HW6

Luo Bingjun 2017013573 Software 71

2019/4/29

Prob-1: KNNL 5.4, KNNL 5.12

```
n=5
X<-matrix(c(1,1,1,1,1,8,4,0,-4,-8), ncol=2)
Y<-matrix(c(7.8,9.0,10.2,11.0,11.7))
```

(1)

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

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

(2)

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

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

(3)

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

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

(4)

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

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

Prob-2: KNNL 5.23

a.

vector of estimated regression coeffecients:

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

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

vector of residuals:

```
H=X%*%solve(t(X)%*%X)%*%t(X)
e=(diag(n)-H)%*%Y
e
```

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

SSR:

```
J=matrix(rep(1,time=n^2),nrow=n)
SSR=t(b)%*%t(X)%*%Y-(1/n)*t(Y)%*%J%*%Y
SSR
```

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

SSE:

```
SSE=t(e)%*%e
MSE=SSE/(n-2)
SSE
```

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

estimated variance-covariance matrix of b:

```
sigma_b=MSE[1,1]*solve(t(X)%*%X)
sigma_b
```

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

point estimate of $E(Y_h)$ when $X_h = -6$:

```
X_h=matrix(c(1, -6))
t(X_h)%*%b
```

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

estimated variance of $\hat{Y}_h$ when $X_h = -6$:

```
t(X_h)%*%sigma_b%*%X_h
```

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

b.

X is symmetric with a mean of 0.

c.

$H =$

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

its rank:

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

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

verifying it is idempotent:

```
all.equal(H%*%H,H)
```

```
## [1] TRUE
```

d.

```
(diag(n)-H)%*%(MSE[1,1]*diag(n))%*%t(diag(n)-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
```

Prob-3: KNNL 6.5

```
n=16
data = read.table(file='CH06PR05.txt', header=F)
data = cbind(data,data$V2*data$V3)
colnames(data) <- c('Y', 'X1', 'X2', 'X12')
```

a.

denote $X = (1_n, X_1, X_2, X_{12})$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})'$, then

$$Y = X\beta + \epsilon$$

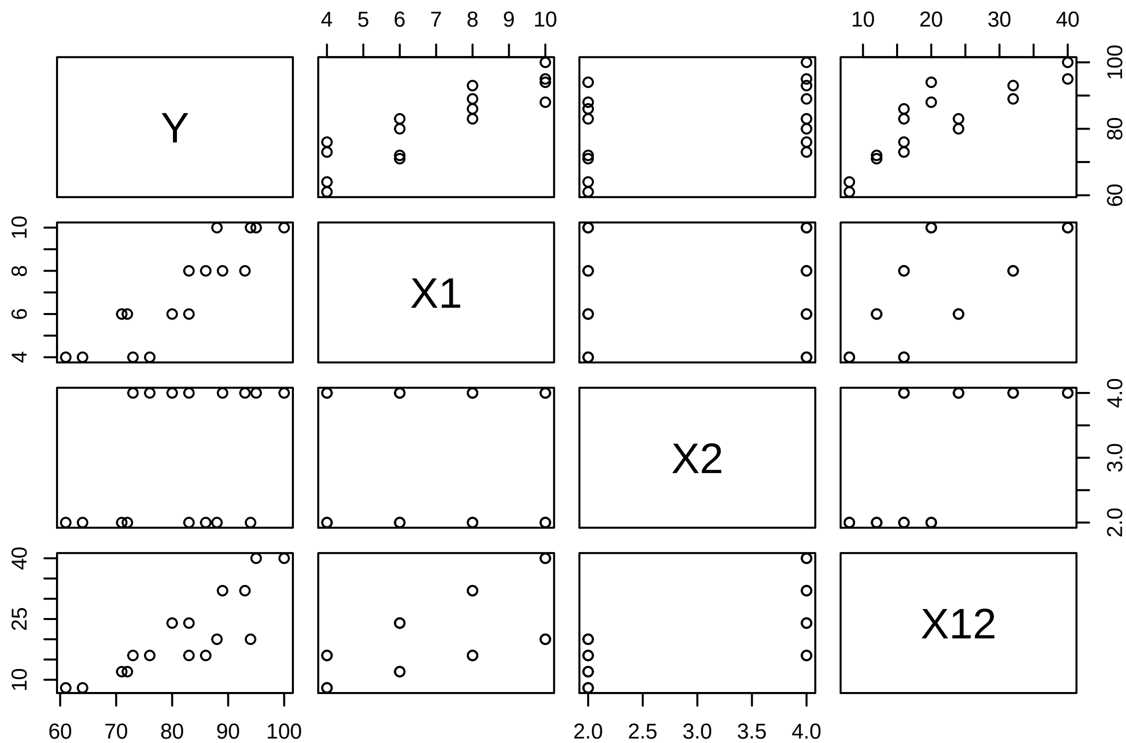
assumptions:

$$\epsilon_i \ i.i.d. \sim N(0, \sigma^2), \ i = 1, 2, \cdots, n$$

b.

scatter plot matrix:

```r
pairs(data[, c("Y", "X1", "X2", "X12")])
```



correlation matrix:

```r
cor(data)
```

```
##             Y        X1        X2       X12
## 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
## X12 0.8565881 0.6741999 0.7035265 1.0000000
```

There appears to be linear relations between Y and X_1, Y and X_1X_2, except for Y and X_2.

c.

the fitted regression model:

```r
reg=lm(Y ~ X1 + X2 + X12, data=data)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X12, data = data)
##
## 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 **
## X12          -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 test results:

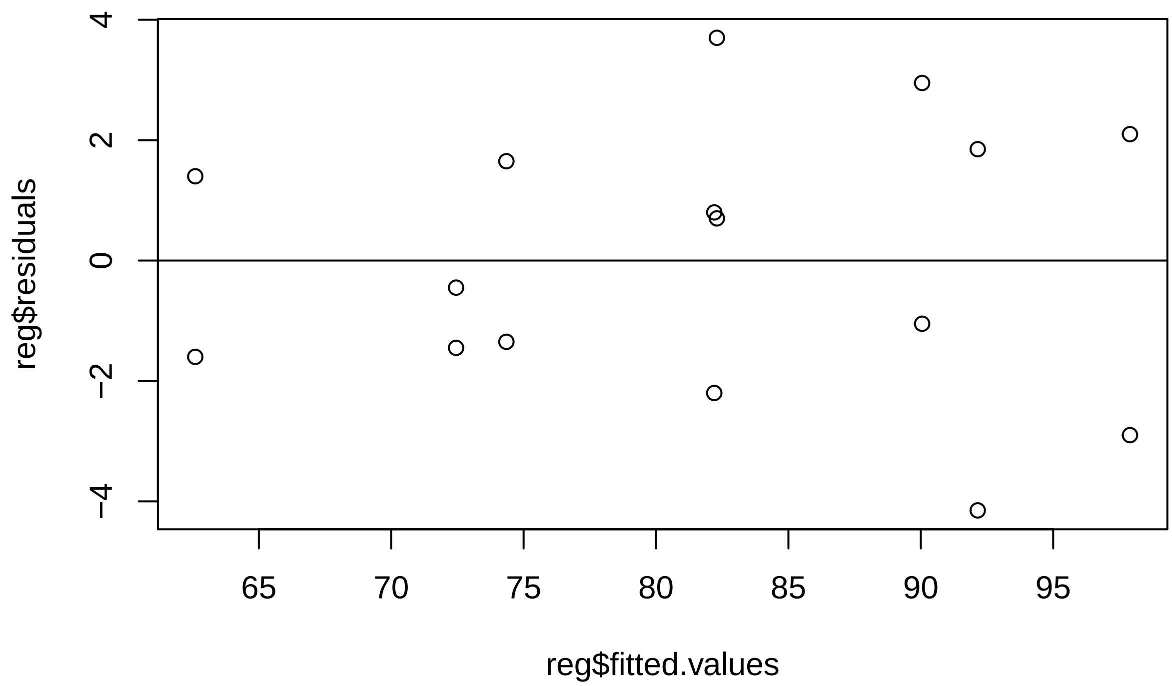
```r
anova(reg)
```

```
## 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 ***
## X12        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
```

R^2 is 0.9622, adjusted R_a^2 is 0.9528, the estimate of error variance is $2.488^2 = 6.190$.

d.

```r
plot(reg$fitted.values, reg$residuals)
abline(0,0)
```



The assumptions appear to fit the data.

e.

reduced model:

denote $X = (1_n, X_1, X_2)$, $\beta = (\beta_0, \beta_1, \beta_2)'$, then

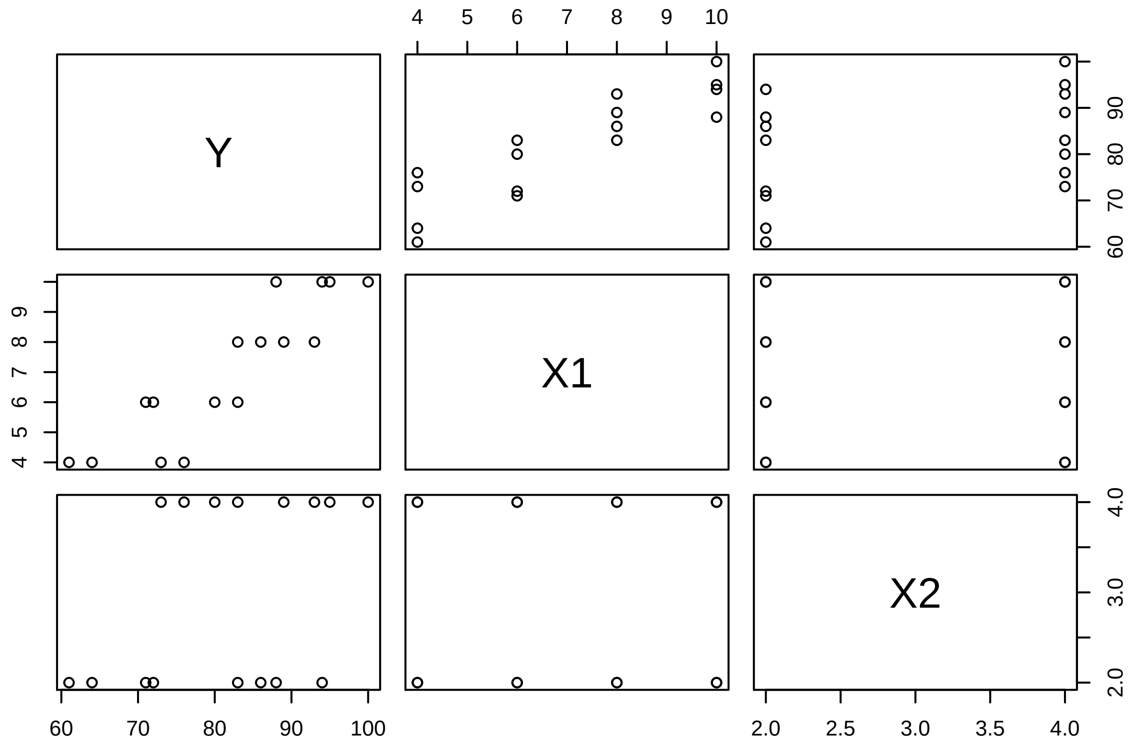
$$Y = X\beta + \epsilon$$

assumptions:

$$\epsilon_i \ i.i.d. \sim N(0, \sigma^2), \ i = 1, 2, \cdots, n$$

scatter plot matrix:

```r
pairs(data[, c("Y", "X1", "X2")])
```



correlation matrix:

```r
cor(data[,1:3])
```

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

the fitted regression model:

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

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

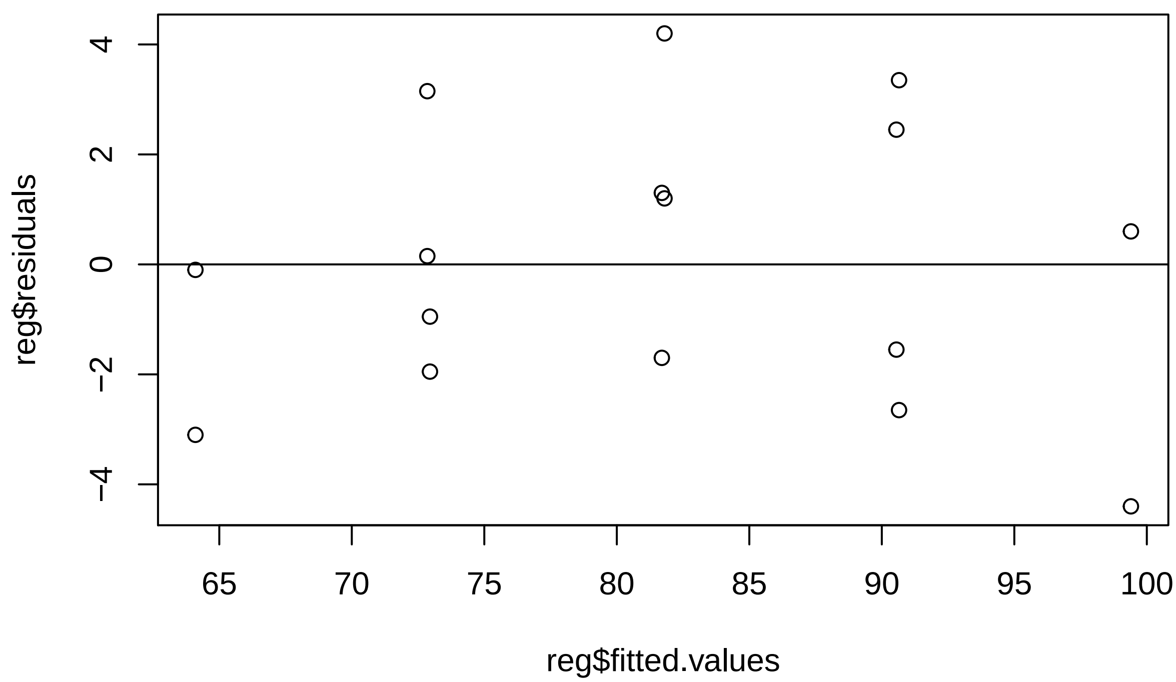
ANOVA test results:

```
anova(reg)
```

```
## 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
```

R^2 is 0.9521, adjusted R_a^2 is 0.447, the estimate of error variance is $2.693^2 = 7.252$.

```
plot(reg$fitted.values, reg$residuals)
abline(0,0)
```



The assumptions still appears to be fit for the data.

f.

I would recommand the non-reduced model, because the added X_1X_2 term seems to fit the linear relation assumptions, and it does have smaller MSE.