Homework 7

Luo Bingjun 2017013573 Software 71

2019/5/16

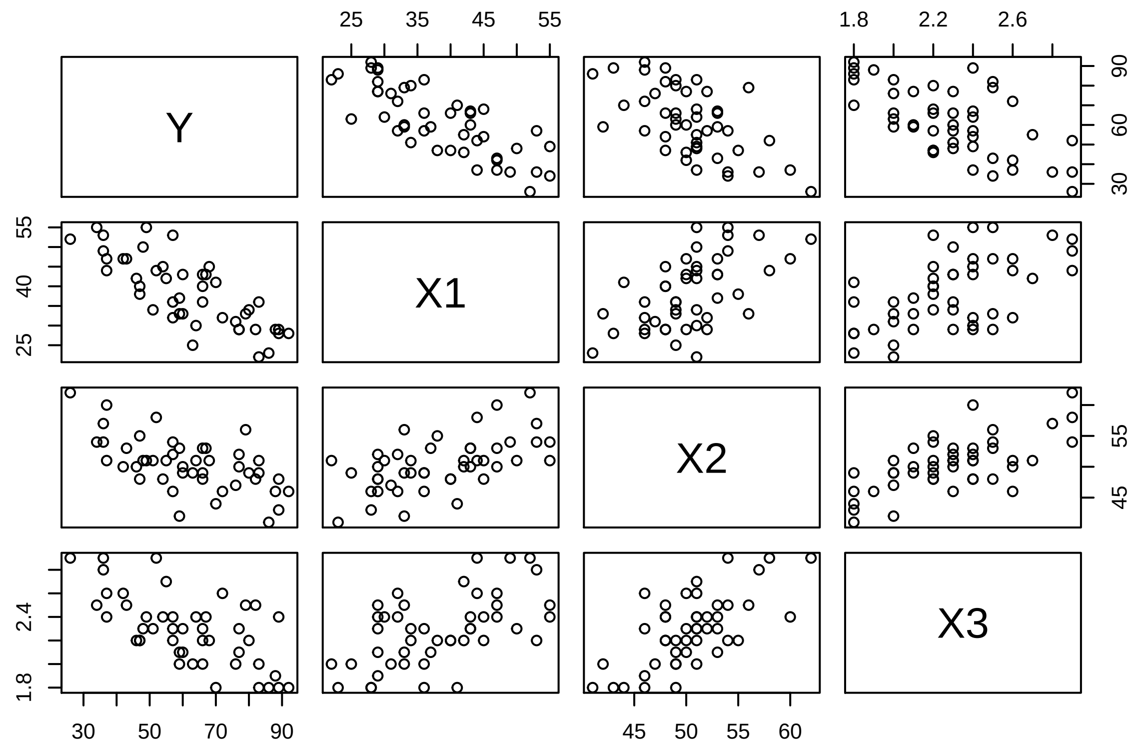
KNNL 6.15

```
n=46
p=4
data = read.table(file='CH06PR15.txt', header=F)
colnames(data) <- c('Y', 'X1', 'X2', 'X3')
attach(data)
```

b.

scatter plot matrix:

```
pairs(data)
```



correlation matrix:

```
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 seems to be negative correlations between Y and X_1,X_2,X_3, but there may also be strong linear relationships among the independent variables, as a potential risk of collinearity.

c.

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

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##     158.491       -1.142       -0.442      -13.470
```

estimated regression function:

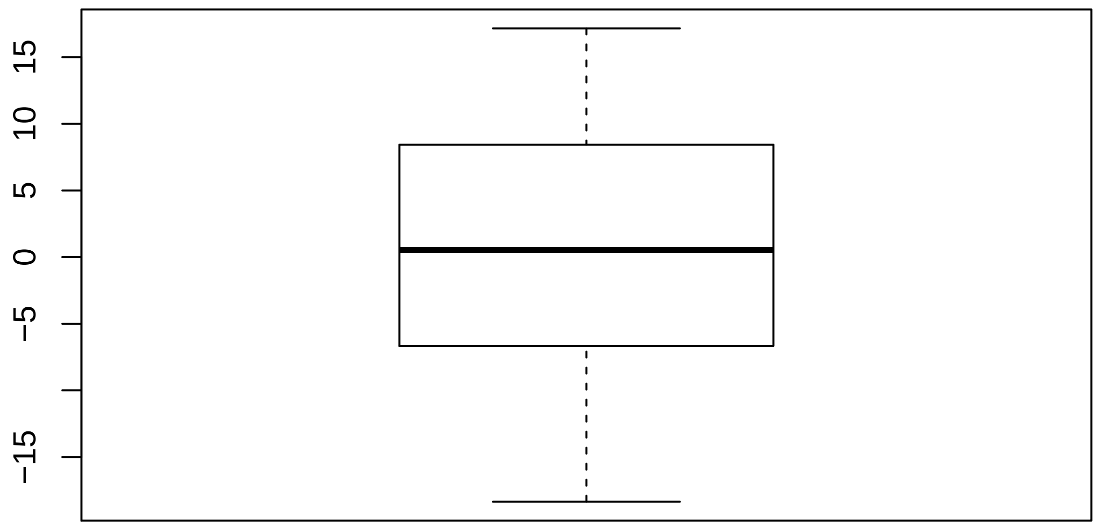
$$Y = 158.491 - 1.142X_1 - 0.442X_2 - 13.470X_3$$

b_2 is interpreted by the fact that patient satisfaction decreases as the severity of the illness goes up with other variables unchanged.

d.

box plot of the residuals:

```r
boxplot(reg$residuals)
```



There appears to be no outlier.

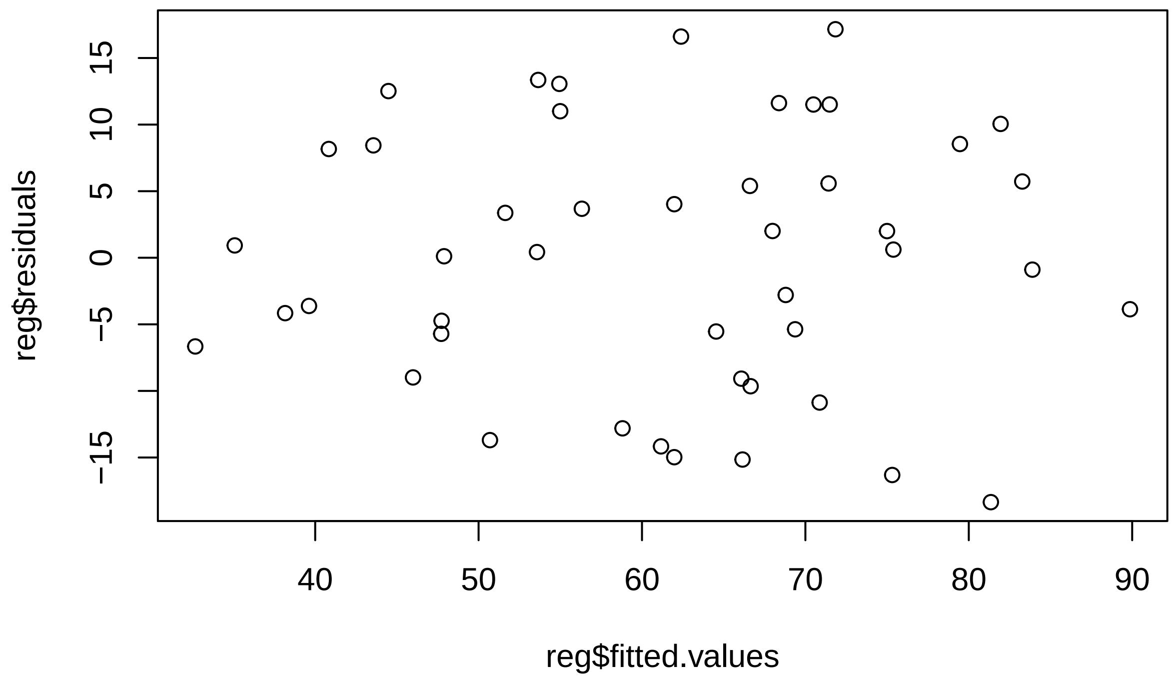
e.

```r
summary(reg)
```

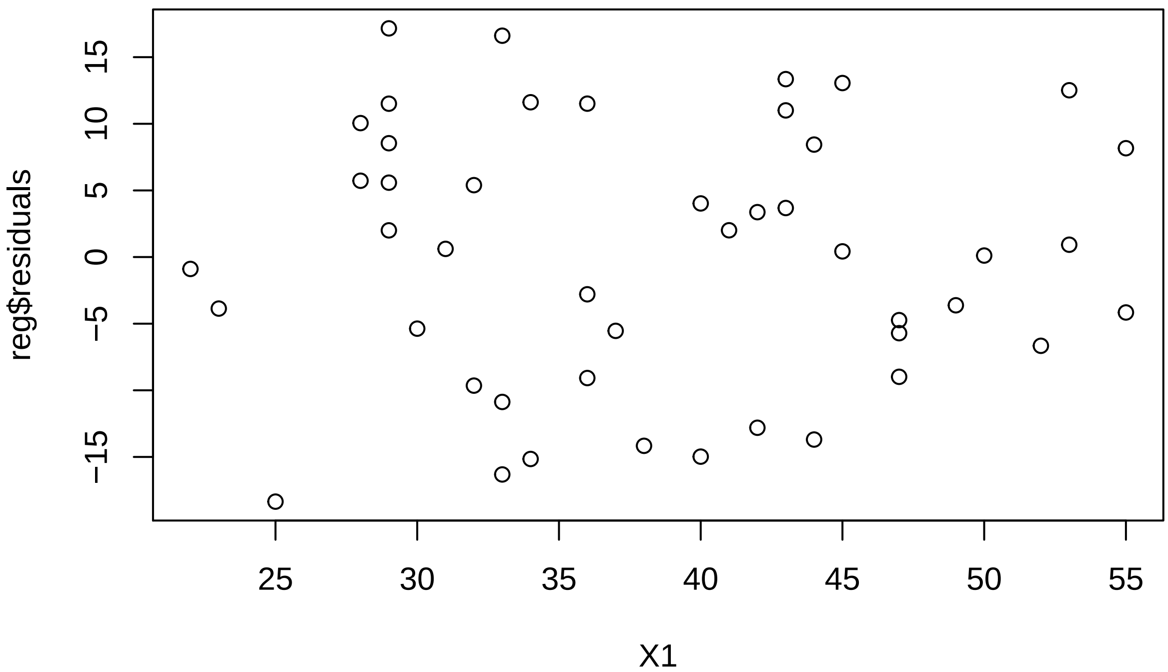
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## X1           -1.1416     0.2148  -5.315 3.81e-06 ***
## X2           -0.4420     0.4920  -0.898   0.3741
## X3          -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
```

```
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

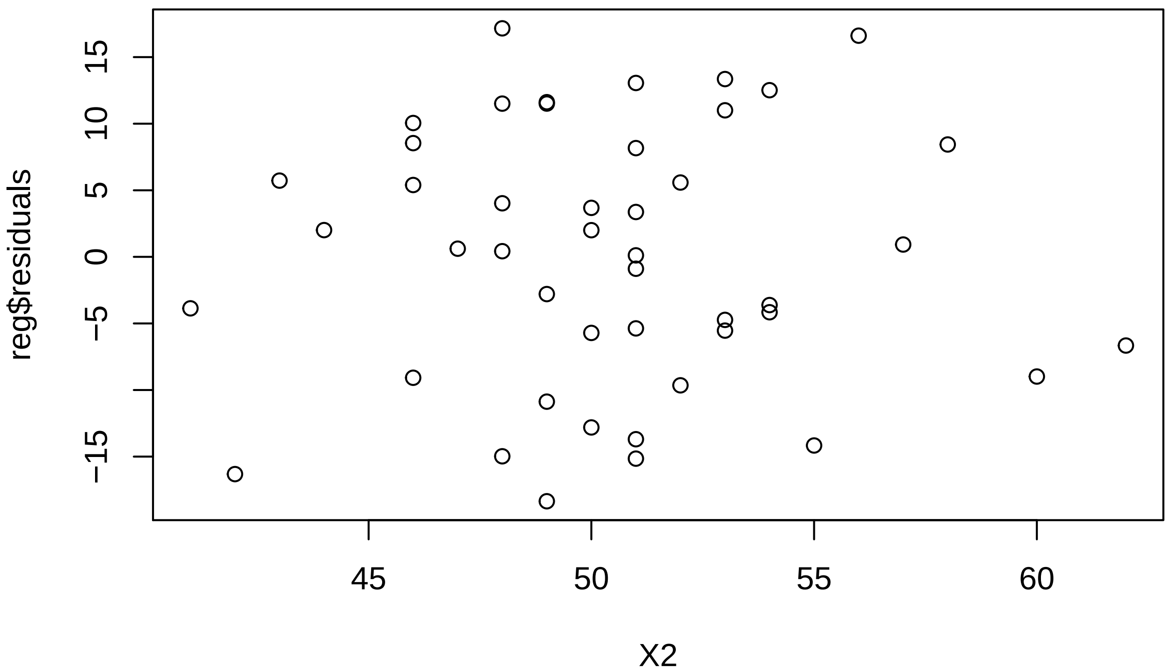
```r
plot(reg$fitted.values, reg$residuals)
```



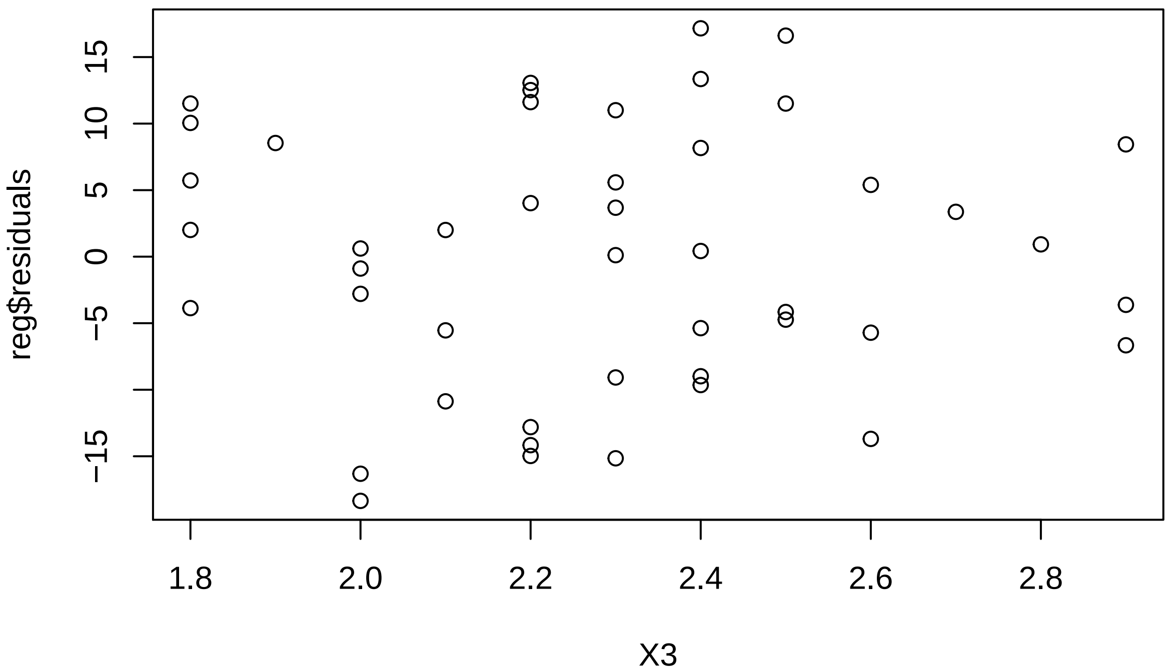

```r
plot(X1, reg$residuals)
```



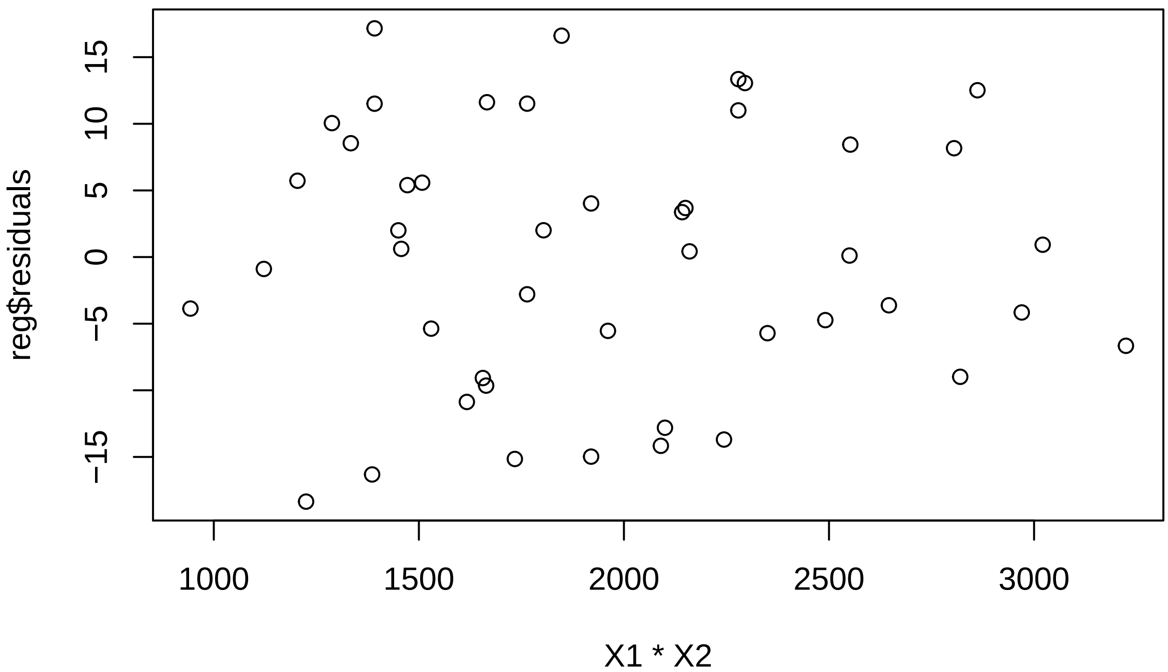

```r
plot(X2, reg$residuals)
```



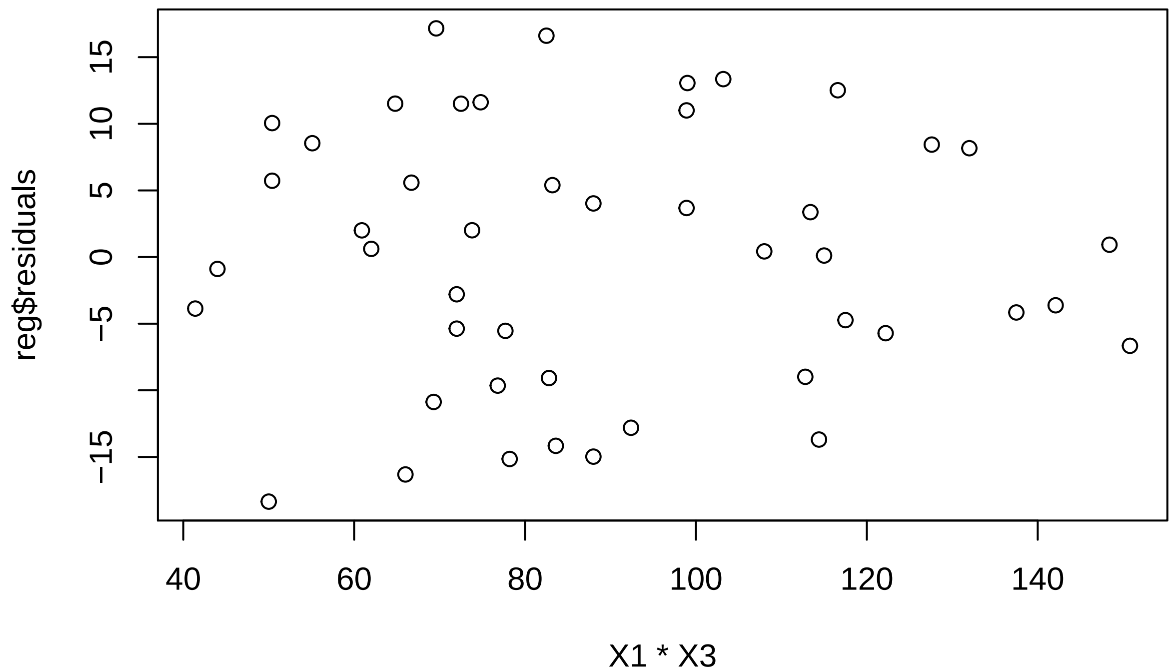

```
plot(X3, reg$residuals)
```



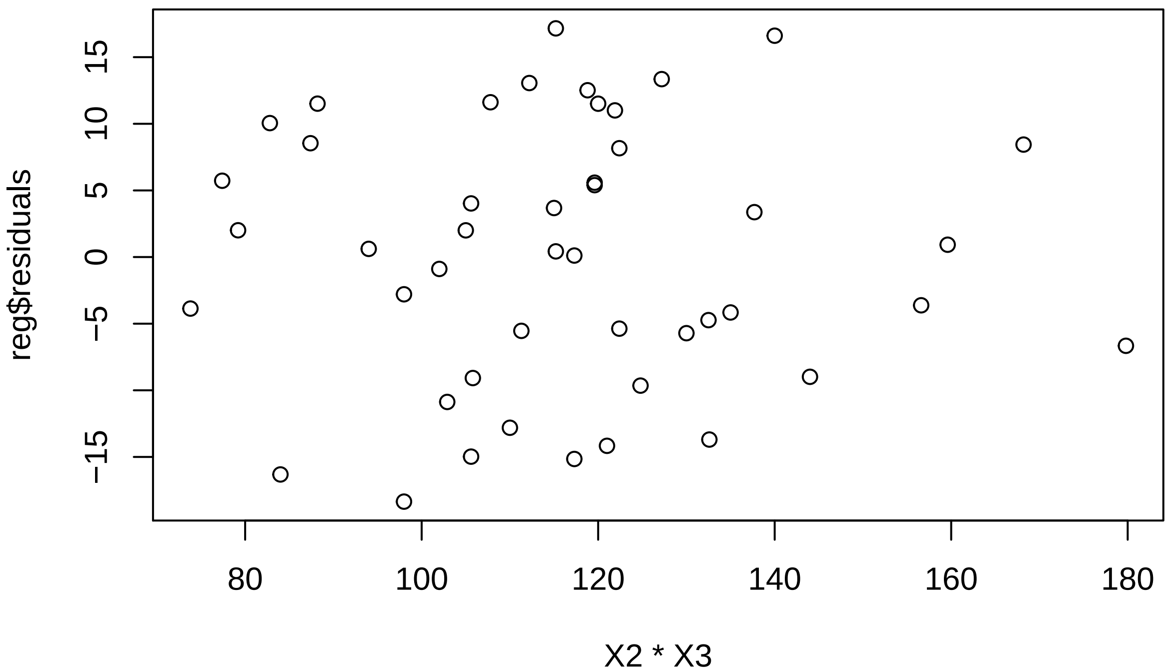

```
plot(X1*X2, reg$residuals)
```



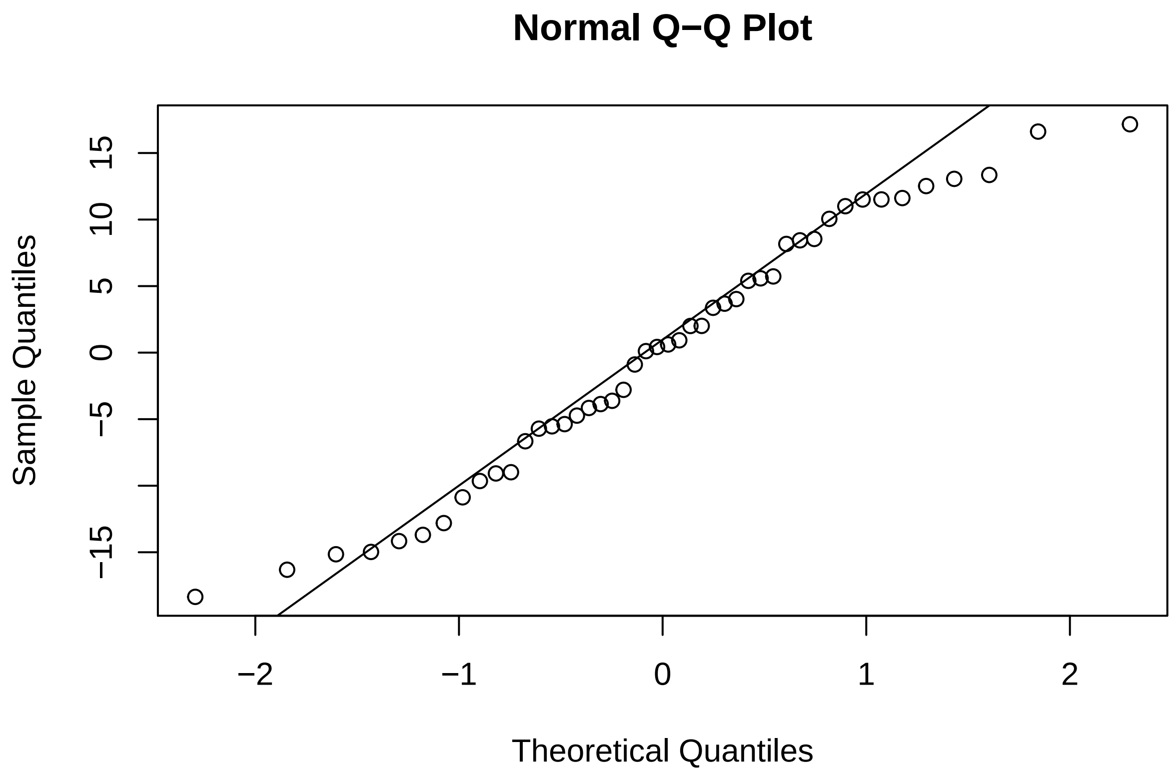

```
plot(X1*X3, reg$residuals)
```




```
plot(X2*X3, reg$residuals)
```




```
qqnorm(reg$residuals)
qqline(reg$residuals)
```



The residuals appear to fit the independence and equal-variance assumptions well, but not so good in normal assumption. And there is an obvious colinearity among X_1, X_2 and X_3, which ruins the regression model a lot.

KNNL 6.16

a.

$$H_0 : \beta_1 = \beta_2 = \beta_3 \leftrightarrow H_1 : \beta_k \neq 0 \ k = 1, 2 \ or \ 3$$

denote the test statistic

$$T^* = \frac{MSM}{MSE} \sim F_{p-1;n-p}$$

```
alpha=0.10
MSM=sum((reg$fitted.values-mean(Y))^2)/(p-1)
MSE=sum((reg$residuals)^2)/reg$df.residual
T=MSM/MSE
T
```

```
## [1] 30.05208
```

```
qf(1-alpha, p-1, n-p)
```

```
## [1] 2.219059
```

$T = 30.05288 > F_{p-1;n-p}(\alpha)$, so we reject H_0, which implies that at least one of β_1, β_2 and β_3 is not 0.

```r
1-pf(T, p-1 ,n-p)
```

```
## [1] 1.541973e-10
```

P-value is 1.5419×10^{-10}.

b.

```r
confint(reg, level=0.90)
```

```
##                    5 %        95 %
## (Intercept) 128.004370 188.9781330
## X1           -1.502893  -0.7803305
## X2           -1.269467   0.3854587
## X3          -25.411454  -1.5288719
```

c.

from the summary of reg we obtain that $R^2 = 0.6822$, so $R = 0.83$, which indicates that there appears to be a regression relation.

KNNL 6.17

a.

```r
predict(reg, newdata=data.frame(X1=35,X2=45,X3=2.2), se.fit=TRUE, interval="confidence", level=0.90)
```

```
## $fit
##        fit      lwr      upr
## 1 69.01029 64.52854 73.49204
##
## $se.fit
## [1] 2.664612
##
## $df
## [1] 42
##
## $residual.scale
## [1] 10.05798
```

b.

```r
predict(reg, newdata=data.frame(X1=35,X2=45,X3=2.2), se.fit=TRUE, interval="predict", level=0.90)
```

```
## $fit
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
##
## $se.fit
## [1] 2.664612
```

```
##
## $df
## [1] 42
##
## $residual.scale
## [1] 10.05798
```

KNNL 7.5

a.

```r
reg1 <- lm(Y ~ X2 + X1 + X3)
anova(reg1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 4860.3  4860.3 48.0439 1.822e-08 ***
## X1         1 3896.0  3896.0 38.5126 2.008e-07 ***
## X3         1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.

$$H_0 : \beta_3 = 0 \leftrightarrow H_1 : \beta_3 \neq 0$$
$$F^* = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)/n - p} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2, X_3)/42} \sim F_{1;42}$$

```r
F=364.2/(4248.8/42)
F
```

```
## [1] 3.600169
```

```r
qf(0.975, 1, 42)
```

```
## [1] 5.403859
```

$F^* = 3.6 < F_{1;42}(0.975)$, so we accept H_0.

P-value is

```r
1-pf(F, 1, 42)
```

```
## [1] 0.06466262
```

KNNL 7.6

$$H_0 : \beta_2, \beta_3 = 0 \leftrightarrow H_1 : \beta_2 \neq 0 \; or \; \beta_3 \neq 0$$
$$F^* = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)/n - p} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2, X_3)/42} \sim F_{1;42}$$

```r
F=(480.9+364.2)/2/(4248.8/42)
F
```

```
## [1] 4.176968
```

```r
qf(0.975, 2, 42)
```

```
## [1] 4.03271
```

$F^* = 4.18 > F_{1;42}(0.975)$, so we reject H_0.

P-value is

```r
1-pf(F, 2, 42)
```

```
## [1] 0.02215814
```

KNNL 7.9

$$H_0 : \beta_1 = -1.0, \beta_2 = 0 \leftrightarrow H_1 : \beta_1 \neq -1.0 \; or \; \beta_2 \neq 0$$

full model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

reduced model:

$$Y = \beta_0 - 1.0X_1 + \beta_3 X_3$$

$$F^* = \frac{(SSE(R) - SSE(F))/2}{SSE(F)/42} \sim F_{2;42}$$

```r
reg2<-lm(Y+X1 ~ X3)
SSE_R=sum(reg2$residuals^2)
SSE_F=sum(reg$residuals^2)
F=(SSE_R-SSE_F)/2/(SSE_F/42)
F
```

```
## [1] 0.8837939
```

```r
qf(0.975, 2, 42)
```

```
## [1] 4.03271
```

$F^* = 0.88 < F_{2;42}(0.975)$, so we accept H_0.

KNNL 7.26

a.

```r
reg3<-lm(Y ~ X1 + X2)
reg3
```

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

```
## (Intercept)           X1           X2
##    156.6719      -1.2677      -0.9208
```

fitted regression function:

$$Y = 156.67 - 1.27X_1 - 0.92X_2$$

b.

β_2 changes a lot while β_1 appears to remain.

c.

No, $SSR(X_1) = 8105.0$, while $SSR(X_1|X_3) = 3309.3$.

No, $SSR(X_2) = 4824.4$, while $SSR(X_2|X_3) = 693.8$

d.

The linear relationship between X_2 and X_3 appears to be relatively strong, which affects β_2 in (b) and suggets colinearity may exists in the data.

KNNL 7.29

a.

$$\begin{aligned} SSR(X_1, X_2, X_3, X_4) &= SSR(X_1) + (SSR(X_1, X_2, X_3) - SSR(X_1)) + (SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)) \\ &= SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3) \end{aligned}$$

b.

$$\begin{aligned} SSR(X_1, X_2, X_3, X_4) &= SSR(X_2, X_3) + (SSR(X_1, X_2, X_3) - SSR(X_2, X_3)) + (SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)) \\ &= SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3) \end{aligned}$$