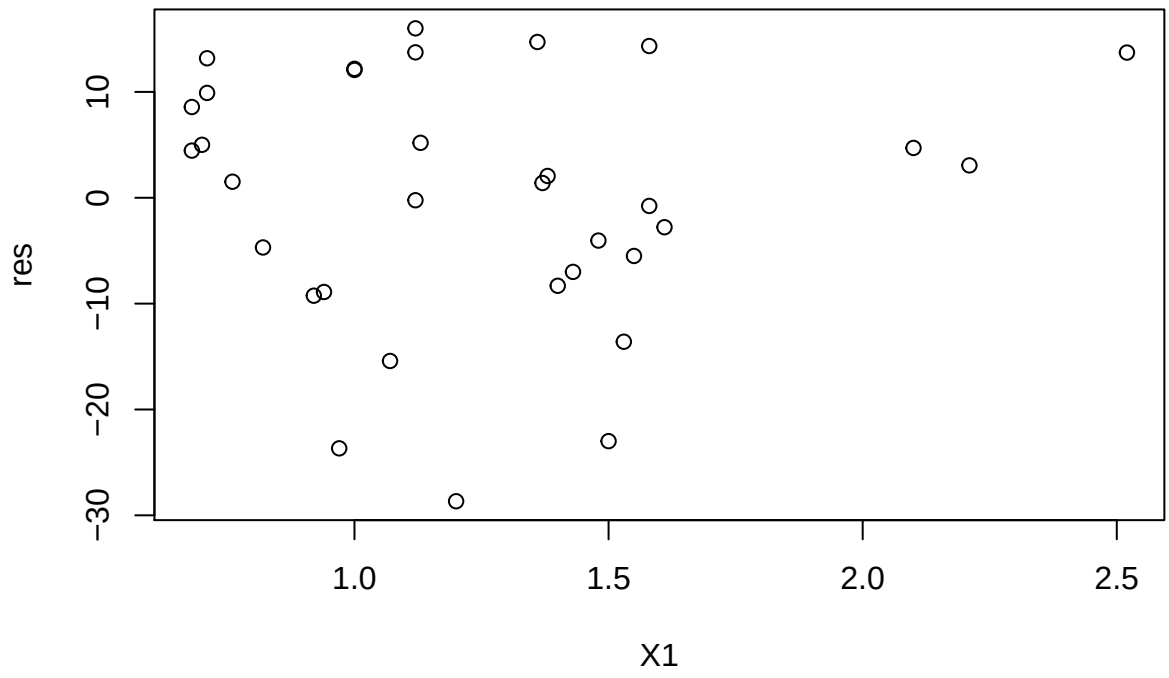
There are serious multicollinearity problems because the average VIF is considerably larger than 1.

b.

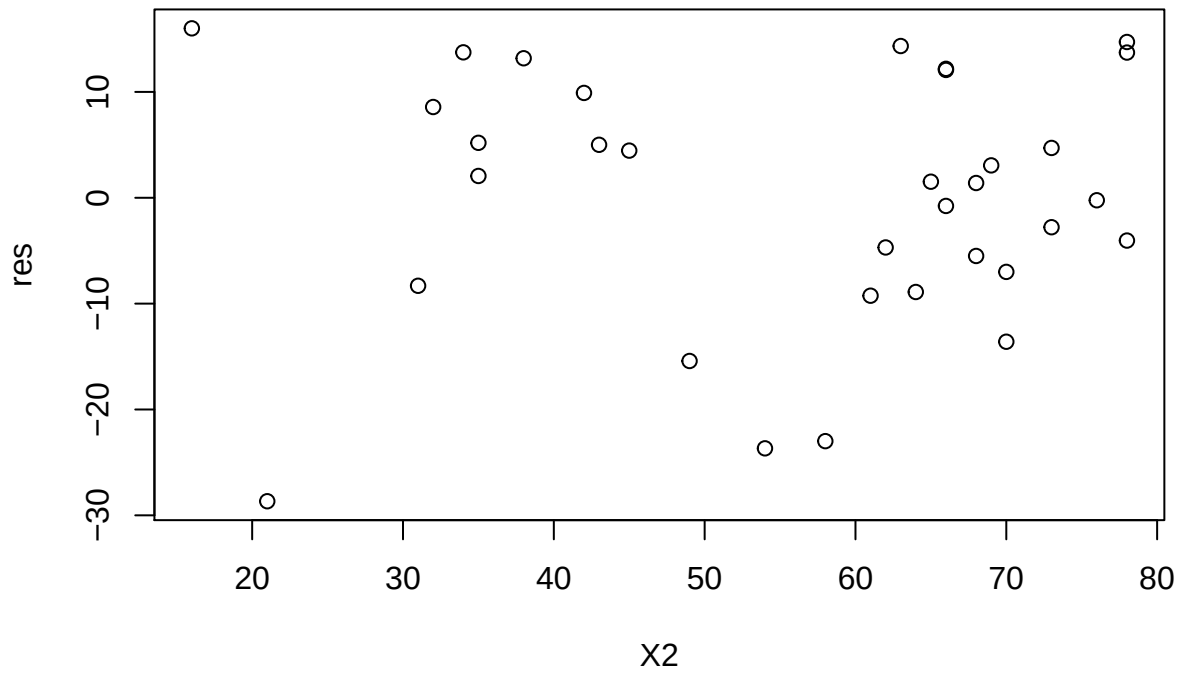
```
res=reg$residuals
fit=reg$fitted.values
plot(fit, res)
```



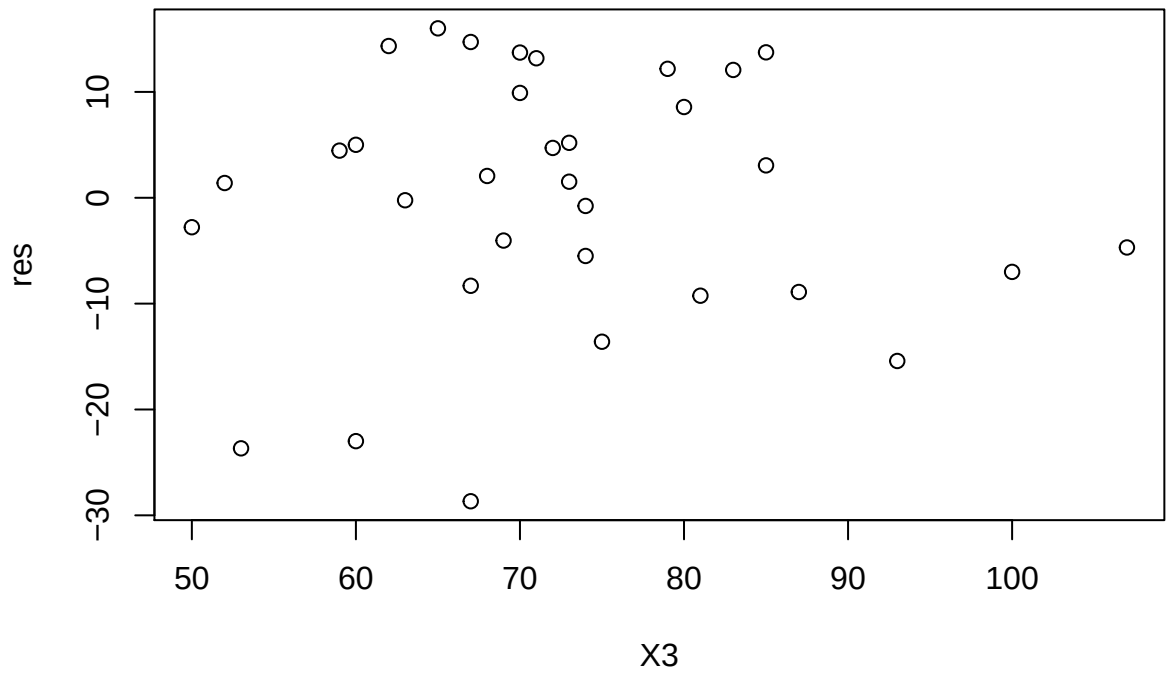

```
plot(X1, res)
```



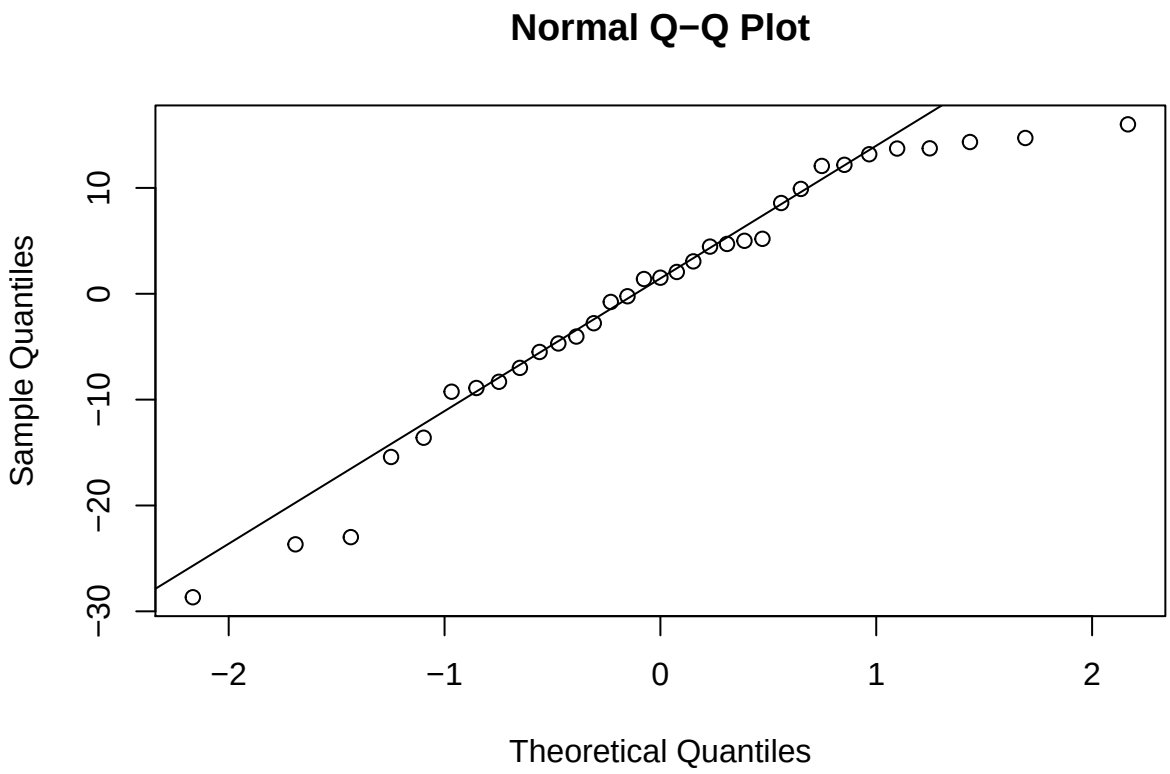

```
plot(X2, res)
```




```
plot(X3, res)
```

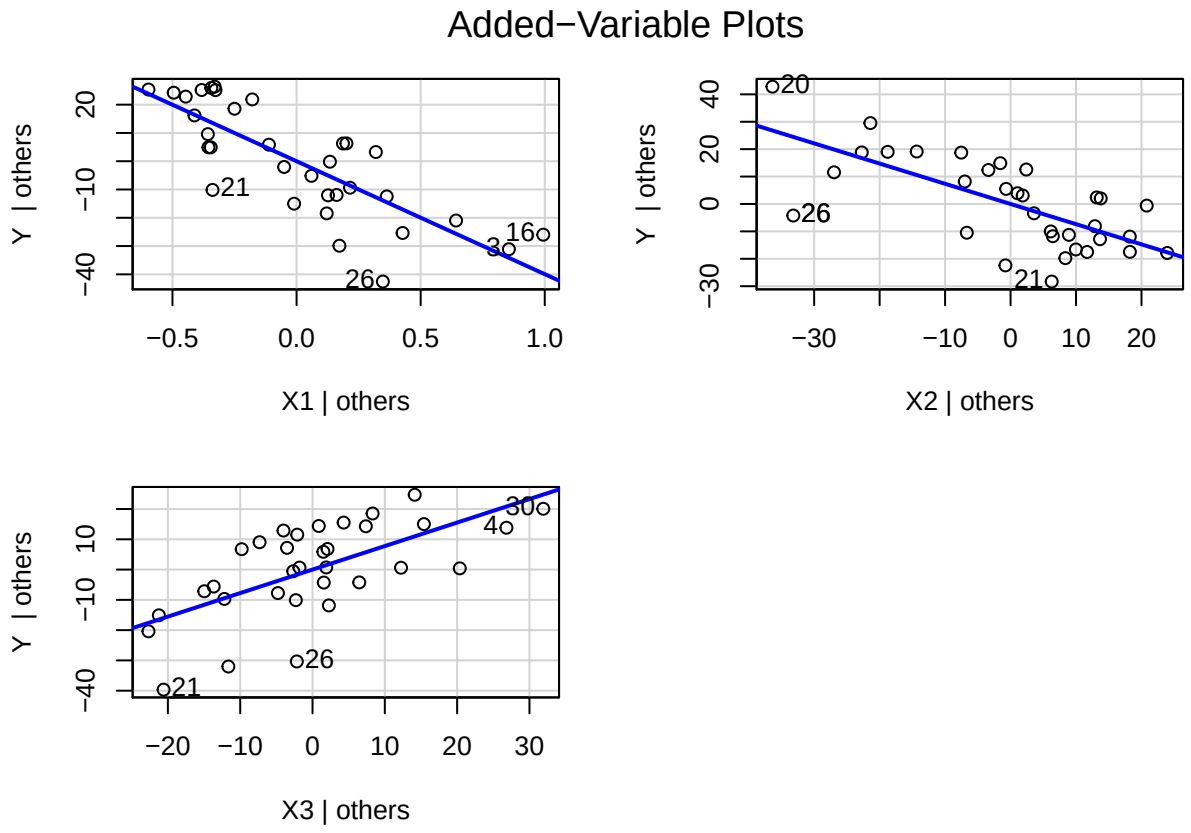



```
qqnorm(res)
qqline(res)
```



c.

```
avPlots(reg)
```



d.

Yes, they suggest that the model should be modified to avoid multicollinearity.

KNNL 10.22

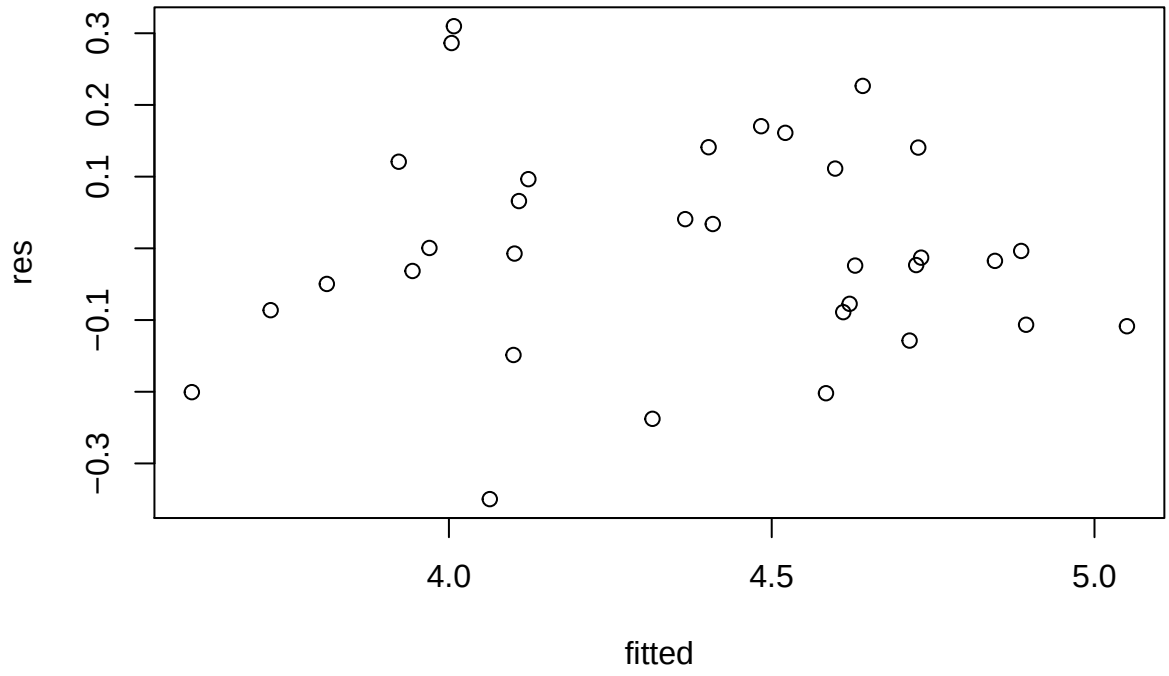
a.

```
reg<-lm(log(Y)~log(X1)+log(140-X2)+log(X3))
reg
```

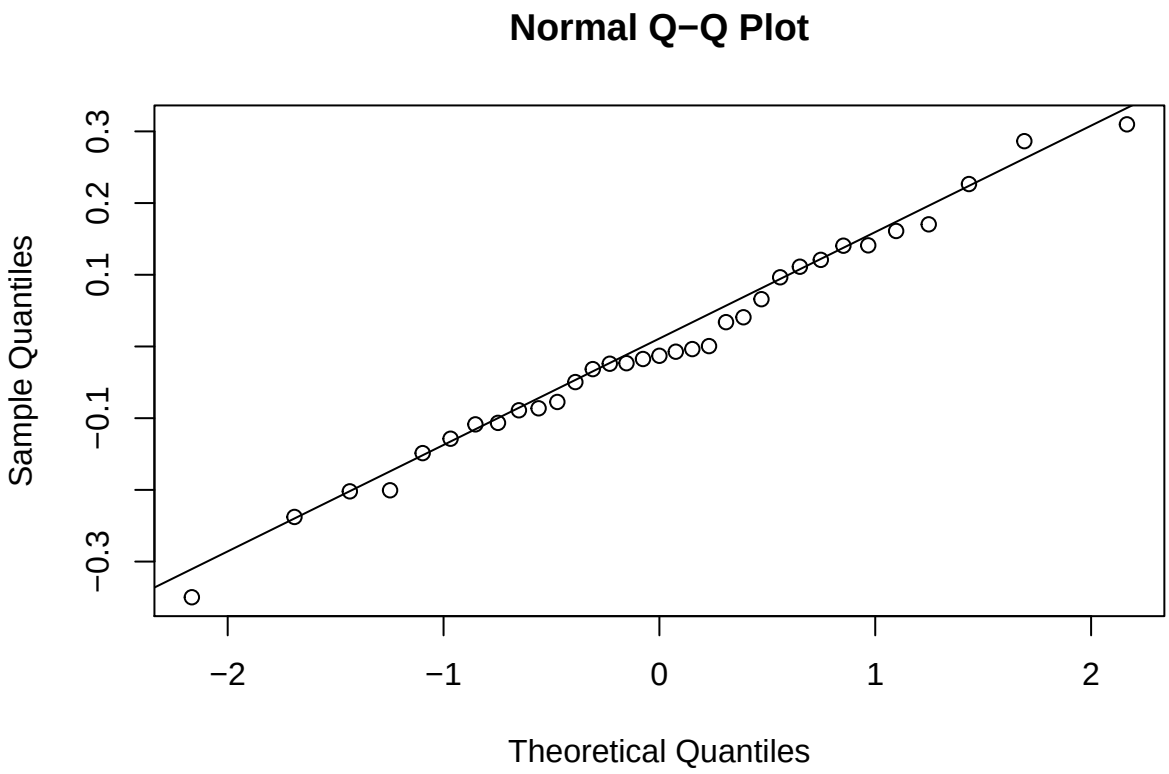
```
##
## Call:
## lm(formula = log(Y) ~ log(X1) + log(140 - X2) + log(X3))
##
## Coefficients:
##   (Intercept)        log(X1)  log(140 - X2)        log(X3)
##       -2.0427         -0.7120         0.7474          0.7574
```

b.

```
res=reg$residuals
fitted=reg$fitted.values
plot(fitted, res)
```




```
qqnorm(res)
qqline(res)
```



c.

```
vif(reg)
```

```
##       log(X1) log(140 - X2)       log(X3)
##      1.339318      1.330109      1.016032
```

Serious multicollearity problems are still here because the average VIF is considerably larger than 1.

d.

```
sturesid = rstudent(reg)
round(sturesid,  digits = 4)
```

```
##       1       2       3       4       5       6       7       8       9
## -0.0240  0.0034 -0.2177  0.2794 -0.1589 -0.1611  0.6500 -0.5807 -0.0468
##      10      11      12      13      14      15      16      17      18
## -0.5113  1.1279 -0.8644 -0.0875 -0.1152  1.0705 -1.4382  0.7346  0.9553
##      19      20      21      22      23      24      25      26      27
##  0.9556  1.6490 -1.6808 -0.6060  0.4257  0.2300 -0.7358 -1.4376 -0.3316
##      28      29      30      31      32      33
##  2.2146 -2.5200 -0.7687 -0.9748  1.9828  0.8290
```

```
outlierTest(reg, cutoff=0.1)
```

```
## No Studentized residuals with Bonferonni p < 0.1
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 29 -2.519951           0.017719      0.58473
```

Case 29 is most likely to be the outlier, but its p-value is still less than 0.1, so we conclude that there is no outlier.