# Assignement3

*Zimeng Ming V00844078*

*2019/3/14*

Question 1:

```
#First of all, Read the data in the system.
LungCancer<-read.csv("LungCancer.csv", header = T)

#using table() to find contingency
#table(LungCancer)
#for look nicely
CancerTable<-table(LungCancer)
colnames(CancerTable)<-c("non-smoker","smoker")
rownames(CancerTable)<-c("NOT Lung Cancer", "Lung Cancer")
CancerTable
```

```
##                  Smoker
## Case              non-smoker smoker
##   NOT Lung Cancer         60    650
##   Lung Cancer             22    687
```

a): For here, is clear that is the sign test so here we use the distribution of chi-square

null-hypothesis H0: The Smoking and Lung Cancer are independent HA:The Smoking and Lung Cancer are not independent

```
chisq.test(LungCancer$Case,LungCancer$Smoker)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  LungCancer$Case and LungCancer$Smoker
## X-squared = 17.664, df = 1, p-value = 2.636e-05
```

As we can see from above, since the p-value is very small,less then 0.001, there is significant eveidence against H0, which we may conclude that The Smoking and Lung Caner are not independent.

b): The assumption is there is association between smoking and lung cancer. According to the result from (a), the assumption is valid.

Question2: first get the AIC function:

```
#for fit1 is fit1<-lm(y,x)
myAIC<-function(fit1){
  #get the sum of squares
  b<-anova(fit1)
  #using the AIC function
  AIC<-length(fit1$residuals)*log((b$"Sum Sq"[2])/(length(fit1$residuals)))+2*(length
(fit1$coefficients)+1)
  return(AIC)

}
```

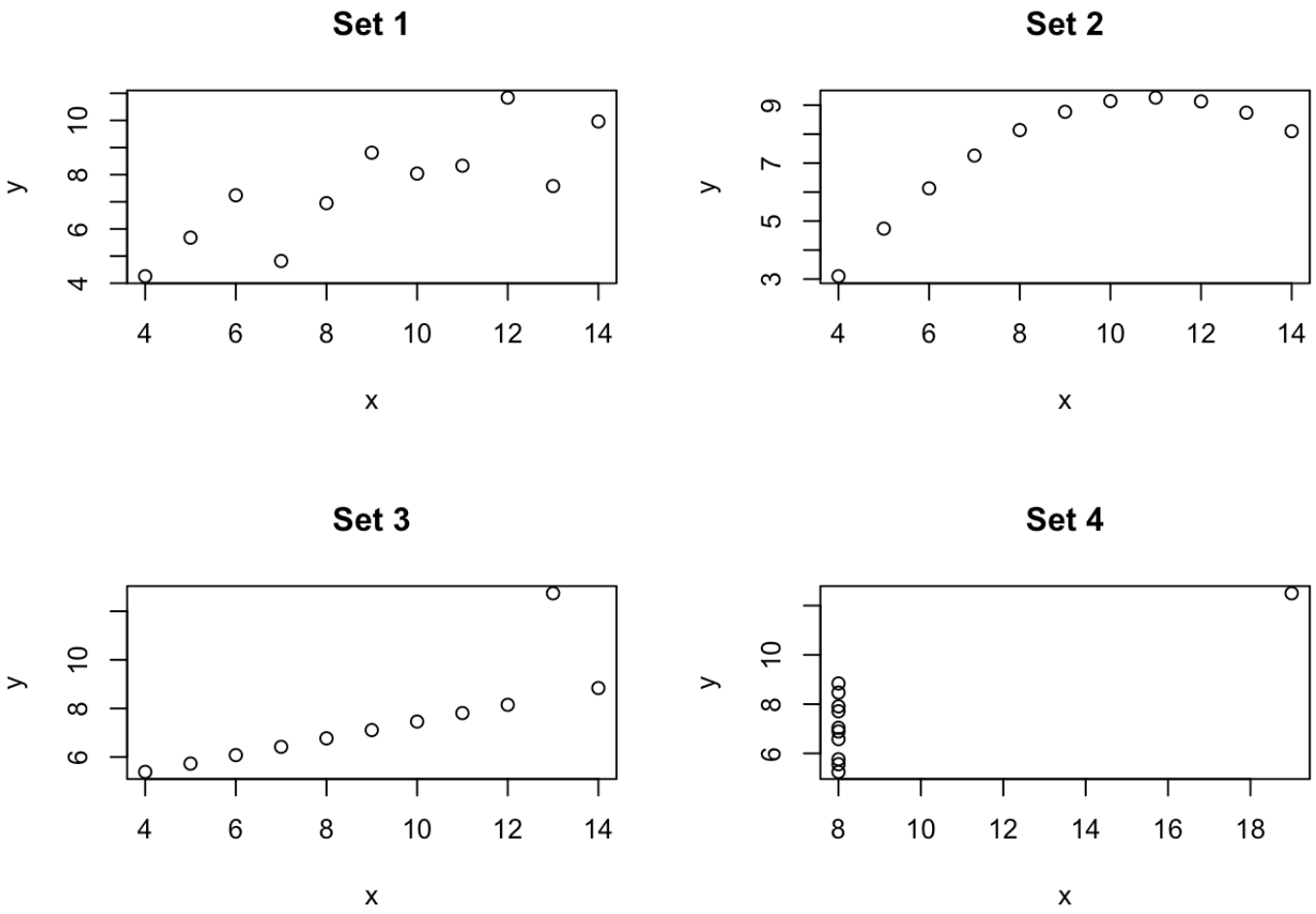
Question3: a):

```
#first read the data
anscombe<-read.csv('anscombe.csv', header=T)
#in order to see the data, plot the data out but commend for the final output

#anscombe
```

Then get the plot

```
par(mfrow=c(2,2))
plot(anscombe[,1],anscombe[,2],main="Set 1",xlab="x", ylab="y")
plot(anscombe[,3],anscombe[,4] ,main="Set 2" ,xlab="x", ylab="y")
plot(anscombe[,5],anscombe[,6] ,main="Set 3" ,xlab="x", ylab="y")
plot(anscombe[,7],anscombe[,8] ,main="Set 4" ,xlab="x", ylab="y")
```




According to above plot, we can notice that for set1, The data are looks like a linear increasing model for x and y. it seems not or a few outliers. For Set2, it is looks like a quadratic model for x and y. For Set3, it is linear increasing model for x and y but with an significantly outlier in the graph. For Set 4, it is not an linear outlier in the graph. becuase there are so many y represent a x data. Howevere, there is an outlier on the right top corner it may have an significant influence on the modeling analysis.

B.

```
#first set the linear regression for the 4 datasets
fit1<-lm(anscombe[,2]~anscombe[,1])
fit2<-lm(anscombe[,4]~anscombe[,3])
fit3<-lm(anscombe[,6]~anscombe[,5])
fit4<-lm(anscombe[,8]~anscombe[,7])
```

```
#then summary the r square

cat("     Data Set              R^2 Values         AIC Values","\n","      1
",summary(fit1)$r.squared, "        ",myAIC(fit1),"\n","      2                ",su
mmary(fit2)$r.squared,"         ",myAIC(fit2),"\n","      3                ",summar
y(fit3)$r.squared,"         ",myAIC(fit3),"\n","      4                ",summary(fi
t4)$r.squared,"        ",myAIC(fit4),"\n")
```

```
##      Data Set             R^2 Values         AIC Values
##        1                   0.6665425          8.464726
##        2                   0.666242           8.475592
##        3                   0.666324           8.459531
##        4                   0.6667073          8.448569
```

According to the result from above, we can see that the R squared values and AIC Values are very close for 4 data sets. According to the analysis from a), we can notice that although that 4 data sets has different conditions for the data, such as outliers, shapes etc, but the linear association between x and y are very close in such different conditions. For instance, graph 3 and 4 which is set 3 and 4 has an siginificant outliers, but the linear association are close to the set which do not have a clearly outliers.

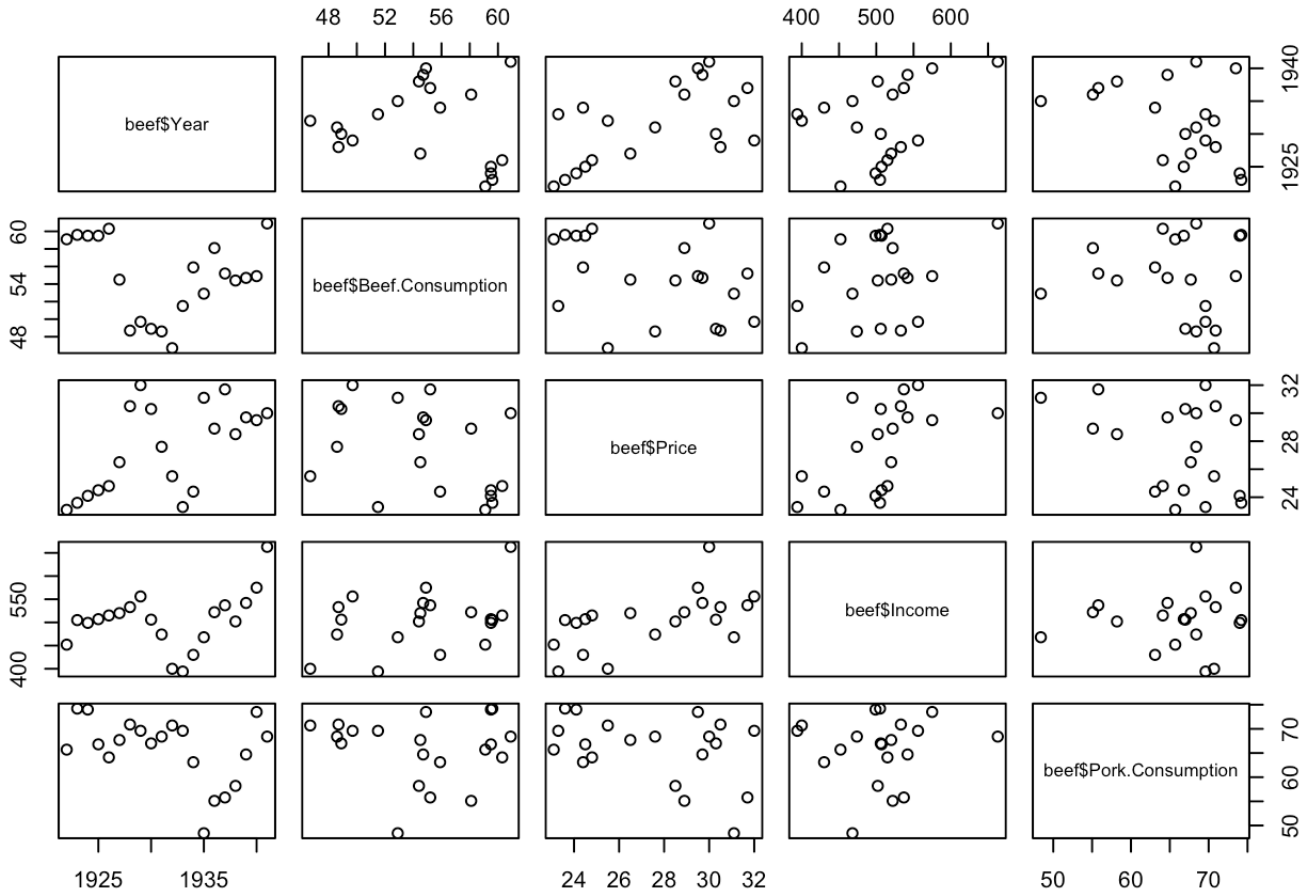
In terms of AIC Valurs, It is similar but if there is an outlier in the data sets, we can see the linear association for x and y are a little bit lower then the set do not have an clearly outliers.

---

Question4:

```
#First load the data in the system.
beef<-read.table("beef.txt",header = T)
#beef
```

```
#first plot the data  and see the relations for data sets.
#here the par
par(mfrow=c(4,3))
pairs(~beef$Year+beef$Beef.Consumption+beef$Price+beef$Income+beef$Pork.Consumption)
```




```
#pairs(cbind(beef$Year,beef$Beef.Consumption,beef$Price,beef$Income,beef$Pork.Consumption))
```

As we can see from the above, beef consumption is looks has an clearly increase linear association with Income. It seems might have an increase linear association with price but it clearly do not have any linear relationship with pork consumption.

know doing the linear model fit

```
#for here we using Price as x1, Income as x2,Pork Consumption as X3 and Beef Consumption as Y
x1<-beef$Price
x2<-beef$Income
x3<-beef$Pork.Consumption
y<-beef$Beef.Consumption

#then we using the model
z1<-x1*x1
z2<-x2*x2
z3<-x3*x3

fit_q4<-lm(y~x1+x2+x3+z1+z2+z3+x1:x2+x1:x3+x2:x3+x1:x2:x3)
summary(fit_q4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + z1 + z2 + z3 + x1:x2 + x1:x3 +
##     x2:x3 + x1:x2:x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41800 -0.28785 -0.08738  0.35150  1.39377
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.148e+02  5.305e+02   1.724    0.119
## x1          -3.078e+01  1.832e+01  -1.680    0.127
## x2          -1.264e+00  9.592e-01  -1.318    0.220
## x3          -1.276e+01  8.408e+00  -1.518    0.163
## z1           4.729e-02  6.713e-02   0.704    0.499
## z2          -1.131e-04  1.187e-04  -0.952    0.366
## z3           1.953e-02  1.243e-02   1.571    0.151
## x1:x2        5.199e-02  3.239e-02   1.605    0.143
## x1:x3        3.491e-01  2.728e-01   1.280    0.233
## x2:x3        1.953e-02  1.411e-02   1.384    0.200
## x1:x2:x3    -6.906e-04  5.086e-04  -1.358    0.208
##
## Residual standard error: 0.9818 on 9 degrees of freedom
## Multiple R-squared:  0.9773, Adjusted R-squared:  0.9521
## F-statistic: 38.73 on 10 and 9 DF,  p-value: 3.532e-06
```

Here we can see that the P-value are very high. z1 and z2 are closing to 0.5 so here we only using income.

```
fit2_q4<-lm(y~x1+x2+x3+z2)
summary(fit2_q4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + z2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6085 -0.6399  0.2083  0.8020  2.2742
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.038e+01  1.540e+01   5.219 0.000104 ***
## x1          -1.856e+00  1.485e-01 -12.501 2.47e-09 ***
## x2           1.234e-01  5.604e-02   2.202 0.043756 *
## x3          -4.094e-01  5.534e-02  -7.398 2.23e-06 ***
## z2          -3.890e-05  5.383e-05  -0.723 0.480992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.393 on 15 degrees of freedom
## Multiple R-squared:  0.9238, Adjusted R-squared:  0.9035
## F-statistic: 45.48 on 4 and 15 DF,  p-value: 3.257e-08
```

However, we can see from here. That the p-value are very low and R-squared are close to 1. But the p-value for x2 and z2 are also very high. So We only use x1+x2+x3

```
fit3_q4<-lm(y~x1+x2+x3)
summary(fit3_q4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44996 -0.87212  0.04715  0.73206  2.47242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.813646   5.266047  17.245 9.28e-12 ***
## x1          -1.849850   0.145990 -12.671 9.32e-10 ***
## x2           0.083190   0.006868  12.113 1.80e-09 ***
## x3          -0.415085   0.053945  -7.695 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 16 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9064
## F-statistic: 62.33 on 3 and 16 DF,  p-value: 4.799e-09
```

Here the p-value for all of the x1 x2 x3 are very low and no strange p-vlaues. for the further check and fit the model, we using the following method to check:

```
summary.lm(fit3_q4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44996 -0.87212  0.04715  0.73206  2.47242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.813646   5.266047  17.245 9.28e-12 ***
## x1          -1.849850   0.145990 -12.671 9.32e-10 ***
## x2           0.083190   0.006868  12.113 1.80e-09 ***
## x3          -0.415085   0.053945  -7.695 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 16 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9064
## F-statistic: 62.33 on 3 and 16 DF,  p-value: 4.799e-09
```

Also we using the AIC function to test residual sums of squares

```
#anova for the AIC
anova(fit3_q4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1  55.172  55.172  29.316 5.731e-05 ***
## x2         1 185.286 185.286  98.455 3.061e-08 ***
## x3         1 111.423 111.423  59.206 9.155e-07 ***
## Residuals 16  30.111   1.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#AIC function
AIC(fit3_q4)
```

```
## [1] 74.94075
```

```
#extract AIC
extractAIC(fit3_q4,0)
```

```
## [1]  4.00000 16.18321
```

```
#backward selection
step(fit3_q4, direction = "backward")
```

```
## Start:  AIC=16.18
## y ~ x1 + x2 + x3
##
##        Df Sum of Sq    RSS    AIC
## <none>               30.11 16.183
## - x3    1    111.42 141.53 45.136
## - x2    1    276.12 306.24 60.572
## - x1    1    302.16 332.27 62.204
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Coefficients:
## (Intercept)           x1           x2           x3
##    90.81365     -1.84985      0.08319     -0.41508
```

```
#then we get to the final model
fit4_q4<-lm(y~x1+x2+x3)
summary(fit4_q4)
```

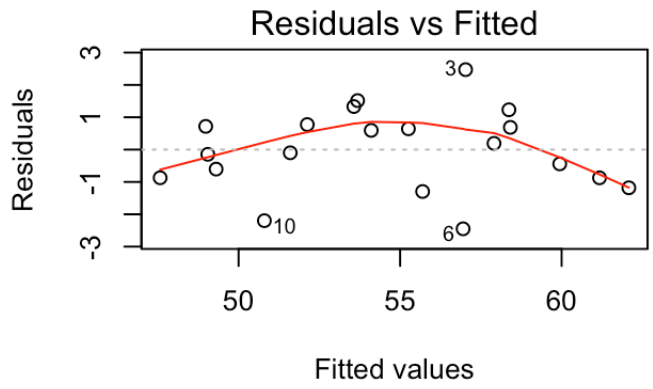
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44996 -0.87212  0.04715  0.73206  2.47242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.813646   5.266047  17.245 9.28e-12 ***
## x1          -1.849850   0.145990 -12.671 9.32e-10 ***
## x2           0.083190   0.006868  12.113 1.80e-09 ***
## x3          -0.415085   0.053945  -7.695 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 16 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9064
## F-statistic: 62.33 on 3 and 16 DF,  p-value: 4.799e-09
```

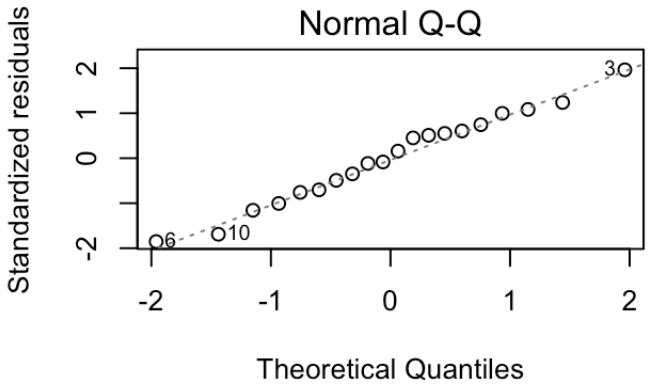
```
#Partial F-tests
anova(fit3_q4,fit4_q4)
```

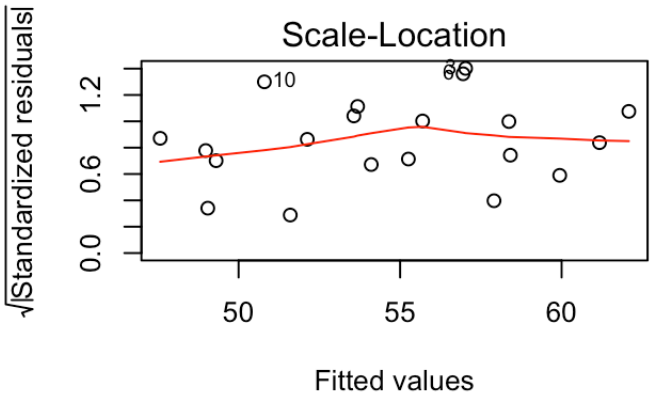
```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     16 30.111
## 2     16 30.111  0         0
```

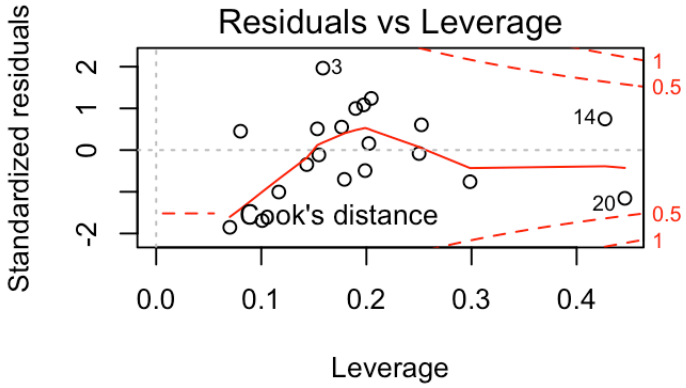
Actuatlly, it is the same for fit3_q4 and fit4_q4

```
#residual plots
par(mfrow=c(2,2))
plot(fit4_q4)
```

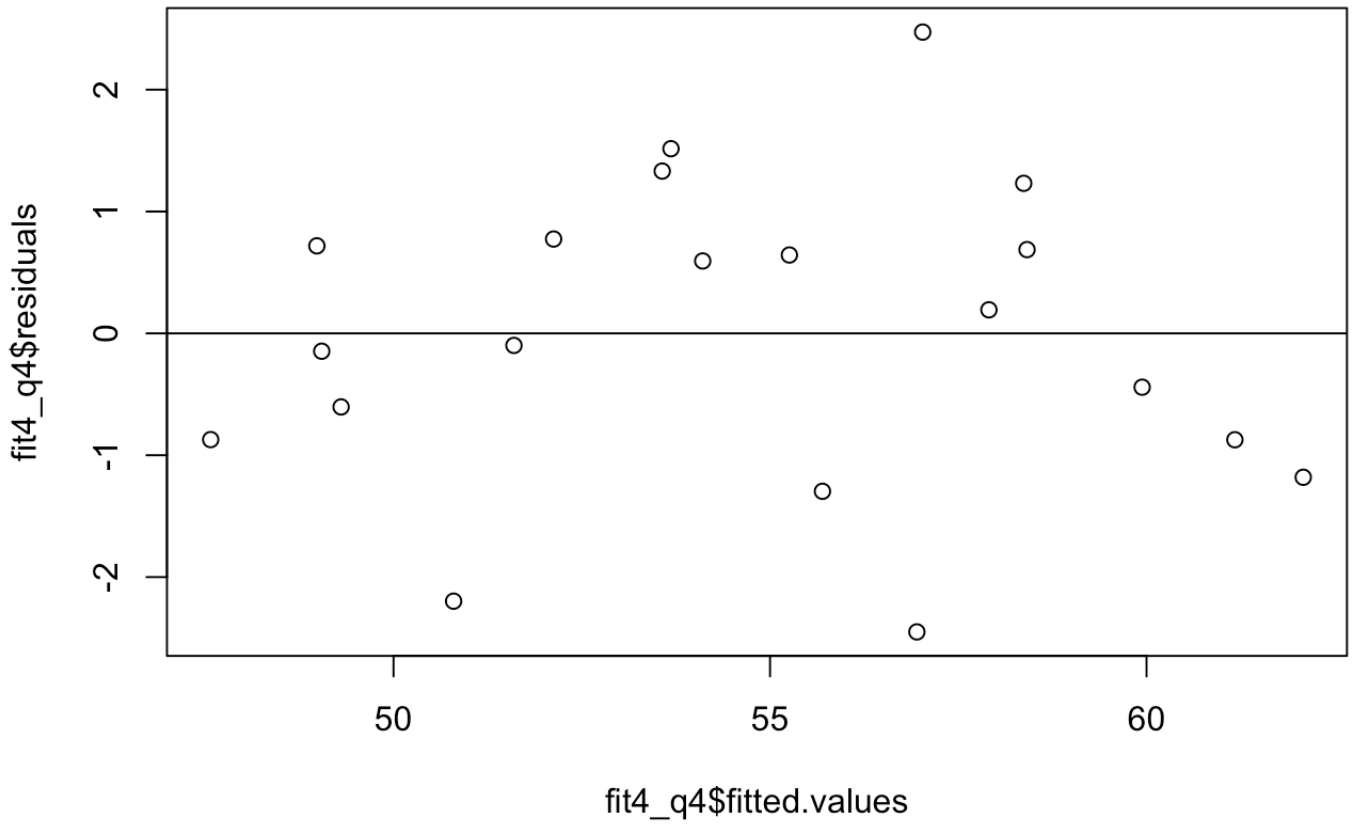


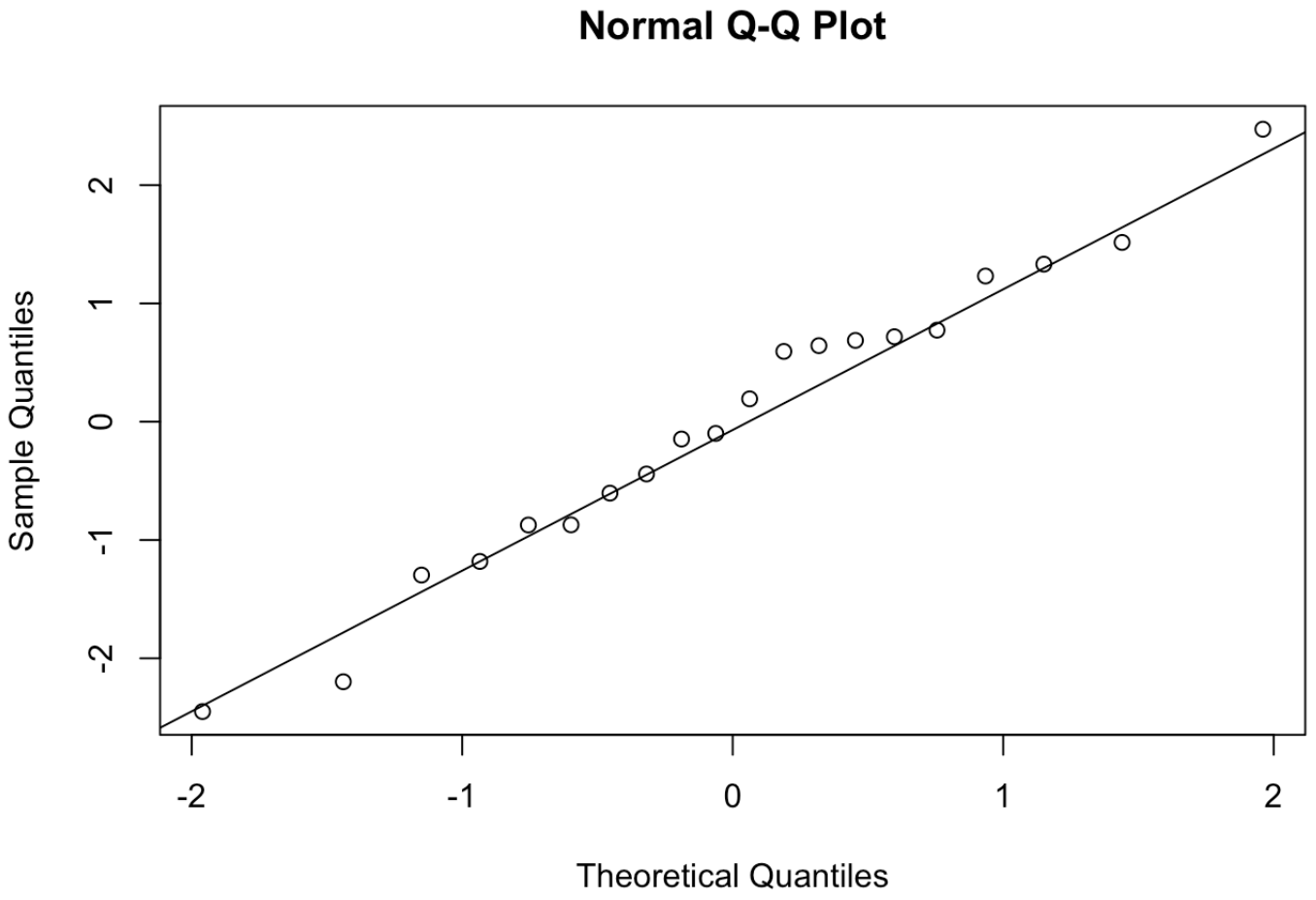




```
plot(fit4_q4$fitted.values,fit4_q4$residuals)
abline(h=0)
```




```
#for qqplot
qqnorm(fit4_q4$residuals)
qqline(fit4_q4$residuals)
```




```
#for confidence interval
confint(fit4_q4)
```

```
##                   2.5 %       97.5 %
## (Intercept) 79.65012611 101.97716683
## x1          -2.15933362  -1.54036544
## x2           0.06863039   0.09774867
## x3          -0.52944332  -0.30072601
```

In conclusion,The model for the question is Beef.Consumption=90.813646-1.849850*Price +0.083190*Income-0.415085*Pork.Consumption

---

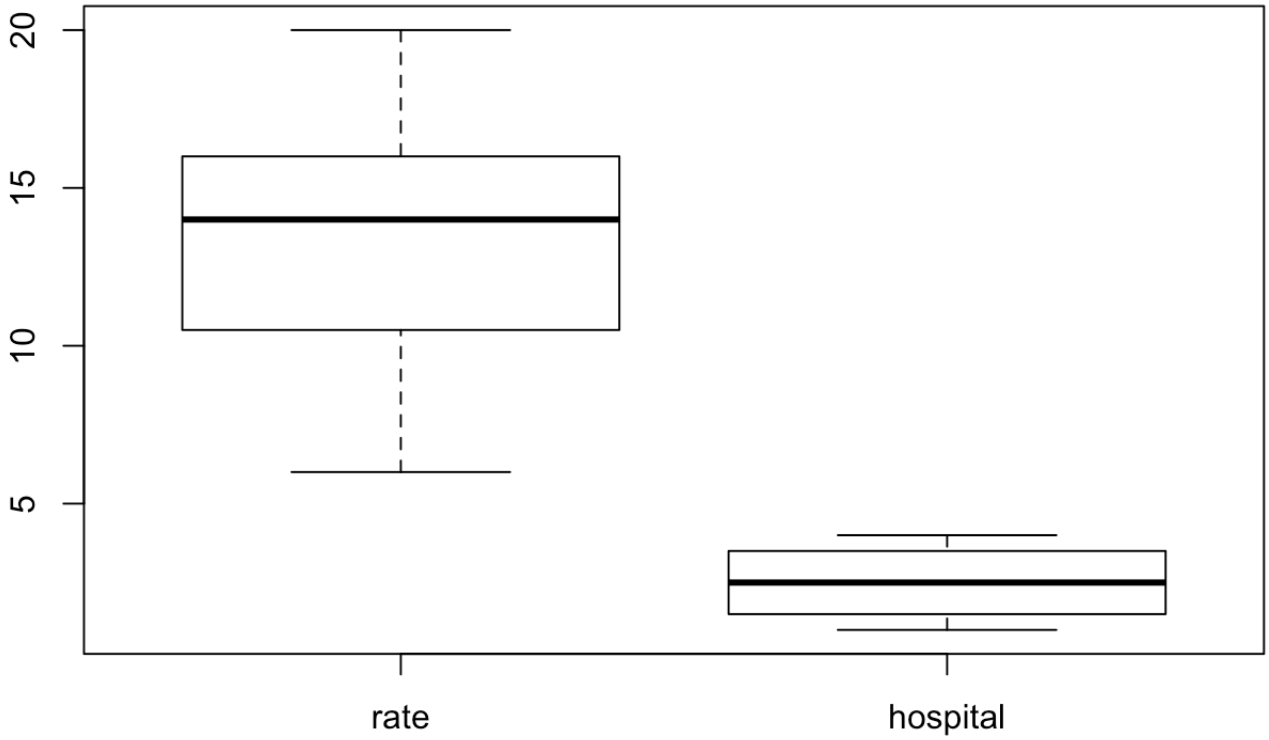
Question5

```
#read the data

Hospital<-read.csv("hospital.csv",header = T)
Hospital
```

```
##   H1 H2 H3 H4
## 1 10 18 14  9
## 2 12 15 16  8
## 3 16 14 16 10
## 4  9 18 14 11
## 5 NA 12 17  6
## 6 NA 13 18 NA
## 7 NA 12 20 NA
## 8 NA 14 NA NA
```

```
stHospital<-stack(Hospital)
names(stHospital)<-c("rate","hospital")
boxplot(stHospital)
```




a)

:The model we selected is one way anova. Because the data has rate and hospital two consideration.

b): H0:The true average ratings for each of the hospitals are equal HA:The true average ratings for each of the hospitals are not equal

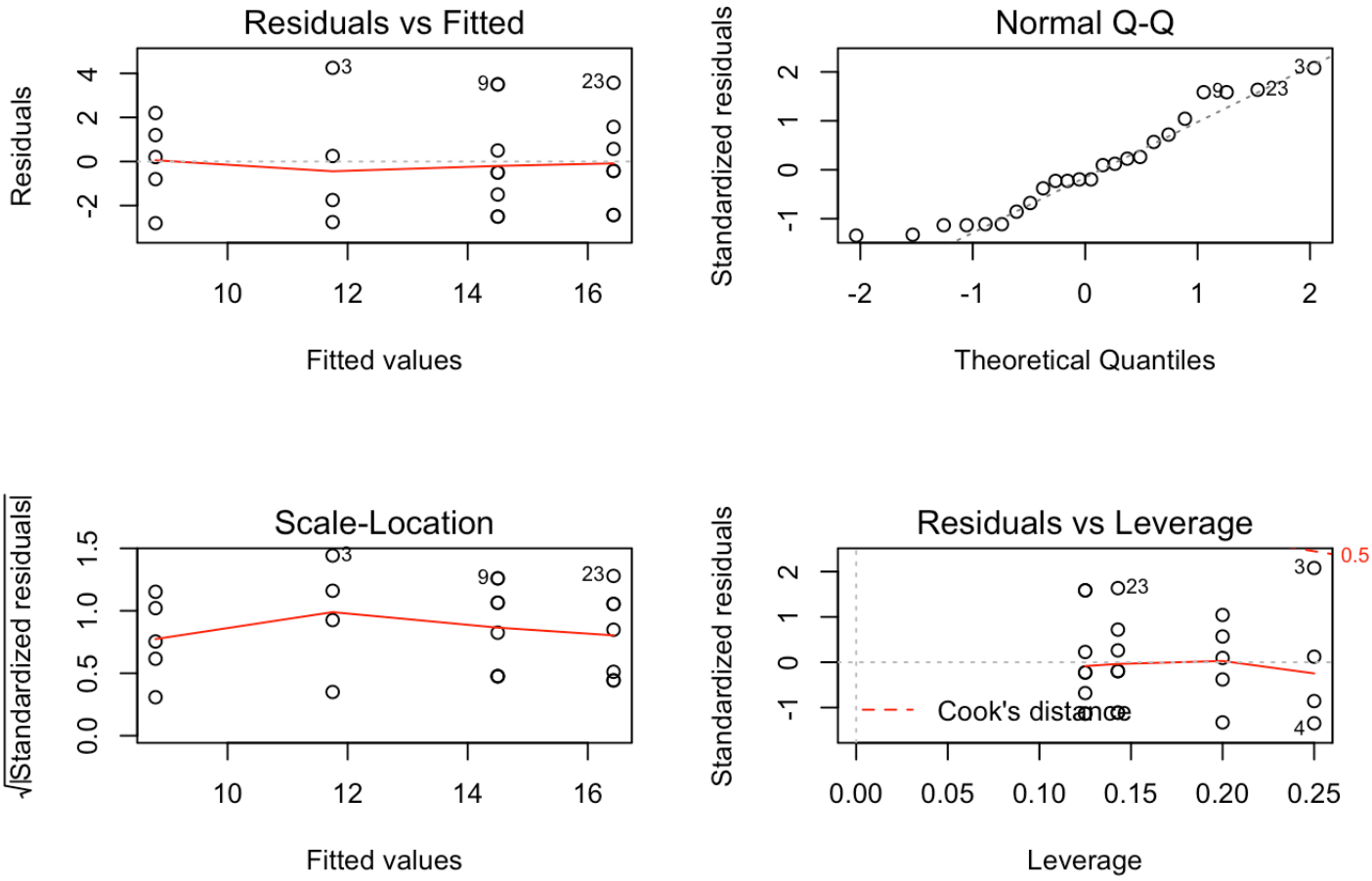
```
aov1<-aov(rate~hospital,data=stHospital)
summary(aov1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## hospital     3  190.6   63.52   11.42 0.00014 ***
## Residuals   20  111.3    5.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 8 observations deleted due to missingness
```

As we can see from above, The p-value is very small and less then 0.001, there is an strong evidence agianst H0,so there is significant evidence shows that the true average ratings for each of the hospitals are not equal.

c):

```
#plot the data
par(mfrow=c(2,2))
plot(aov1)
```



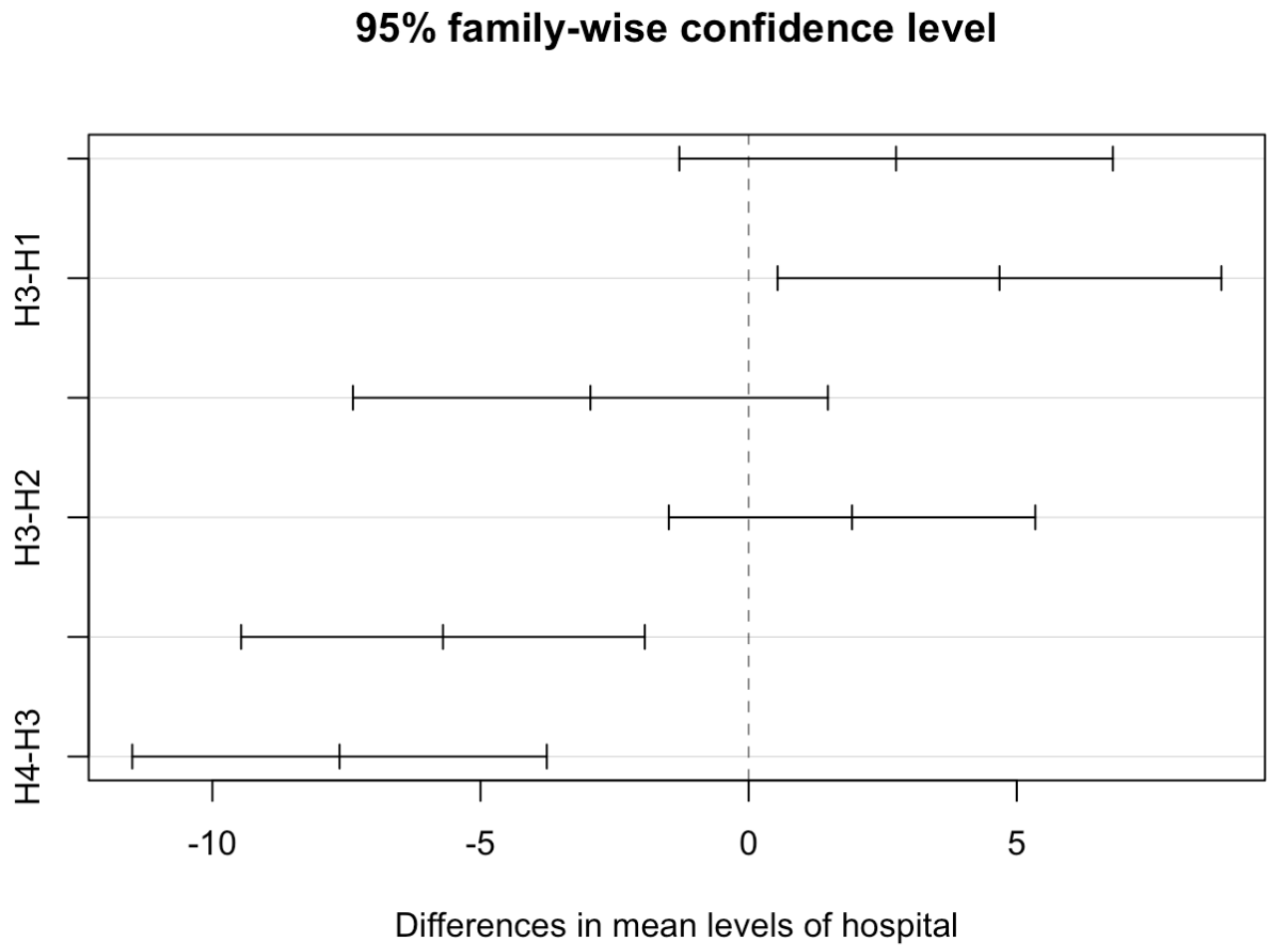

So as we can see that the plot information are support the aov result that the normal qq line is linear and residuals vs fitted are an straight line on 0 resuduals.

D):

```
#for the Tukey
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rate ~ hospital, data = stHospital)
##
## $hospital
##            diff        lwr       upr     p adj
## H2-H1  2.750000  -1.292700  6.792700 0.2579686
## H3-H1  4.678571   0.540736  8.816407 0.0231842
## H4-H1 -2.950000  -7.378556  1.478556 0.2742946
## H3-H2  1.928571  -1.488134  5.345277 0.4118317
## H4-H2 -5.700000  -9.463549 -1.936451 0.0020977
## H4-H3 -7.628571 -11.494132 -3.763011 0.0001143
```

```
plot(TukeyHSD(aov1))
```




It is much more easier to see from the plot, The confidence interval of H3-H1,H4-H2 and H4-H3 do not contain 0, so there is an difference in mean in this 3 caculation. so those above has mean differ.