Applied Linear Statistical Models

Homework 5

Qiuyang Yin, 2015011468

April 18th, 2017

Problem 1

Preparations

```r
rm(list=ls())
x=c(8,4,0,-4,-8)
y=c(7.8,9.0,10.2,11.0,11.7)
X=matrix(c(rep(1,5),x),nrow = 5,ncol = 2)
Y=t(t(y))
```

KNNL 5.4

- $Y'Y$, $X'X$ and $X'Y$

```r
t(Y) %*% Y
```

```
##        [,1]
## [1,] 503.77
```

```r
t(X) %*% X
```

```
##      [,1] [,2]
## [1,]    5    0
## [2,]    0  160
```

```r
t(X) %*% Y
```

```
##       [,1]
## [1,]  49.7
## [2,] -39.2
```

KNNL 5.12

- $(X'X)^{-1}$

```r
solve(t(X) %*% X)
```

```
##      [,1]    [,2]
## [1,]  0.2 0.00000
## [2,]  0.0 0.00625
```

Problem 2

a

- vectors of estimated regression coefficients
- vectors of residuals
- SSR
- SSE
- estimated variance-covariance matrix of **b**
- point estimate of EY_h when $X_h = -6$
- estimated variance of $\hat{Y}_h$ when $X_h = -6$
- using the matrix methods

regression coefficients

```
(b=solve(t(X) %*% X) %*% t(X) %*% Y)
```

```
##         [,1]
## [1,]  9.940
## [2,] -0.245
```

residuals

```
(e=Y-X %*% b)
```

```
##       [,1]
## [1,] -0.18
## [2,]  0.04
## [3,]  0.26
## [4,]  0.08
## [5,] -0.20
```

SSR

```
J=matrix(rep(1,25),nrow = 5)
H= X %*% (solve(t(X) %*% X)) %*% t(X)
(SSR=t(Y) %*% (H-(1/5)*J) %*% Y)
```

```
##       [,1]
## [1,] 9.604
```

SSE

```
(SSE=t(e) %*% e)
```

```
##       [,1]
## [1,] 0.148
```

variance-covariance matirx of b

```
MSE=as.numeric(SSE/(5-2))
MSE * solve(t(X) %*% X)
```

```
##             [,1]         [,2]
## [1,] 0.009866667 0.0000000000
## [2,] 0.000000000 0.0003083333
```

point estimate

```
Xh=as.matrix(c(1,-6))
t(Xh) %*% b
```

```
##       [,1]
## [1,] 11.41
```

variance of the point estimate

```
MSE*(t(Xh) %*% solve(t(X) %*% X) %*% Xh)
```

```
##            [,1]
## [1,] 0.02096667
```

verification using another way

```
fit=(lm(y~x))
fit$coefficients
```

```
## (Intercept)           x
##       9.940      -0.245
```

```
fit$residuals
```

```
##     1     2     3     4     5
## -0.18  0.04  0.26  0.08 -0.20
```

```
(SSR=sum((fit$fitted.values-mean(fit$fitted.values))^2))
```

```
## [1] 9.604
```

```
(SSE=sum(fit$residuals^2))
```

```
## [1] 0.148
```

```
(conf=predict(fit,se.fit = T,data.frame(x=-6),
              interval = "confidence",level = 0.95))
```

```
## $fit
##     fit      lwr      upr
## 1 11.41 10.94919 11.87081
##
## $se.fit
## [1] 0.1447987
##
## $df
## [1] 3
##
## $residual.scale
## [1] 0.2221111
```

```
conf$se.fit^2
```

```
## [1] 0.02096667
```

b

- Simplication arose from spacing of X levels

In this experiment, Xs are in the same intervals, and the mean value of X is 0.

According to the equations in KNNL section 4.7:

$$\sigma^2 b_0 = \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}]$$

So when $\bar{X} = 0$, the variance of intercept is minimize to $\sigma^2(\frac{1}{n})$

c

- Find hat matrix and its rank
- verify that H is idempotent

```
(H= X %*% (solve(t(X) %*% X)) %*% t(X))
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.6  0.4  0.2  0.0 -0.2
## [2,]  0.4  0.3  0.2  0.1  0.0
## [3,]  0.2  0.2  0.2  0.2  0.2
## [4,]  0.0  0.1  0.2  0.3  0.4
## [5,] -0.2  0.0  0.2  0.4  0.6
```

And the rank of H matrix is:

```r
qr(H)$rank
```

```
## [1] 2
```

H is idempotent,since:

```r
H %*% H
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.6  0.4  0.2  0.0 -0.2
## [2,]  0.4  0.3  0.2  0.1  0.0
## [3,]  0.2  0.2  0.2  0.2  0.2
## [4,]  0.0  0.1  0.2  0.3  0.4
## [5,] -0.2  0.0  0.2  0.4  0.6
```

Is identical to H matrix

d

- find $S^2\{e\}$

```r
MSE*(diag(5)-H)
```

```
##              [,1]         [,2]         [,3]         [,4]         [,5]
## [1,]  0.019733333 -0.019733333 -0.009866667  0.000000000  0.009866667
## [2,] -0.019733333  0.034533333 -0.009866667 -0.004933333  0.000000000
## [3,] -0.009866667 -0.009866667  0.039466667 -0.009866667 -0.009866667
## [4,]  0.000000000 -0.004933333 -0.009866667  0.034533333 -0.019733333
## [5,]  0.009866667  0.000000000 -0.009866667 -0.019733333  0.019733333
```

Problem 3

a

- Write down the linear regresison model in matrix form togerther with proper assumptions

```r
dat=read.table("CH06PR05_83630456.txt")
colnames(dat)=c("Y","X1","X2")
dat$X1X2=dat$X1*dat$X2
```

The model:

$$\mathbf{Y} = \mathbf{X}_{5\times4}\beta_{4\times1} + \epsilon_{5\times1}$$

where

$$X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,1}X_{1,2} \\ 1 & X_{2,1} & X_{2,2} & X_{2,1}X_{2,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{5,1} & X_{5,2} & X_{5,1}X_{5,2} \end{pmatrix}$$

and

$$\beta = (\beta_0, \beta_1, \cdots, \beta_4)'$$
$$\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_5)$$

Model Assumptions:

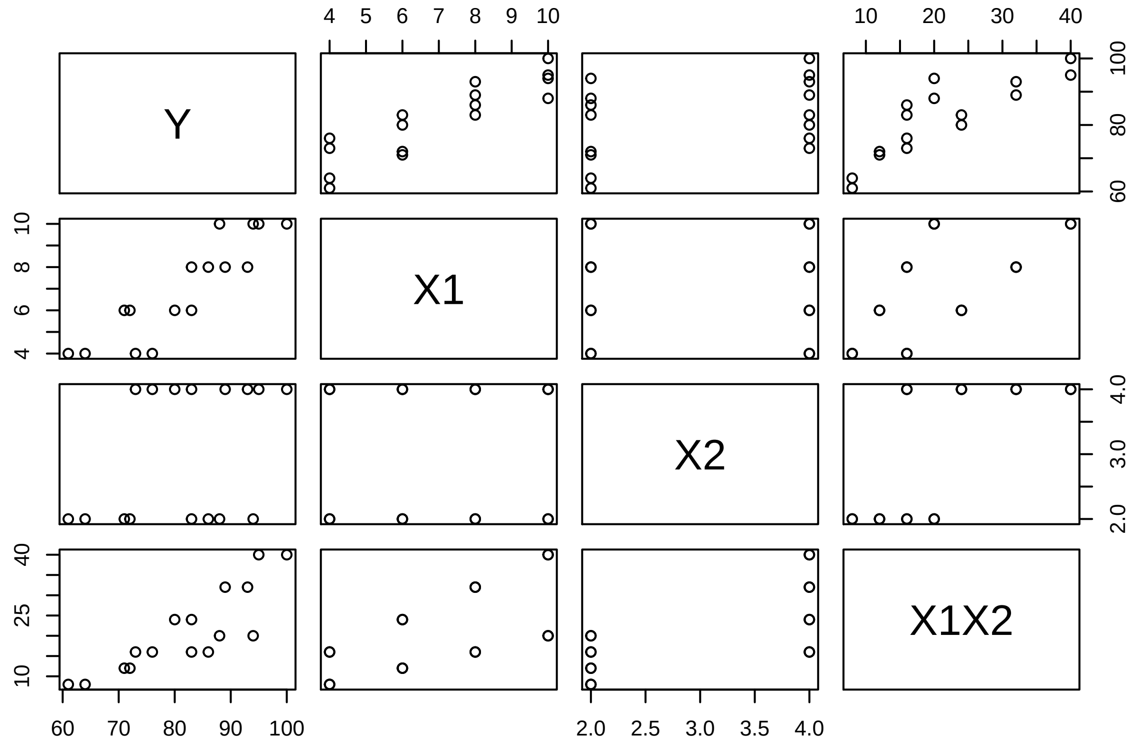
$$\epsilon \sim N(0, \sigma^2 I_n)$$
$$Y \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

i.e residuals are iids.

b

- Obtain the scatter plot and correlation matrix.
- What infomation provided?

```
pairs(dat)
```



```
cor(dat)
```

```
##              Y        X1        X2      X1X2
## Y    1.0000000 0.8923929 0.3945807 0.8565881
## X1   0.8923929 1.0000000 0.0000000 0.6741999
## X2   0.3945807 0.0000000 1.0000000 0.7035265
## X1X2 0.8565881 0.6741999 0.7035265 1.0000000
```

It seems that Y shows a strong positive relationship with X1 and X1X2. Besides, X1 and X2 are uncorrelated.

c

- Fit the model
- Report the fitted regression model
- ANOVA test results
- R^2 and R_a^2
- estimate of error variance

```
fit=lm(Y~X1+X2+X1X2,data = dat)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1X2, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.150 -1.488  0.125  1.700  3.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.1500     6.4648   4.200  0.00123 **
## X1            5.9250     0.8797   6.735 2.09e-05 ***
## X2            7.8750     2.0444   3.852  0.00230 **
## X1X2         -0.5000     0.2782  -1.797  0.09749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 12 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9528
## F-statistic: 101.9 on 3 and 12 DF,  p-value: 8.379e-09
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 1566.45 1566.45 252.9933 1.984e-09 ***
## X2         1  306.25  306.25  49.4616 1.370e-05 ***
## X1X2       1   20.00   20.00   3.2301   0.09749 .
## Residuals 12   74.30    6.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fitted model

So the fitted model is:

$$\hat{Y}i = 27.15 + 5.925X1 + 7.875X2 - 0.5X1X2$$

ANOVO test results

$$F = 101.9, p - value = 8.379e - 09$$

reject H_0 and maintain that at least one of the parameter is useful.

R^2 and R_a^2

$$R^2 = 0.9622, R_a^2 = 0.9528$$

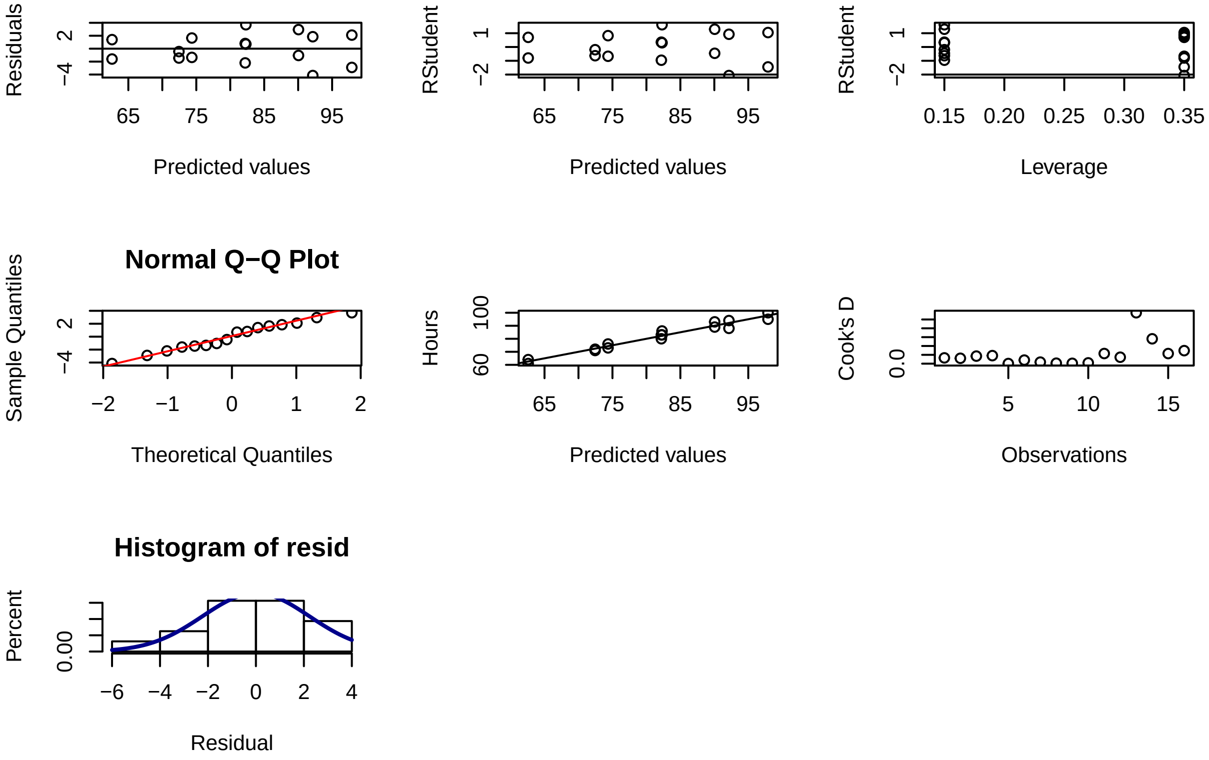
estimate of error variance

$$MSE = 2.488^2 = 6.19$$

d

- obtain the residual plots
- comment on the assumptions of the regression model

```r
source("residualPlot.r")
residualplot(fit)
```



From the first two figures, we don't see any trend , and the variance is constant. So we can conclude that the residuals are iid with constant variance.

In addition, from histogram and Normal quantile plot, we can conclude that the residuals follow approximately normal distribution.

e

```
newdata=dat[,-4]
```

write down matrix

The model:

$$\mathbf{Y} = \mathbf{X}_{5\times 3}\beta_{3\times 1} + \epsilon_{5\times 1}$$

where

$$X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{5,1} & X_{5,2} \end{pmatrix}$$

and

$$\beta = (\beta_0, \beta_1, \cdots, \beta_3)'$$

$$\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_4)$$

Model Assumptions:

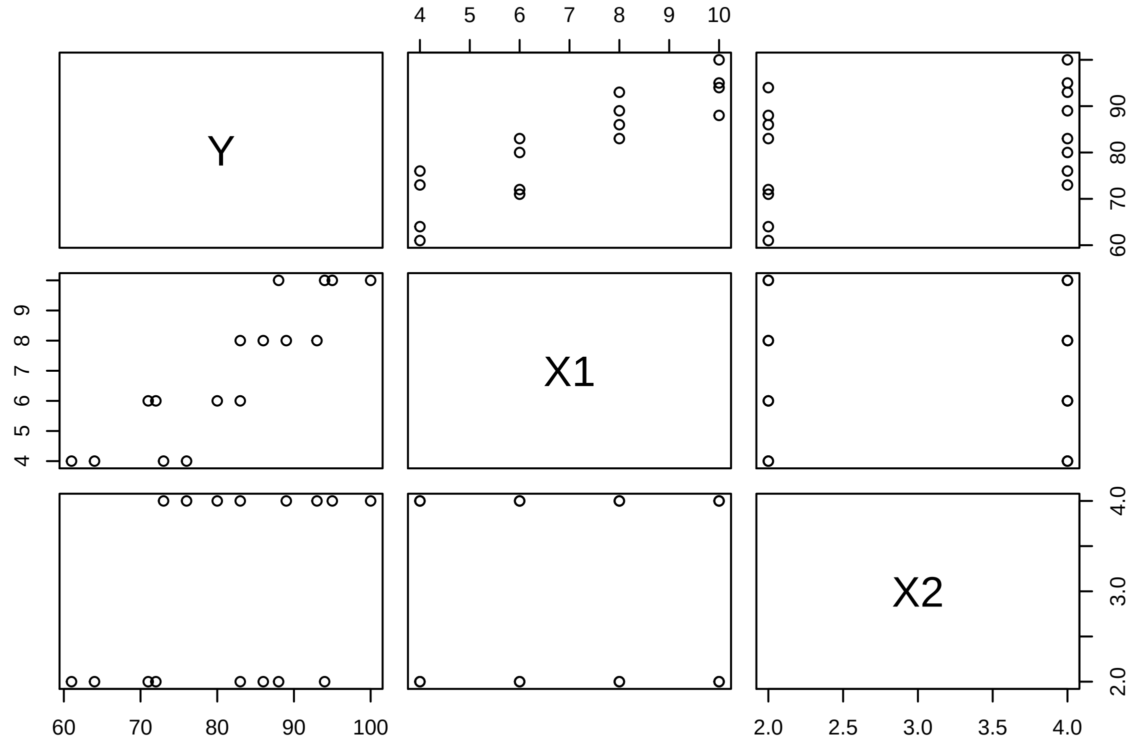
$$\epsilon \sim N(0, \sigma^2 I_n)$$

$$Y \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

i.e residuals are iids.

scatter plot and correlation

```
pairs(newdata)
```



```r
cor(newdata)
```

```
##            Y        X1        X2
## Y  1.0000000 0.8923929 0.3945807
## X1 0.8923929 1.0000000 0.0000000
## X2 0.3945807 0.0000000 1.0000000
```

regression results

```r
reduced=lm(Y~X1+X2,data = newdata)
summary(reduced)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

```r
anova(reduced)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the regression model is:

$$\hat{Y} = 37.65 + 4.425 + X_1 + 4.3750X_2$$

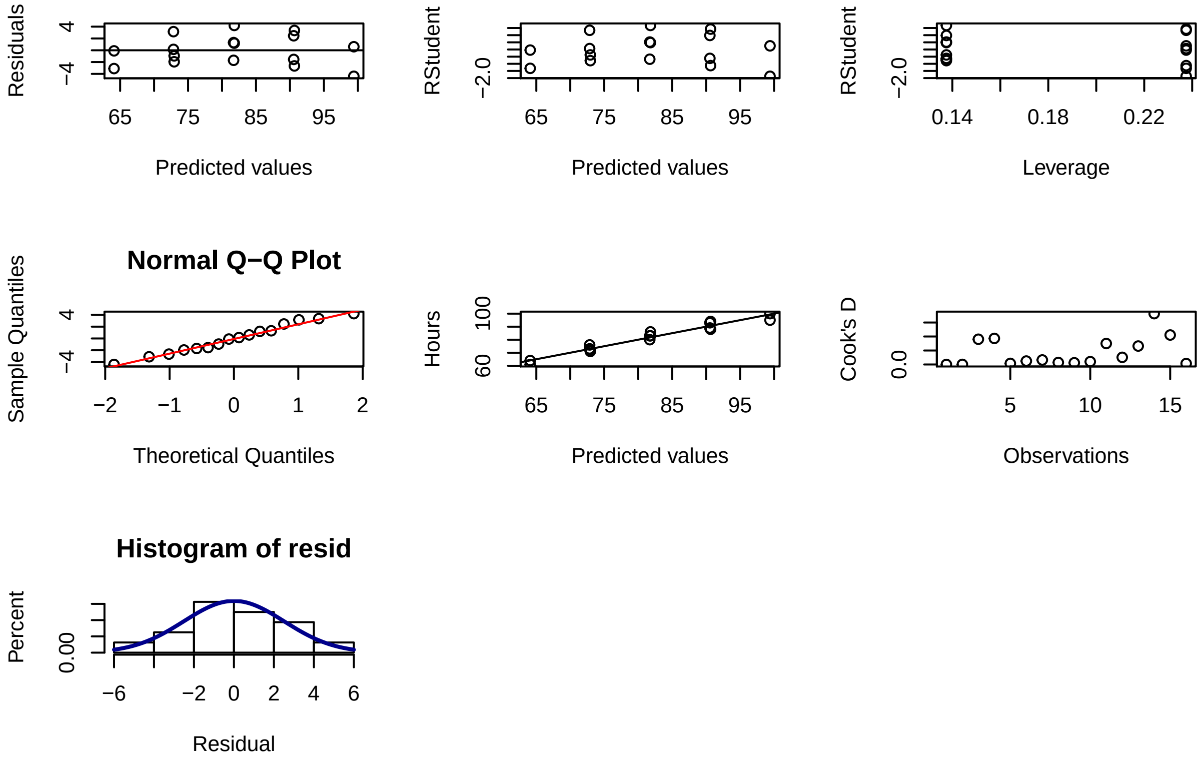
And the anova test is:

$$F = 129.1, p - value = 2.658e - 09$$

we reject H_0 conclude that at least one of the parameter is useful.

residual plots

```r
residualplot(reduced)
```



The conclusion is similar to the one of full model. The variance is constant and residuals approximately follow the normal distribution.

f

```
anova(reduced,fit)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X1X2
##   Res.Df  RSS Df Sum of Sq      F  Pr(>F)
## 1     13 94.3
## 2     12 74.3  1        20 3.2301 0.09749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since P-value is 0.09, We do not reject H_0, and use reduced model without X1X2 term