

HW4: 563 Regression Analysis

HW4

1.

```
> Y<-c(64,78,83,88,89,99,101,102)
```

```
> Y
```

```
[1] 64 78 83 88 89 99 101 102
```

```
> X<-matrix(c(1,1,1,1,1,1,1,1,58,84,78,81,82,102,85,102,111,131,158,147,121,165,174,169),ncol=3)
```

```
> X
```

```
      [,1] [,2] [,3]
[1,]    1   58  111
[2,]    1   84  131
[3,]    1   78  158
[4,]    1   81  147
[5,]    1   82  121
[6,]    1  102  165
[7,]    1   85  174
[8,]    1  102  169
```

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

```
      [,1] [,2] [,3]
[1,]    8   672  1176
[2,]  672  57822 100453
[3,] 1176 100453 176758
```

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

```
      [,1]
[1,]   704
[2,] 60251
[3,] 105288
```

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

```
      [,1] [,2] [,3]
[1,] 6.34817305 -0.0317491208 -0.0241921558
[2,] -0.03174912  0.0015216522 -0.0006535351
[3,] -0.02419216 -0.0006535351  0.0005380211
```

```
> B<-solve((t(X)%*%X))%*%t(X)%*%Y
```

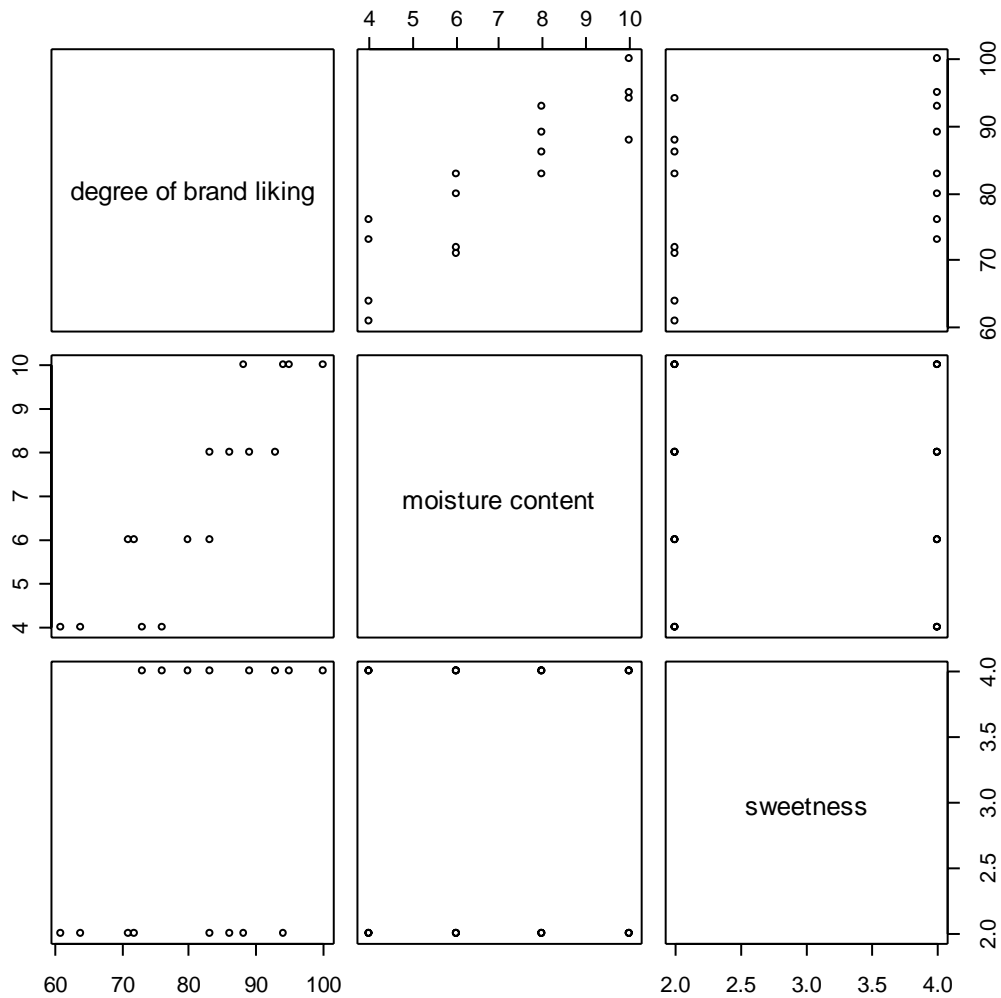
```
> B
```

```
      [,1]
[1,] 9.0538499
[2,] 0.5202790
[3,] 0.2397464
```

6.5

(a)

```
> mod<-read.table("CH06PR05.txt",head=T)
> mod.lm<-lm(mod$Y~mod$X1+mod$X2)
> Z<-cbind(mod$Y,mod$X1,mod$X2)
> pairs(Z,labels=c("degree of brand liking","moisture content","sweetness"))
```



```
> cor(Z)
      [,1] [,2] [,3]
[1,] 1.0000000 0.8923929 0.3945807
[2,] 0.8923929 1.0000000 0.0000000
[3,] 0.3945807 0.0000000 1.0000000
#the correlation matrix is the above matrix
```

(b)

```
> summary(mod.lm)
```

Call:

```
lm(formula = mod$Y ~ mod$X1 + mod$X2)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.400	-1.762	0.025	1.588	4.200

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	37.6500	2.9961	12.566	1.20e-08	***
mod\$X1	4.4250	0.3011	14.695	1.78e-09	***
mod\$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

#the estimated regression function is: $\hat{Y} = 37.65 + 4.425X_1 + 4.375X_2$

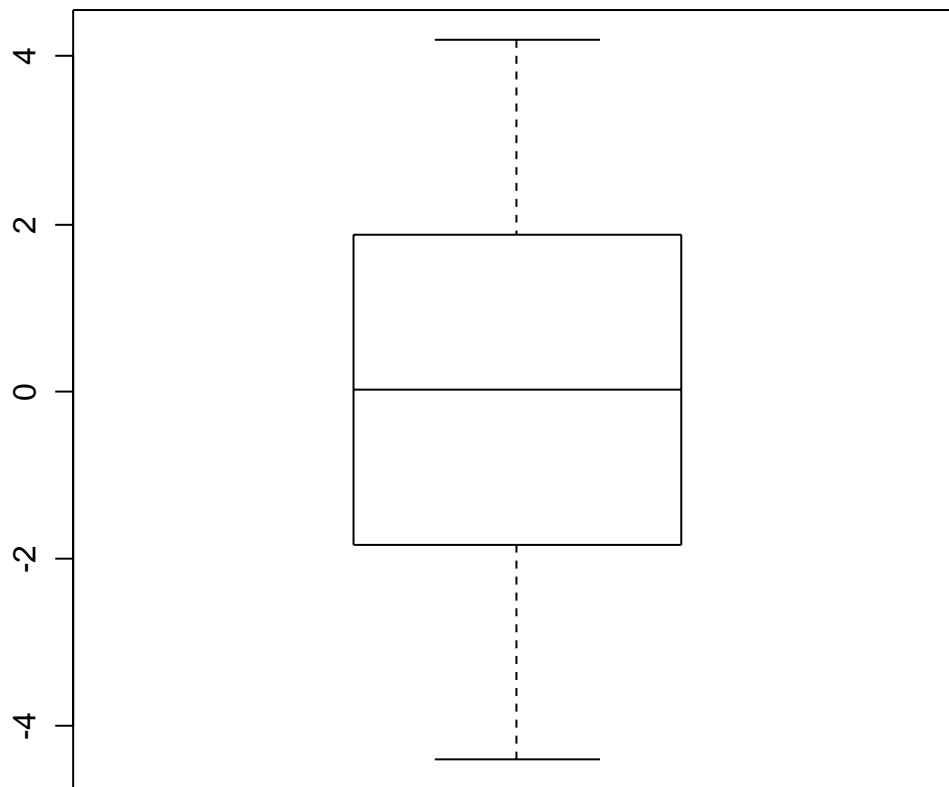
#b1 is the increase of the mean of degree of brand liking when fixing sweetness and increasing moisture content by 1

(c)

```
> mod.lm$residuals
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-0.10	0.15	-3.10	3.15	-0.95	-1.70	-1.95	1.30	1.20	-1.55	4.20	2.45	-2.65	-4.40	3.35	0.60

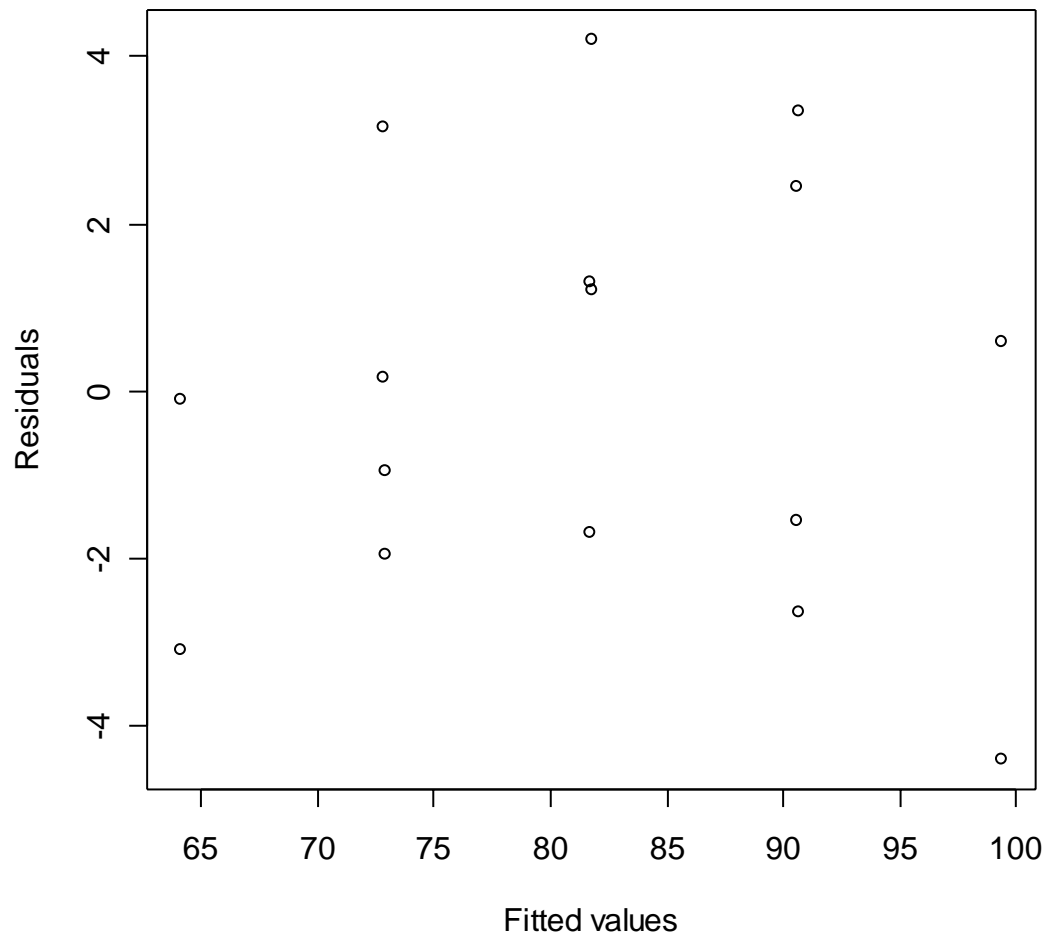
```
> boxplot(mod.lm$res)
```



#the above box plot tells that the maximum and minimum of residuals, the first and third quartiles, and the median
#residual. And the median is located in the middle of the central box, which indicates that the residuals are
#symmetrically distributed.

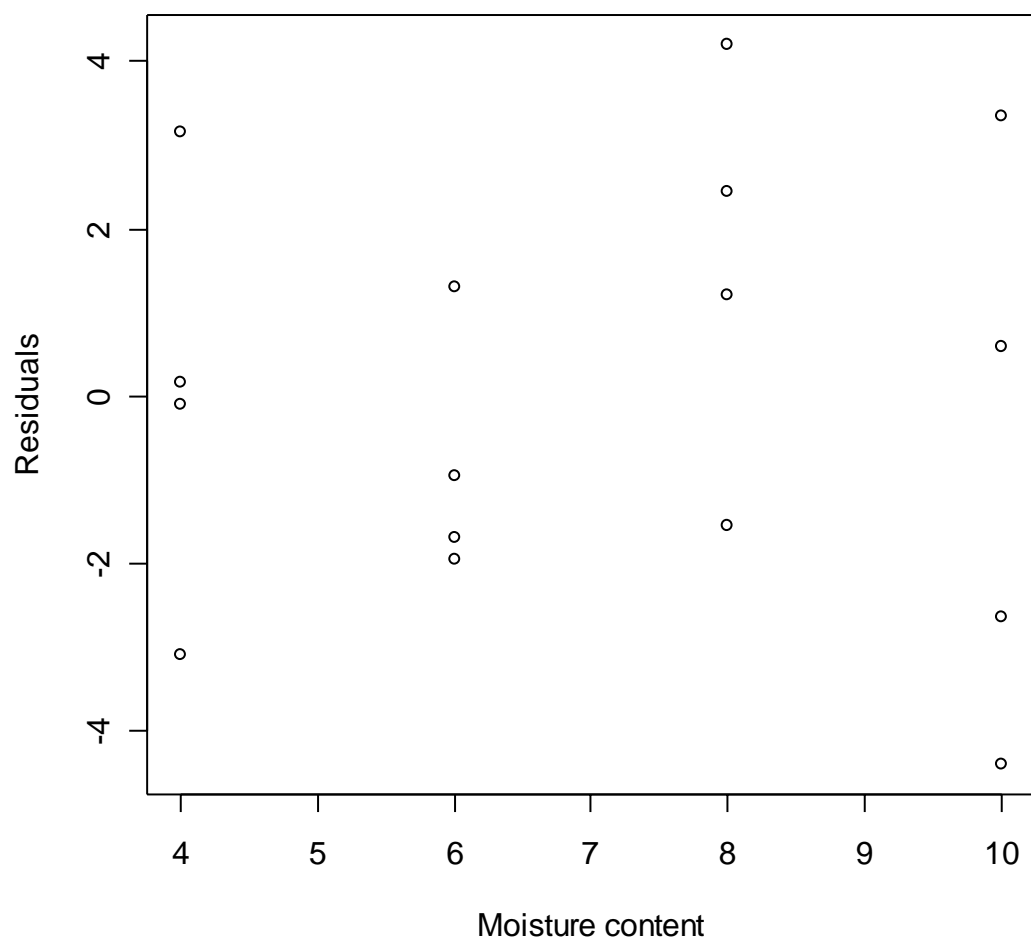
(d)

```
> plot(mod.lm$fit,mod.lm$res,xlab="Fitted values",ylab="Residuals")
```



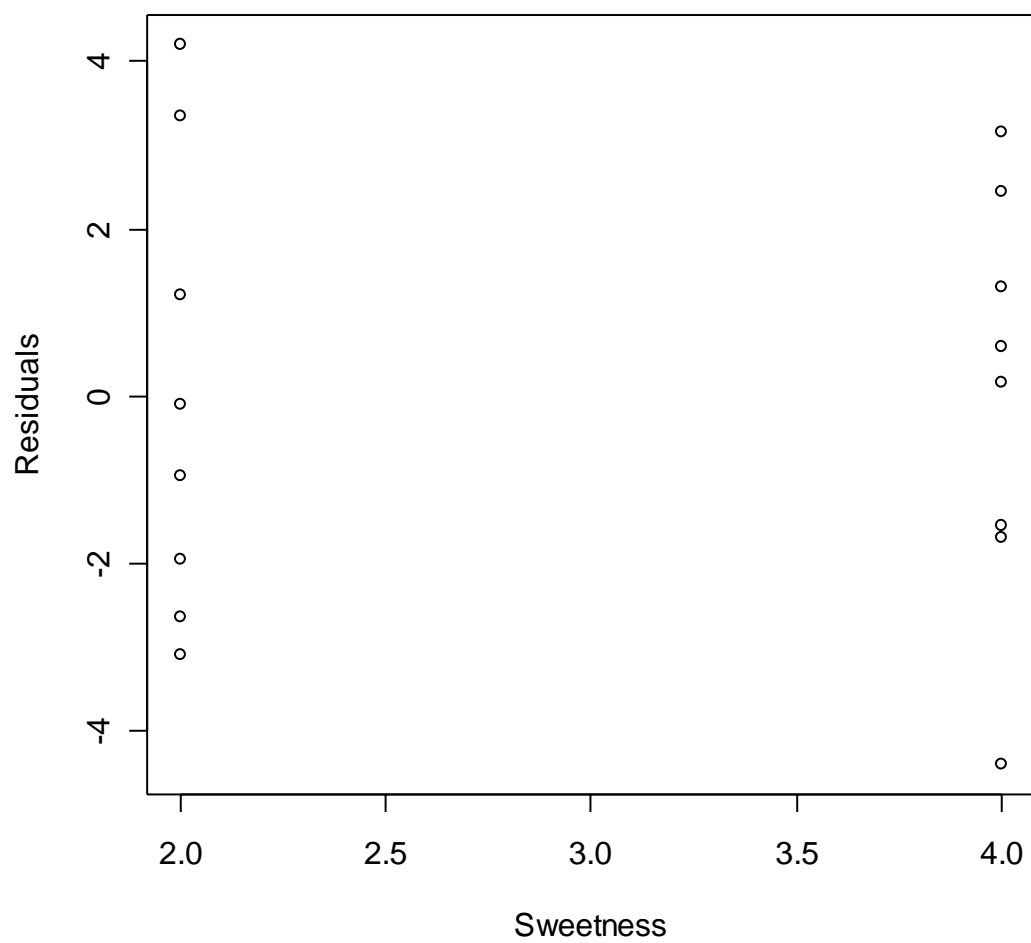
#with respect to fitted values, the residuals have constant variance.

```
> plot(mod$X1,mod.lm$res,xlab="Moisture content",ylab="Residuals")
```



#with respect to the predictor variable “Moisture content”, the residuals do not have constant variance.

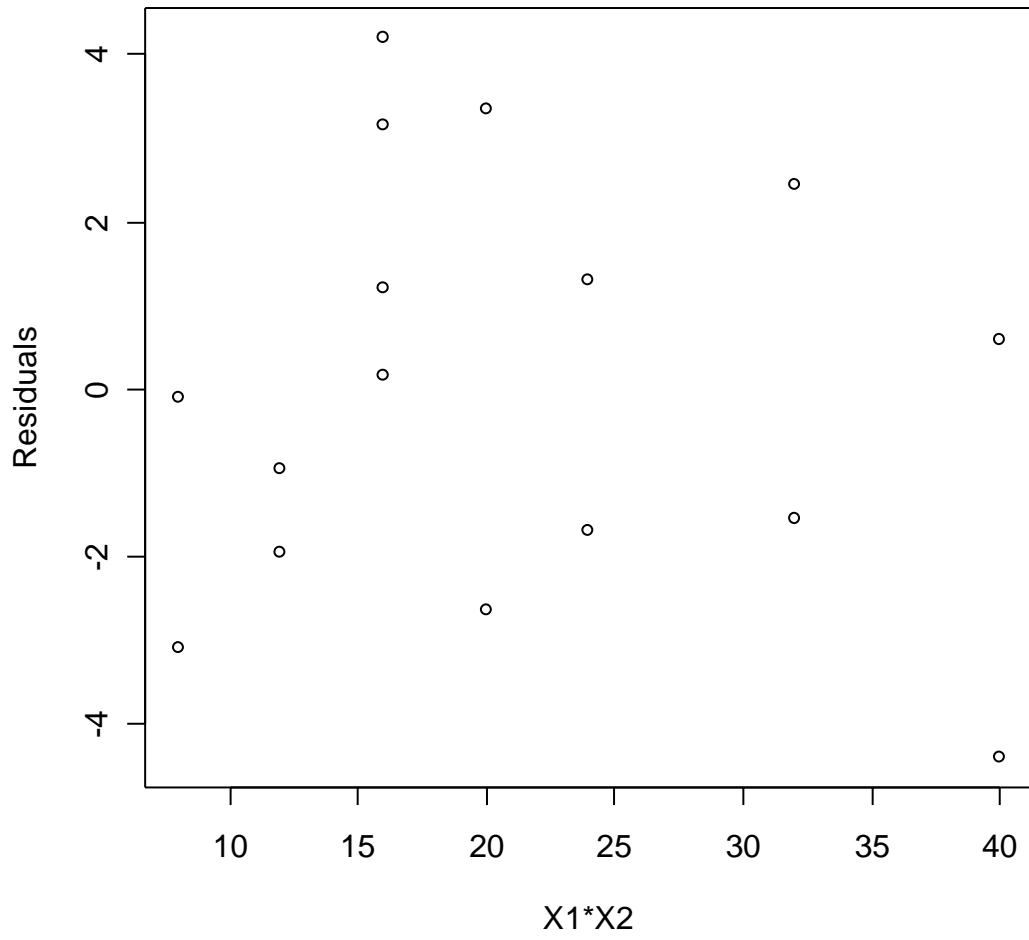
```
> plot(mod$X2,mod.lm$res,xlab="Sweetness",ylab="Residuals")
```



#with respect to the predictor variable “Sweetness”, the residuals have constant variance

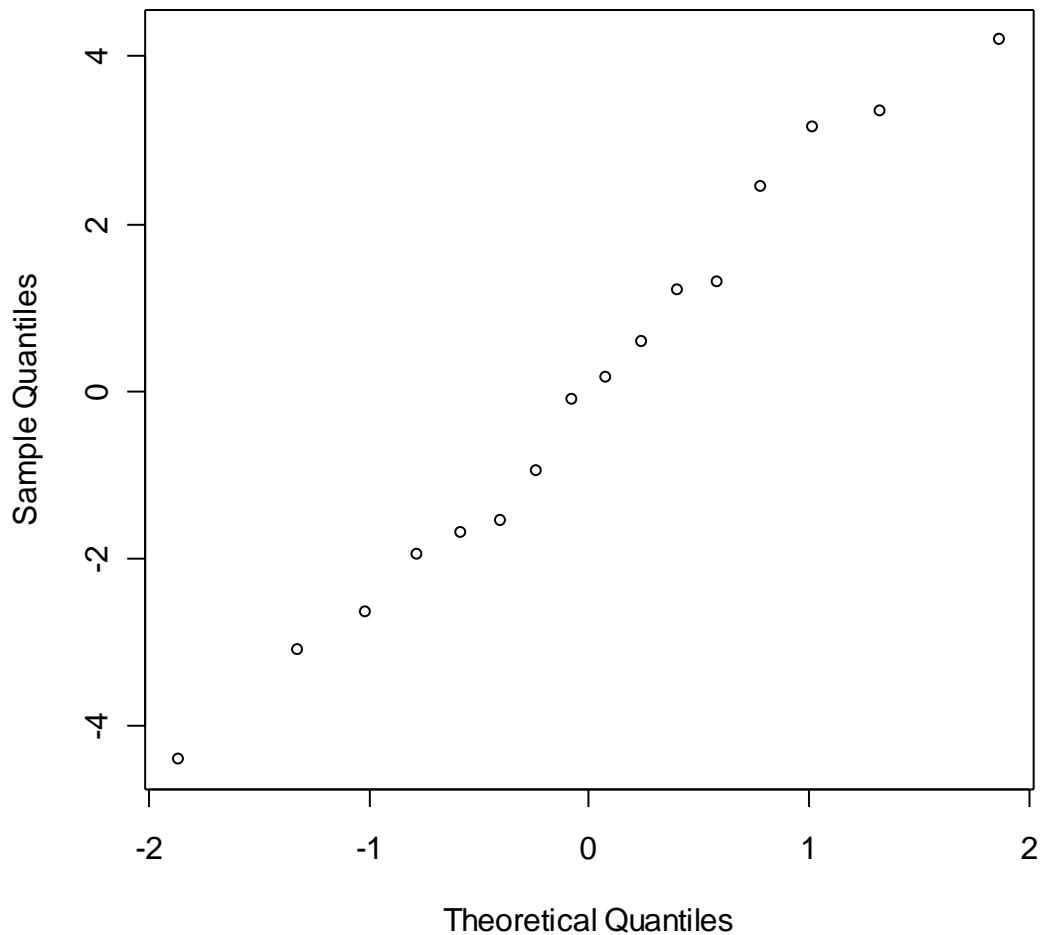
```
> X1X2<-mod$X1*mod$X2
```

```
> plot(X1X2,mod.lm$res,xlab="X1*X2",ylab="Residuals")
```



```
> qqnorm(mod.lm$res)
```

Normal Q-Q Plot



#the error terms are normally distributed.

6.6

(a)

```
> qf(0.99,2,13)
```

```
[1] 6.700965
```

#from the summary of mod.lm, we know that F-statistic is 129.1

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : \beta_1^2 + \beta_2^2 \neq 0$, if F-statistic > qf(0.99,2,13), reject H_0 , say, multiple regression model is sufficient.

#in this model, obviously F-statistic > qf(0.99,2,13), say, $\beta_1^2 + \beta_2^2 \neq 0$.

(b)

#from the summary of mod1, we get the p-value is 2.658e-09, which is much less than 0.01, indicating rejecting H_0

(c)

```
> mod<-read.table("CH06PR05.txt",head=T)
> mod.lm<-lm(mod)
> summary(mod.lm)
```

Call:

```
lm(formula = mod)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-4.400 -1.762   0.025   1.588   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

```
> X<-cbind(rep(1,16),mod$X1,mod$X2)
```

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

```
      [,1]      [,2]      [,3]
[1,] 1.2375 -8.750000e-02 -1.875000e-01
[2,] -0.0875  1.250000e-02  1.295260e-17
[3,] -0.1875  1.486543e-18  6.250000e-02
```

```
> 4.4250-qt(0.9975,14)*2.693*solve(t(X)%*%X)[2,2]
```

```
[1] 4.313049
```

```
> 4.4250+qt(0.9975,14)*2.693*solve(t(X)%*%X)[2,2]
```

```
[1] 4.536951
```

```
> 4.3750-qt(0.9975,14)*2.693*solve(t(X)%*%X)[3,3]
```

```
[1] 3.815244
```

```
> 4.3750+qt(0.9975,14)*2.693*solve(t(X)%*%X)[3,3]
```

```
[1] 4.934756
```

#the joint confidence interval for β_1, β_2 , using a 99% family confidence coefficient is [4.313049,4.536951] and

#[3.815244,4.934756]

#if repeated samples are selected and intervals estimates for β_1, β_2 are calculated, 99% of the samples would lead to

#confidence intervals that both are correct.

6.19

(a)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_a : \beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 > 0$$

#we use the statistic: $F^* = MSR/MSE$

#the decision rule to control the Type I error at $\alpha = 0.05$ is

$$\text{#if } F^* \leq F(0.95; 4, 76), \text{conclude } H_0$$
$$F^* > F(0.95; 4, 76), \text{conclude } H_a$$

```
> data<-read.table("CH06PR18.txt",head=T)
```

```
> mod<-lm(data)
```

```
> summary(mod)
```

Call:

```
lm(formula = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.18719	-0.59105	-0.09095	0.55794	2.94414

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

```
> qf(0.95,4,76)
```

```
[1] 2.492049
```

#since $F^* = 26.76 > F(0.95; 4, 76) = 2.492$, conclude H_a , say, the conclusion.

#the p-value is 7.272e-14

```

(b)
>
print(c("Beta1",mods$coef[2,1]-mods$coef[2,2]*qt(1-0.05/4,df=76),mods$coef[2,1]+mods$coef[2,2]*qt(1-0.05/4,df=76)))
[1] "Beta1"          "-0.190837292566318"  "-0.0932299944501798"
>
print(c("Beta2",mods$coef[3,1]-mods$coef[3,2]*qt(1-0.05/4,df=76),mods$coef[3,1]+mods$coef[3,2]*qt(1-0.05/4,df=76)))
[1] "Beta2"          "0.137561788345887"  "0.426471271556097"
>
print(c("Beta3",mods$coef[4,1]-mods$coef[4,2]*qt(1-0.05/4,df=76),mods$coef[4,1]+mods$coef[4,2]*qt(1-0.05/4,df=76)))
[1] "Beta3"          "-1.86584609270146"  "3.10453309962935"
>
print(c("Beta4",mods$coef[5,1]-mods$coef[5,2]*qt(1-0.05/4,df=76),mods$coef[5,1]+mods$coef[5,2]*qt(1-0.05/4,df=76)))
[1] "Beta4"          "4.7577682574292e-06"  "1.10908354946753e-05"
#95% of the samples would lead to confidence intervals that all are correct.

```

(c)

$R^2 = 0.5847$, indicating that 58.47% of the total variation can be explained by the multiple linear regression model.

6.20

#we compare the two procedures, one of which uses simultaneous Scheffe prediction limits, while the other uses #Bonferroni simultaneous prediction limits.

#simultaneous Scheffe prediction limits:

```

> data1<-read.table("CH06PR20.txt",head=T)
> b<-matrix(mods$coef[,1])
> X1<-matrix(t(cbind(1,data1[1,])))
> X2<-matrix(t(cbind(1,data1[2,])))
> X3<-matrix(t(cbind(1,data1[3,])))
> X4<-matrix(t(cbind(1,data1[4,])))
> Y1<-(t(X1)%*%b)[1,1]
> Y2<-(t(X2)%*%b)[1,1]
> Y3<-(t(X3)%*%b)[1,1]
> Y4<-(t(X4)%*%b)[1,1]
> MSE<-mods$sigma^2
> X<-cbind(1,data$X1,data$X2,data$X3,data$X4)
> spread1<-sqrt(MSE*(1+t(X1)%*%solve(t(X)%*%X)%*%X1))[1,1]
> spread2<-sqrt(MSE*(1+t(X2)%*%solve(t(X)%*%X)%*%X2))[1,1]
> spread3<-sqrt(MSE*(1+t(X3)%*%solve(t(X)%*%X)%*%X3))[1,1]

```

```

> spread4<-sqrt(MSE*(1+t(X4)% solve(t(X)% X)% X4))[1,1]
> print(Y1-sqrt(4*qt(0.95,4,76))*spread1)
[1] 12.10289
> print(Y1+sqrt(4*qt(0.95,4,76))*spread1)
[1] 19.49337
> print(Y2-sqrt(4*qt(0.95,4,76))*spread2)
[1] 12.36164
> print(Y2+sqrt(4*qt(0.95,4,76))*spread2)
[1] 19.69343
> print(Y3-sqrt(4*qt(0.95,4,76))*spread3)
[1] 12.24341
> print(Y3+sqrt(4*qt(0.95,4,76))*spread3)
[1] 19.55804
> print(Y4-sqrt(4*qt(0.95,4,76))*spread4)
[1] 12.16190
> print(Y4+sqrt(4*qt(0.95,4,76))*spread4)
[1] 19.52487
#using a 95% family confidence coefficient by simultaneous Scheffe prediction limites, the prediction intervals are
#[12.10289, 19.49337]
#[12.36164, 19.69343]
#[12.24341, 19.55804]
#[12.16190, 19.52487]

```

Bonferroni simultaneous prediction limits:

```

> print(Y1-qt(0.99375,76)*spread1)
[1] 12.80361
> print(Y1+qt(0.99375,76)*spread1)
[1] 18.79265
> print(Y2-qt(0.99375,76)*spread2)
[1] 13.05680
> print(Y2+qt(0.99375,76)*spread2)
[1] 18.99827
> print(Y3-qt(0.99375,76)*spread3)
[1] 12.93694
> print(Y3+qt(0.99375,76)*spread3)
[1] 18.86451
> print(Y4-qt(0.99375,76)*spread4)
[1] 12.86002
> print(Y4+qt(0.99375,76)*spread4)
[1] 18.82675
#using a 95% family confidence coefficient by Bonferroni simultaneous prediction limits, the prediction intervals are
#[12.80361, 18.79265]
#[13.05680, 18.99827]
#[12.93694, 18.86451]

```

```
#[12.86002, 18.82675]
```

#obviously, the latter method leads to narrower prediction intervals, indicating it is the most efficient procedure.

6.21

```
> data2<-read.table("CH06PR21.txt",head=T)
> X1<-matrix(t(cbind(1,data2[1,])))
> X2<-matrix(t(cbind(1,data2[2,])))
> X3<-matrix(t(cbind(1,data2[3,])))
> Y1<-(t(X1)%*%b)[1,1]
> Y2<-(t(X2)%*%b)[1,1]
> Y3<-(t(X3)%*%b)[1,1]
> spread1<-sqrt(MSE*(1+t(X1)%*%solve(t(X)%*%X)%*%X1))
> spread2<-sqrt(MSE*(1+t(X2)%*%solve(t(X)%*%X)%*%X2))
> spread3<-sqrt(MSE*(1+t(X3)%*%solve(t(X)%*%X)%*%X3))
> print(c(Y1-qt(0.975,76)*spread1,Y1+qt(0.975,76)*spread1))
[1] 12.85249 17.44450
> print(c(Y2-qt(0.975,76)*spread2,Y2+qt(0.975,76)*spread2))
[1] 13.24504 17.83994
> print(c(Y3-qt(0.975,76)*spread3,Y3+qt(0.975,76)*spread3))
[1] 14.53469 19.29299
```

#the fairly precisely predicted separate prediction intervals for the rental rates of these three properties are:

```
#[12.85249,17.44450]
#[13.24504,17.83994]
#[14.53469,19.29299]
```

#using a 95% family confidence coefficient by simultaneous Scheffe prediction limites, the prediction intervals

```
> print(c(Y1-sqrt(3*qf(0.95,3,76))*spread1,Y1+sqrt(3*qf(0.95,3,76))*spread1))
[1] 11.85245 18.44454
> print(c(Y2-sqrt(3*qf(0.95,3,76))*spread2,Y2+sqrt(3*qf(0.95,3,76))*spread2))
[1] 12.24437 18.84061
> print(c(Y3-sqrt(3*qf(0.95,3,76))*spread3,Y3+sqrt(3*qf(0.95,3,76))*spread3))
[1] 13.49843 20.32925
```

#using a 95% family confidence coefficient by Bonferroni simultaneous prediction limits, the prediction intervals

```
> print(c(Y1-qt(1-0.05/6,76)*spread1,Y1+qt(1-0.05/6,76)*spread1))
[1] 12.32630 17.97069
> print(c(Y2-qt(1-0.05/6,76)*spread2,Y2+qt(1-0.05/6,76)*spread2))
[1] 12.71852 18.36646
> print(c(Y3-qt(1-0.05/6,76)*spread3,Y3+qt(1-0.05/6,76)*spread3))
[1] 13.98944 19.83824
```

6.27

```
> data<-read.table("CH06PR27.txt",head=T)
> mod<-lm(data)
> mods<-summary(mod)
> b<-matrix(mods$coef[,1])
> b
```

```
      [,1]
[1,] 33.9321033
[2,]  2.7847614
[3,] -0.2644189
```

```
> X<-cbind(1,data$X1,data$X2)
> H<-X%*%solve(t(X)%*%X)%*%t(X)
> H
```

```
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.23143293 0.25167585 0.21178735 0.1488684 -0.05475543 0.21099091
[2,] 0.25167585 0.31240459 0.09437844 0.2662773 -0.14787283 0.22313666
[3,] 0.21178735 0.09437844 0.70442026 -0.3191744 0.10446672 0.20412159
[4,] 0.14886839 0.26627729 -0.31917435 0.6142563 0.14143492 0.14833743
[5,] -0.05475543 -0.14787283 0.10446672 0.1414349 0.94039955 0.01632707
[6,] 0.21099091 0.22313666 0.20412159 0.1483374 0.01632707 0.19708635
```

```
> Y<-matrix(data$Y)
> e<-Y-X%*%b
> e
```

```
      [,1]
[1,] -2.699608
[2,] -1.229973
[3,] -1.637353
[4,] -1.329860
[5,] -0.089998
[6,]  6.986792
```

```
> J<-rbind(1,1,1,1,1,1)%*%cbind(1,1,1,1,1,1)
> SSR<-(t(Y)%*%(H-1/6*J)%*%Y)[1,1]
> SSR
[1] 3009.926
```

```
> VCM<-(mods$sigma)^2*solve(t(X)%*%X)
> VCM
      [,1]      [,2]      [,3]
[1,] 715.47114 -34.1589166 -13.5949371
[2,] -34.15892  1.6616664  0.6440674
[3,] -13.59494  0.6440674  0.2624678
```

$S^2\{b\}$ is the above variance-covariance matrix.

```
> Xh1<-10
```

```
> Xh2<-30
```

```
> Xh<-rbind(1,Xh1,Xh2)
```

```
> Yh<-(t(Xh)%*%b)[1,1]
```

```
> Yh
```

```
[1] 53.84715
```

```
#  $\hat{Y}_h = 53.84715$ 
```

```
> print(S2Yh<-((mods$sigma)^2*t(Xh)%*%solve(t(X)%*%X)%*%Xh)[1,1])
```

```
[1] 5.42462
```

```
#  $S^2\{\hat{Y}_h\} = 5.42462$ 
```