MATH 564 - Assignment3

Mohammed Wasim R D(A20497053)

Problem1

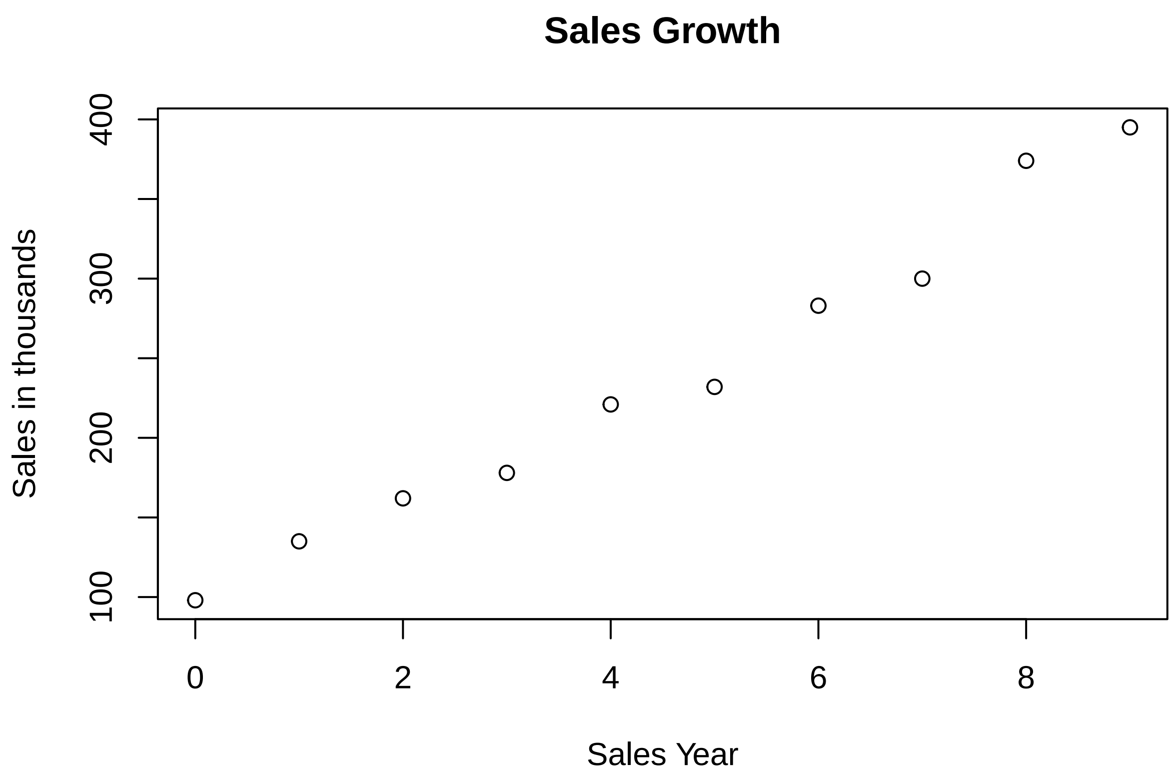
Reading data and renaming columns

```
data1<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%203%20Data%20Sets/CH03PR1
colnames(data1)[1] ="sales"
colnames(data1)[2]="year"
data1
```

```
##    sales year
## 1     98    0
## 2    135    1
## 3    162    2
## 4    178    3
## 5    221    4
## 6    232    5
## 7    283    6
## 8    300    7
## 9    374    8
## 10   395    9
```

a)Scatter Plot

```
plot(sales~year, data=data1, main='Sales Growth',xlab = "Sales Year", ylab = "Sales in thousands")
```



##

Linear relation appears adequate here

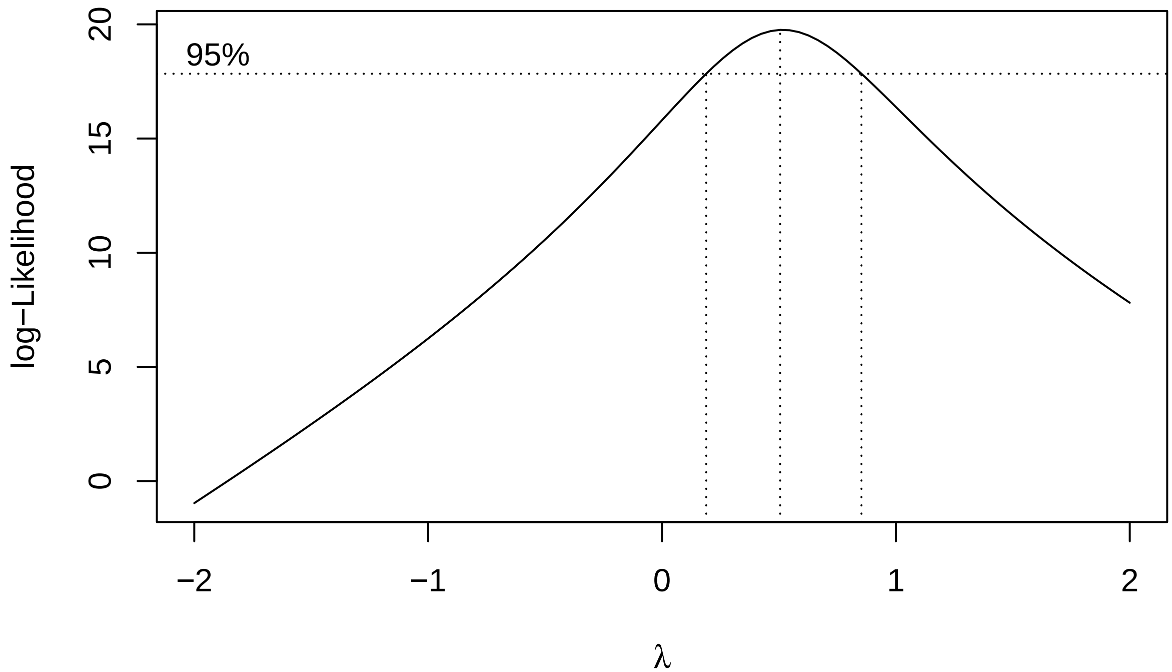
---

b) Boxplot

```
library(MASS)
lm <- lm(sales~year,data=data1)
summary(lm)
```

```
##
## Call:
## lm(formula = sales ~ year, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.049  -9.177   2.446   9.814  22.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   91.564      8.814   10.39 6.38e-06 ***
## year          32.497      1.651   19.68 4.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 8 degrees of freedom
## Multiple R-squared:  0.9798, Adjusted R-squared:  0.9772
## F-statistic: 387.4 on 1 and 8 DF,  p-value: 4.62e-08
```

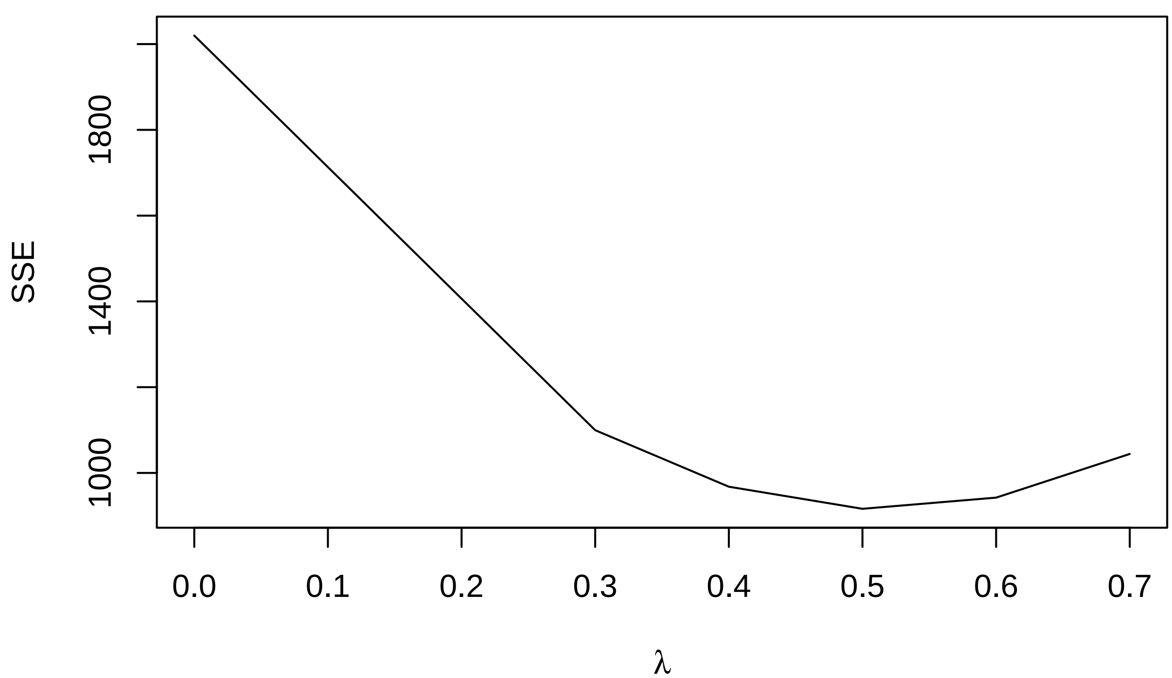
```
bc <- boxcox(lm)
```




```
library(ALSM)
```

```
## Loading required package: leaps
## Loading required package: SuppDists
## Loading required package: car
## Loading required package: carData
```

```
sse<-boxcox.sse(data1$year,data1$sales,l= seq(0.3,0.7,by=0.1))
```




```
sse
```

```
##   lambda       SSE
## 6    0.0 2019.8767
## 1    0.3 1099.7093
## 2    0.4  967.9088
## 3    0.5  916.4048
## 4    0.6  942.4498
## 5    0.7 1044.2384
```

```
peak_val = bc$x[which.max(bc$y)]
peak_val
```

```
## [1] 0.5050505
```

The suggested transformation is where lambda=0.5050505

---

c) Linear Regression function for transformed data

```
sales.prime <- sqrt(data1$sales)
lm1<-lm(sales.prime~year,data=data1)
lm1
```

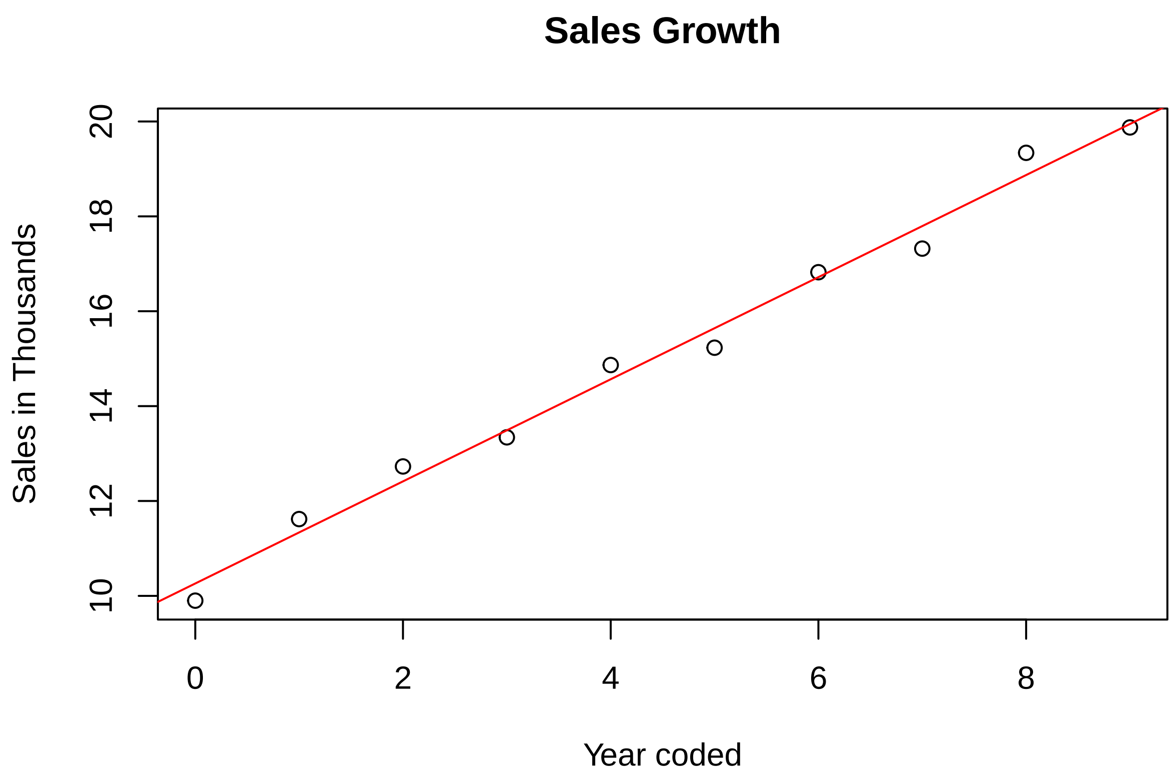
```
##
## Call:
## lm(formula = sales.prime ~ year, data = data1)
##
## Coefficients:
## (Intercept)         year
##      10.261        1.076
```

Y^=10.26093+1.076X is the regression function for transformed data

---

d) Regression Line for transformed data

```
plot(sales.prime ~ year, data=data1, main='Sales Growth', xlab="Year coded", ylab = "Sales in Thousands
abline(lm1, col='red')
```



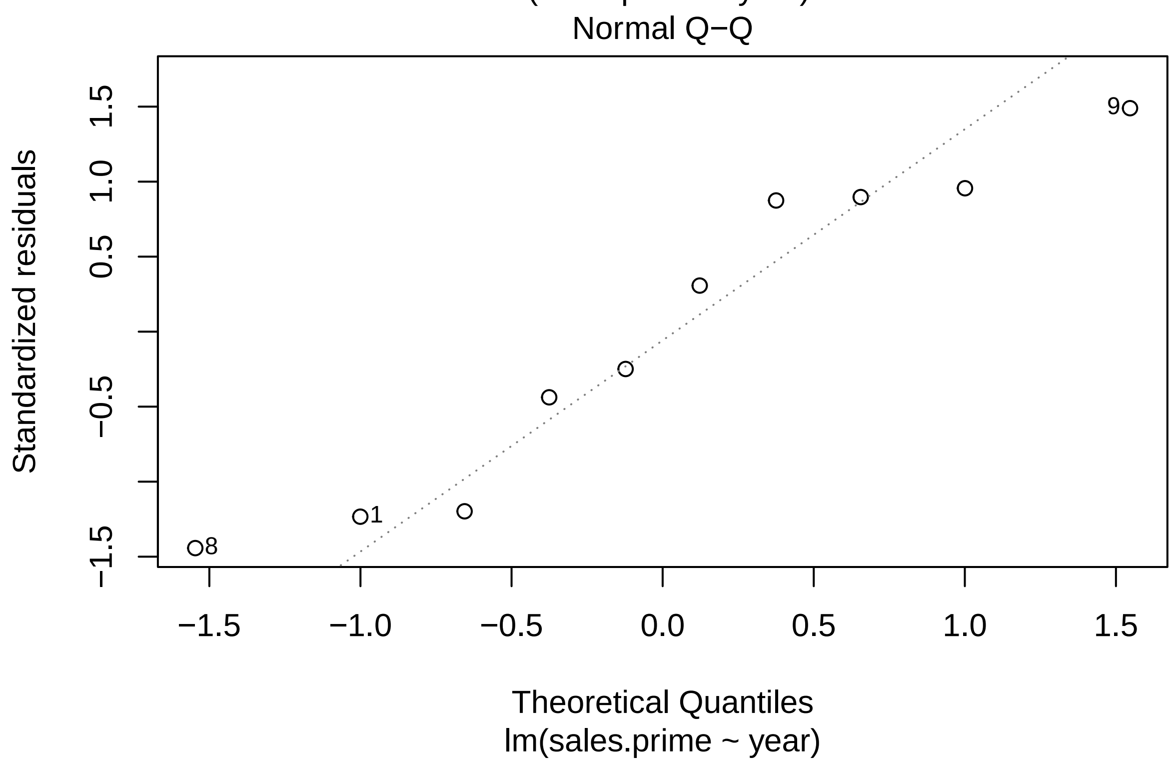
##

Linear Regression line for transformed data appears to be good fit

---

e) Obtain residuals and plot them against fitted values

```
sales.res <- lm1$residuals
sales.fitted <- lm1$fitted.values
plot(lm1, which= c(1,2))
```



The residuals v. Fitted values plot points to the error in the linear regression lining up well with the differences between the expected values and the observed values. Additionally, the sum of the residuals is zero which supports the use of this transformation on the data for linear regression analysis.

The Q-Q plot helps us determine if the standardized residuals from the linear model are normally distributed. The points do not perfectly line up along the line y=x; however they appear to generally follow this line therefore, we conclude that the residuals are normally distributed.

---

f) Estimated Regression function

```r
lm1
```

```
##
## Call:
## lm(formula = sales.prime ~ year, data = data1)
##
## Coefficients:
## (Intercept)         year
##      10.261        1.076
```

Estimated Regression function is $\hat{Y}$=10.261+1.076X

——————————Problem 2———————————————————

a)confidence intervals for mean of interest using working hotelling procedure and 95 percent family confidence co efficient

```
data2= read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR27
colnames(data2)[1] ="musclemass"
colnames(data2)[2]="age"
head(data2)
```

```
##   musclemass age
## 1        106  43
## 2        106  41
## 3         97  47
## 4        113  46
## 5         96  45
## 6        119  41
```

```
linear <- lm(musclemass ~ age, data=data2)
summary(linear)
```

```
##
## Call:
## lm(formula = musclemass ~ age, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

```
fit<- linear$fitted.values
fit
```

```
##         1         2         3         4         5         6         7         8
## 105.17676 107.55675 100.41678 101.60677 102.79677 107.55675 100.41678 107.55675
##         9        10        11        12        13        14        15        16
##  99.22678  99.22678 106.36675 100.41678 105.17676 103.98676 106.36675  90.89681
##        17        18        19        20        21        22        23        24
##  88.51682  89.70682  86.13683  88.51682  92.08681  93.27680  94.46680  93.27680
##        25        26        27        28        29        30        31        32
##  92.08681  84.94683  86.13683  95.65679  86.13683  88.51682  75.42687  81.37685
##        33        34        35        36        37        38        39        40
##  84.94683  81.37685  81.37685  80.18685  77.80686  78.99686  84.94683  78.99686
##        41        42        43        44        45        46        47        48
##  78.99686  74.23687  83.75684  73.04688  75.42687  63.52691  63.52691  63.52691
```

```
##        49        50        51        52        53        54        55        56
##  70.66689  73.04688  69.47689  65.90691  63.52691  63.52691  71.85688  67.09690
##        57        58        59        60
##  64.71691  65.90691  70.66689  65.90691
```

```
age<-c(45,55,65)
value<-predict.lm(linear,data.frame(age=c(45,55,65)), se.fit=TRUE, level=0.95)
value
```

```
## $fit
##         1         2         3
## 102.79677  90.89681  78.99686
##
## $se.fit
##        1        2        3
## 1.714578 1.146901 1.148083
##
## $df
## [1] 58
##
## $residual.scale
## [1] 8.173177
```

```
work<-rep(sqrt(2 * qf(0.95,2,58)), length(age))
work
```

```
## [1] 2.512342 2.512342 2.512342
```

```
value
```

```
## $fit
##         1         2         3
## 102.79677  90.89681  78.99686
##
## $se.fit
##        1        2        3
## 1.714578 1.146901 1.148083
##
## $df
## [1] 58
##
## $residual.scale
## [1] 8.173177
```

```
rbind(value$fit - work * value$se.fit, value$fit + work * value$se.fit)
```

```
##               1        2        3
## [1,]  98.48916 88.01540 76.11248
## [2,] 107.10437 93.77822 81.88123
```

$B = t(.99375; 6) = 2.558541$ $W = 3.4168$ $F(.95; 5, 65) = 2.3349$

$15.79813 \pm 2.558541(0.2780832)$ $15.08664 \le E\{Y_h\} \le 16.50962$ $16.02754 \pm 2.558541(0.2359255)$ $15.42391 \le E\{Y_h\} \le 16.63116$ $15.90072 \pm 2.558541(0.2221593)$ $15.33232 \le E\{Y_h\} \le 16.46913$ $15.84339 \pm 2.558541(0.2591281)$ $15.18040 \le E\{Y_h\} \le 16.50638$

b).Is the Working-Hotelling procedure the most efficient one to be employed in part (a)? Explain

```
alpha = 0.05
B <- rep(qt(1 - alpha/(2 *length(age)),58), length(age))
B
```

```
## [1] 2.465398 2.465398 2.465398
```

As B = 2.465398, Working-Hotelling procedure is not the most effecient to be employed in part(a)

---

c). Three additional women aged 48, 59, and 74 have contacted the nutritionist. Predict the muscle mass for each of these three women using the Bonferroni procedure and a 95 percent family confidence coefficient

```
age_new<-c(48,59,74)
new_val<-predict.lm(linear,data.frame(age=c(48,59,74)), se.fit=TRUE, level=0.95)
new_val
```

```
## $fit
##        1        2        3
## 99.22678 86.13683 68.28690
##
## $se.fit
##        1        2        3
## 1.510501 1.058874 1.646728
##
## $df
## [1] 58
##
## $residual.scale
## [1] 8.173177
```

```
alpha = 0.05
B = rep(qt(1 - alpha/(2 *length(age_new)),58), length(age_new))
B
```

```
## [1] 2.465398 2.465398 2.465398
```

```
rbind(new_val$fit - B * new_val$se.fit, new_val$fit + B * new_val$se.fit)
```

```
##              1        2        3
## [1,]  95.50279 83.52628 64.22706
## [2,] 102.95077 88.74737 72.34674
```

d)Subsequently, the nutritionist wishes to predict the muscle mass for a fourth woman aged 64, with a family confidence coefficient of 95 percent for the four predictions. Will the three prediction intervals in part (c) have to be recalculated? Would this also be true if the Scheffe procedure had been used in constructing the prediction intervals?

Yes, the three prediction in part(c) have to be recalculated and also the Scheffe procedure has been used in constructing the prediction interval

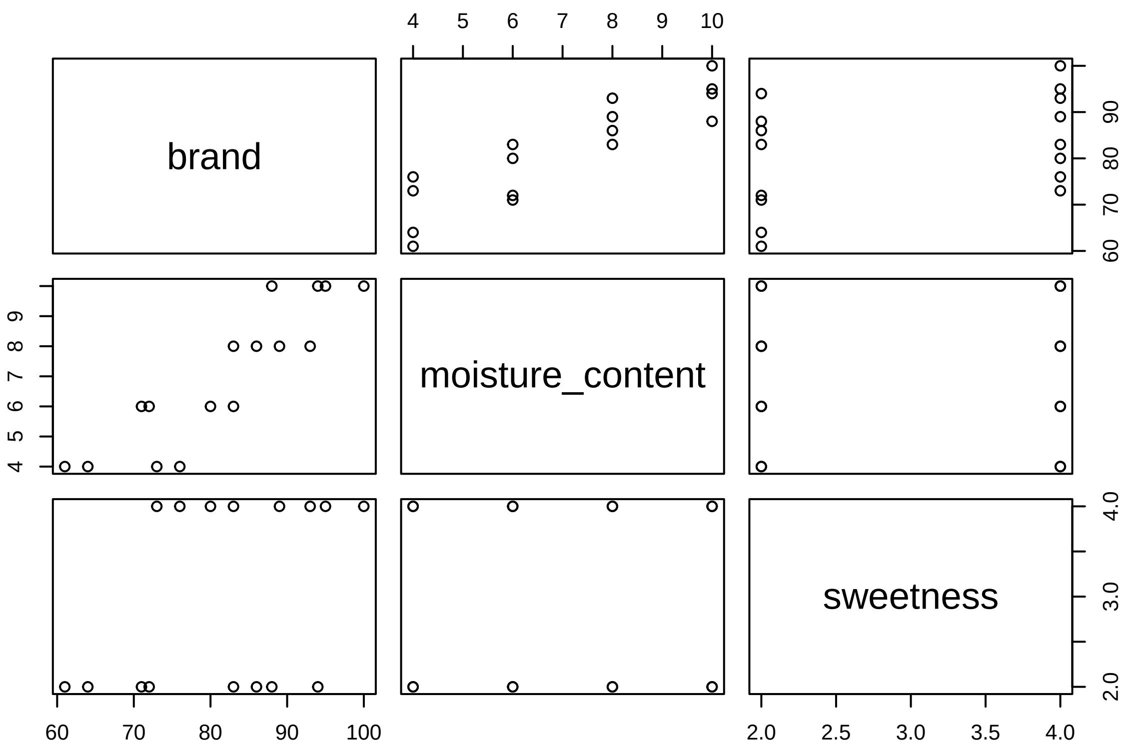
Problem 3

```
data3<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%206%20Data%20Sets/CH06PR0
colnames(data3)[1] ="brand"
colnames(data3)[2]="moisture_content"
colnames(data3)[3] ="sweetness"
head(data3)
```

```
##   brand moisture_content sweetness
## 1    64                4         2
## 2    73                4         4
## 3    61                4         2
## 4    76                4         4
## 5    72                6         2
## 6    80                6         4
```

a)Scatter Plot and Correlation Matrix

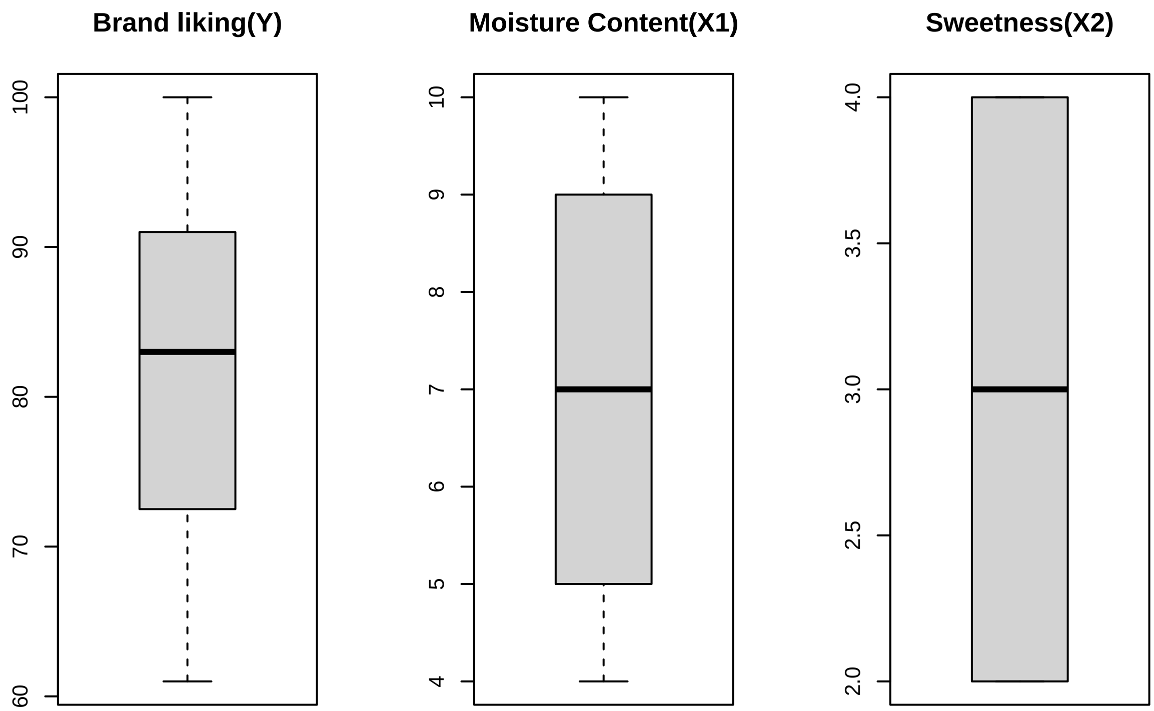
```
plot(data3)
```




```
cor(data3)
```

```
##                      brand moisture_content sweetness
## brand            1.0000000        0.8923929 0.3945807
## moisture_content 0.8923929        1.0000000 0.0000000
## sweetness        0.3945807        0.0000000 1.0000000
```

```
par(mfrow=c(1,3))
boxplot(data3$brand,main="Brand liking(Y)");boxplot(data3$moisture_content,main="Moisture Content(X1)")
```



The diagnostic aids show that firstly, there are no outliers and the distribution for each variable is normal. Additionally, looking at the correlation matrix, Y and X1 have significant positive correlation, Y and X2 are positively correlated, but less so than Y and X1 and there's no corrleation between X1 and X2.

---

b) Regression Model

```
lm3<-lm(brand~moisture_content+sweetness,data=data3)
summary(lm3)
```

```
##
## Call:
## lm(formula = brand ~ moisture_content + sweetness, data = data3)
##
## 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 ***
## moisture_content   4.4250     0.3011  14.695 1.78e-09 ***
## sweetness          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
```

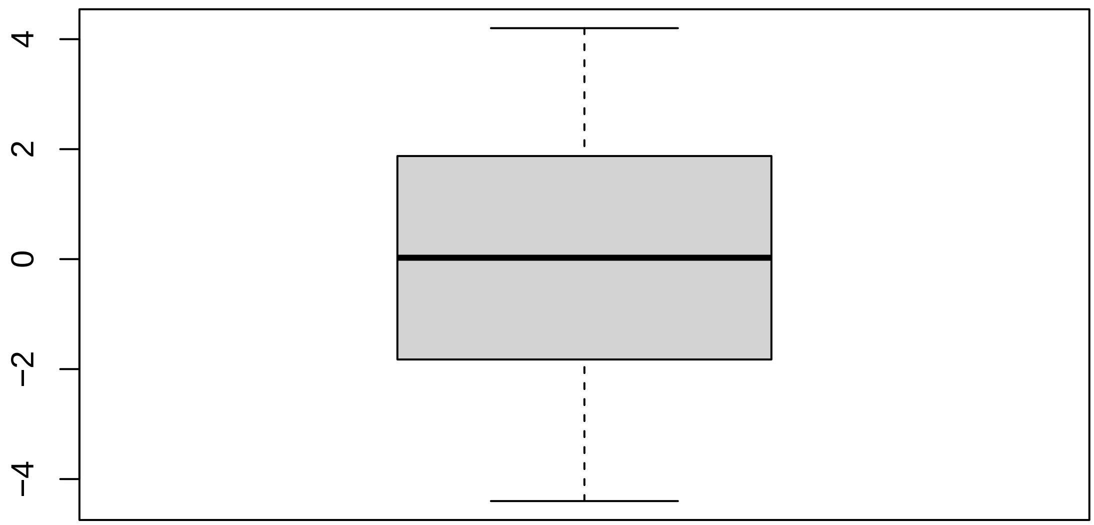
```
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

The regression model yields the equation Y= 37.65 + 4.425X1 + 4.375X2. Holding the other variable constant, Increasing one unit of X1 leads to an increase in the brand liking degree by 4.425, and holding X1 constant, an one unit increase in X2 leads to an increase of the brand liking degreee of 4.375. Both X1 and X2 are significant as the P values for each variable are $<$ 0.05.

---

c) Residuals and box plot

```
resid<-lm3$residuals
boxplot(resid)
```

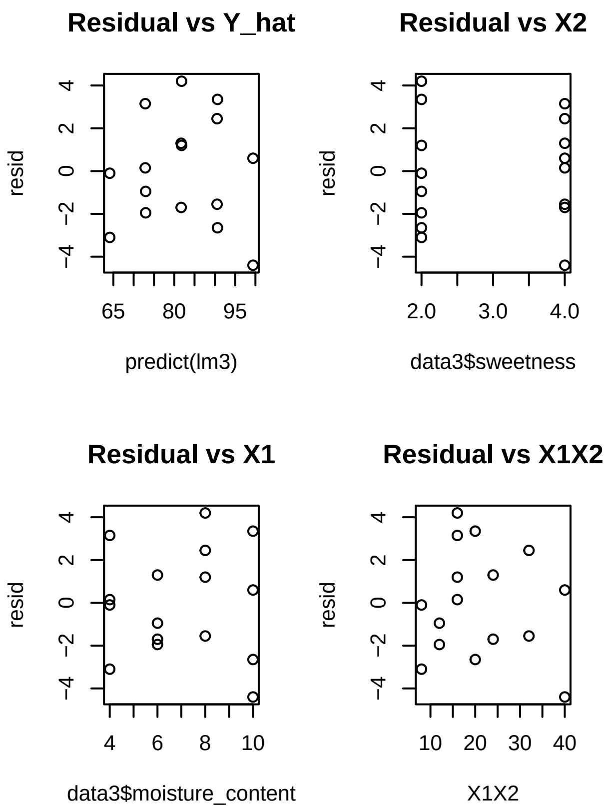


There are no outliers and errors are normally distribute

---

d) Plot residuals against Y,X1,X2

```
par(mfcol=c(2,4))
X1X2 = data3$moisture_content * data3$sweetness
plot(resid ~ predict(lm3) , main="Residual vs Y_hat")
plot(resid ~ data3$moisture_content, main="Residual vs X1")
plot(resid ~ data3$sweetness, main="Residual vs X2")
plot(resid ~ X1X2, main="Residual vs X1X2")
```

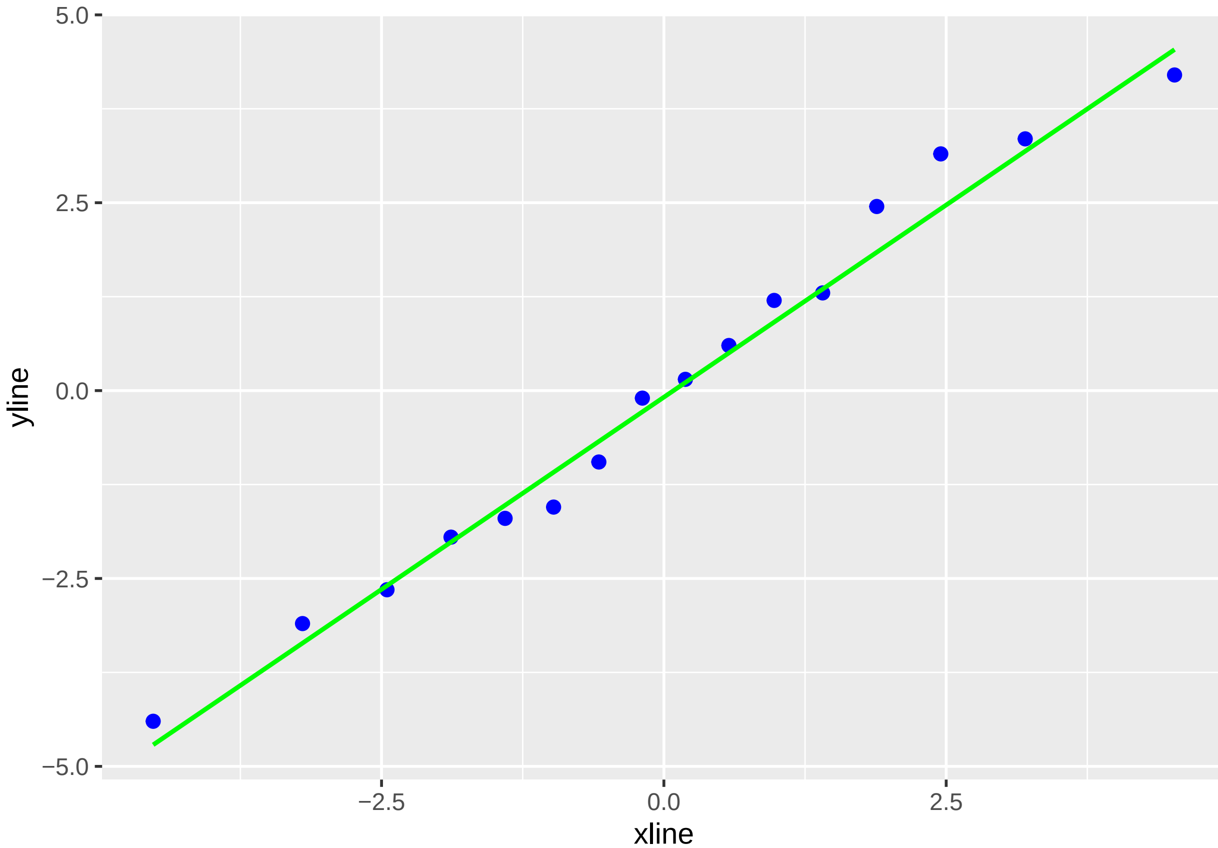



```r
library(ggplot2)
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```r
ggplot(mapping= aes(sample = resid)) + stat_qq_point(size=2, color="blue") + stat_qq_line(color='green'
```



e) The alternatives are as follows:

$$H_0 : \gamma_1 = \gamma_2 = 0$$
$$H_a : \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0$$

The decision rule is reject H_0 if:

$$\chi^2_{BP} > \chi^2_{BP}(0.99, 2)$$

```
bp = lm(resid^2 ~ data3$moisture_content + data3$sweetness)
summary(bp)
```

```
##
## Call:
## lm(formula = resid^2 ~ data3$moisture_content + data3$sweetness)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.724 -3.732 -1.961  2.987 11.276
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.1588     6.8599   0.169    0.868
## data3$moisture_content   0.9175     0.6894   1.331    0.206
## data3$sweetness         -0.5625     1.5416  -0.365    0.721
##
```

```
## Residual standard error: 6.167 on 13 degrees of freedom
## Multiple R-squared:  0.1278, Adjusted R-squared:  -0.006434
## F-statistic: 0.9521 on 2 and 13 DF,  p-value: 0.4113
```

```r
anova(bp)
```

```
## Analysis of Variance Table
##
## Response: resid^2
##                        Df Sum Sq Mean Sq F value Pr(>F)
## data3$moisture_content  1  67.34  67.344  1.7710 0.2061
## data3$sweetness         1   5.06   5.063  0.1331 0.7211
## Residuals              13 494.35  38.027
```

```r
bp_SSR = sum(anova(bp)$"Sum Sq ") - deviance(bp)
bp_SSR
```

```
## [1] -494.3472
```

```r
bp_SSE = deviance(lm3)
bp_SSE
```

```
## [1] 94.3
```

```r
bp_chisqr  =  (bp_SSR/2)/(bp_SSE/length(lm3$model$brand))^2
bp_chisqr
```

```
## [1] -7.115717
```

```r
bp_critical =  qchisq(0.99,2)
bp_critical
```

```
## [1] 9.21034
```

$\chi^2_{BP} is -7.115717, and the critical value is", 9.21034l$ **We cannot reject the null hypothesis, the model has constant error variance.**

f)

The Alternative are as follows :

$H_0 : E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $H_1 : E(Y) \neq \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

```r
moisture<-factor(data3$moisture_content)
sweetness_ <-factor(data3$sweetness)
reduced_model = lm(brand ~ moisture *sweetness_ ,data3)
anova(reduced_model)
```

```
## Analysis of Variance Table
##
## Response: brand
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## moisture             3 1581.50  527.17 73.9883 3.554e-06 ***
## sweetness_           1  306.25  306.25 42.9825 0.0001773 ***
## moisture:sweetness_  3   22.25    7.42  1.0409 0.4253674
## Residuals            8   57.00    7.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm3, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: brand ~ moisture_content + sweetness
## Model 2: brand ~ moisture * sweetness_
##   Res.Df  RSS Df Sum of Sq     F Pr(>F)
## 1     13 94.3
## 2      8 57.0  5      37.3 1.047  0.453
```

Reject null Hypothesis H_0 if the test statistics is larger than F, that says that the regression function is not linear

As the MSLF = 7.46 and MSPE = 7.125, F^* Statistics is $(MSLF/MSPE)7.46/7.125 = 1.047$ which is far less than $F(0.99; 5, 8)$ *critical value* $= 6.6318$ As $F^* <= 6.63$, Conclude H_0 so that we can conclude that there is no lack of fit ---

—————————————Problem 4—————————————————————————

```
data4 = read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%206%20Data%20Sets/CH06PR1
colnames(data4)[1] ="X"
colnames(data4)[2]="B"
colnames(data4)[3] ="C"
colnames(data4)[4]="D"
colnames(data4)[5]="F"
head(data4)
```

```
##      X  B     C    D      F
## 1 13.5  1  5.02 0.14 123000
## 2 12.0 14  8.19 0.27 104079
## 3 10.5 16  3.00 0.00  39998
## 4 15.0  4 10.70 0.05  57112
## 5 14.0 11  8.97 0.07  60000
## 6 10.5 15  9.45 0.24 101385
```

```
#Taking columns names as X,B,C,D,F as reference
# X ---> RENTAL RATES
# B ---> AGE
# C ---> OPEARTING EXPENSES AND TAXES
# D ---> VACANCY RATES
# F ---> TOTAL SQUARE FOOTAGE
```

a) Stem and leaf plot for each predictor variable

```
stem(data4$B)
```

```
##
##   The decimal point is at the |
##
##    0 | 0000000000000000
##    2 | 000000000000000000000000
##    4 | 00000
##    6 | 0
##    8 | 0
##   10 | 00
##   12 | 00000
##   14 | 0000000000000
##   16 | 0000000000
##   18 | 000
##   20 | 00
```

```
stem(data4$C)
```

```
##
##   The decimal point is at the |
##
##    2 | 0
##    4 | 080003358
##    6 | 012613
##    8 | 00001223456001555689
##   10 | 01334456667777812334466668
##   12 | 00011115777889002
```

```
##   14 | 6
```

```
stem(data4$D)
```

```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   0 | 0000000000000000000000000000002333333333334444445555556678889
##   1 | 023444469
##   2 | 1223477
##   3 | 3
##   4 |
##   5 | 7
##   6 | 0
##   7 | 3
```

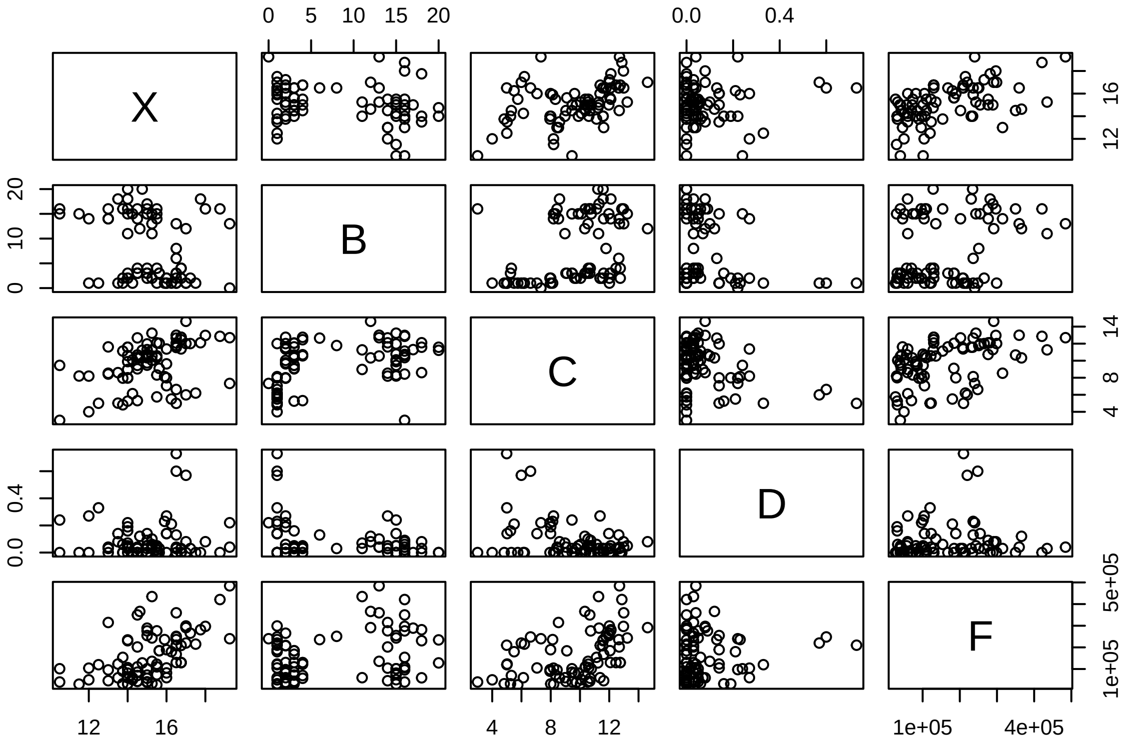
```
stem(data4$F)
```

```
##
##   The decimal point is 5 digit(s) to the right of the |
##
##   0 | 333333444444
##   0 | 555666667778899
##   1 | 000001111222333334
##   1 | 578889
##   2 | 011122334444
##   2 | 555788899
##   3 | 002
##   3 | 567
##   4 | 23
##   4 | 8
```

---

b) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings

```
par(mfrow=c(3,2))
pairs(data4)
```




```
cor(data4)
```

```
##             X          B          C           D          F
## X  1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## B -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## C  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## D  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## F  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

The scatter plot matrix reveals that there aren't many outliers, especially for the vacancy rates, and that the plot is discrete.Along with this, there is a strong positive relationship between rental rates and vacancy rates, as well as between rental rates and total square footage. Other than these, there is a bad correlation between age and rental rates. Both operating expenses and taxes have a favorable correlation with rental rates.

##c) Linear Regression Function

```
lm4 <-lm(X ~ B+C+D+F, data=data4)
summary(lm4)
```

```
##
## Call:
## lm(formula = X ~ B + C + D + F, data = data4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
```

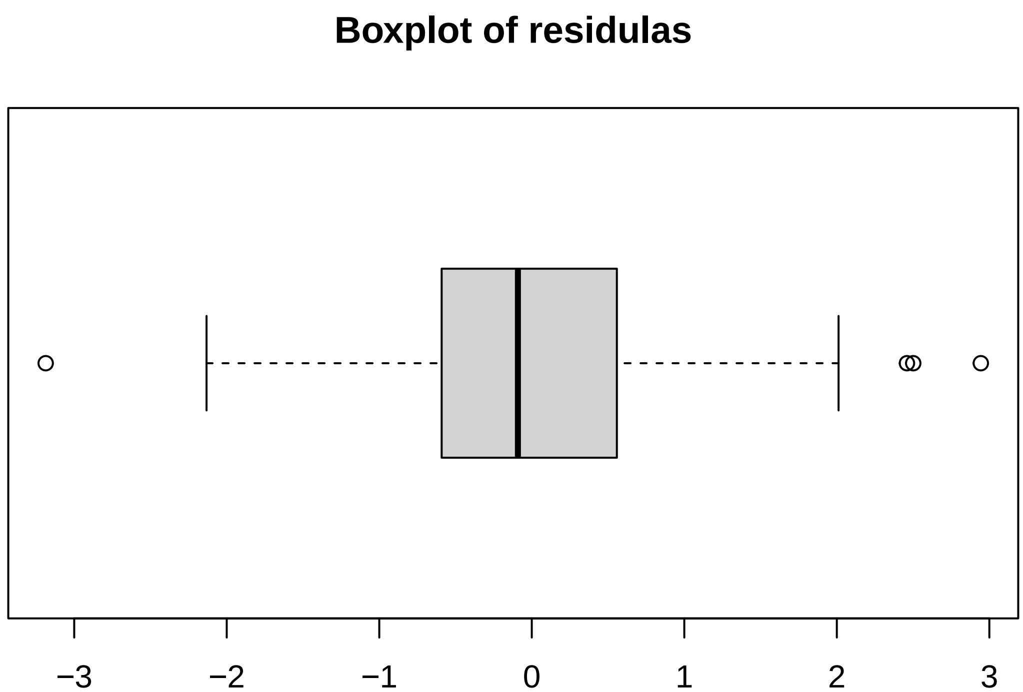
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## B           -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## C            2.820e-01  6.317e-02   4.464 2.75e-05 ***
## D            6.193e-01  1.087e+00   0.570     0.57    
## F            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
```

Regression function is $Y = 1.220 * 10^{01} - 1.420 * 10^{01} X_1 + 2.820 * 10^{-01} X_2 + 6.193 * 10^{-01} X_3 + 7.924 * 10^{-06} X_3$

---

d)Obtain the residuals and prepare a boxplot of the residuals.Does the distribution appear to be fairly symmetrical?

```
resid <-as.numeric(lm4$residuals)
boxplot(resid, main="Boxplot of residulas",horizontal = T)
```

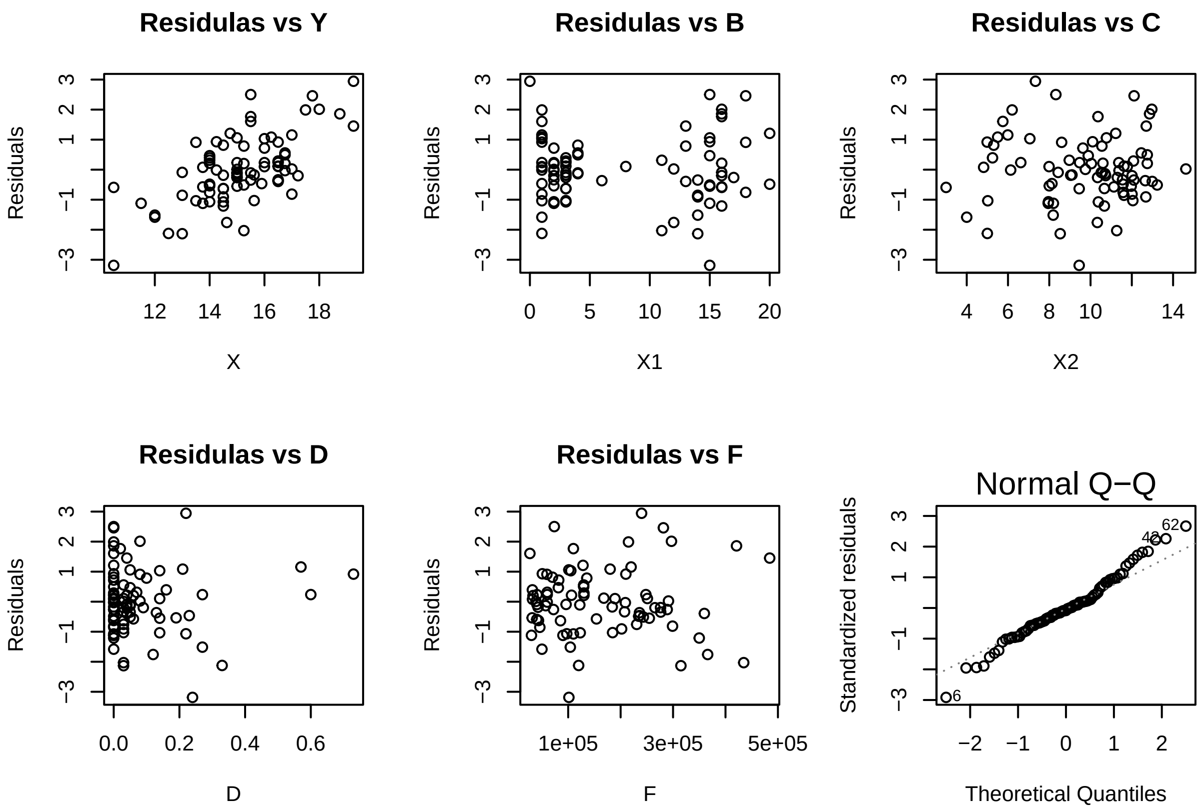


Although the disturbance and residual are typical, there are outliers on both. However, the right side has more outliers than the left side.

---

e)Plot the residuals against $\hat{Y}$, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze your plots and summarize your findings

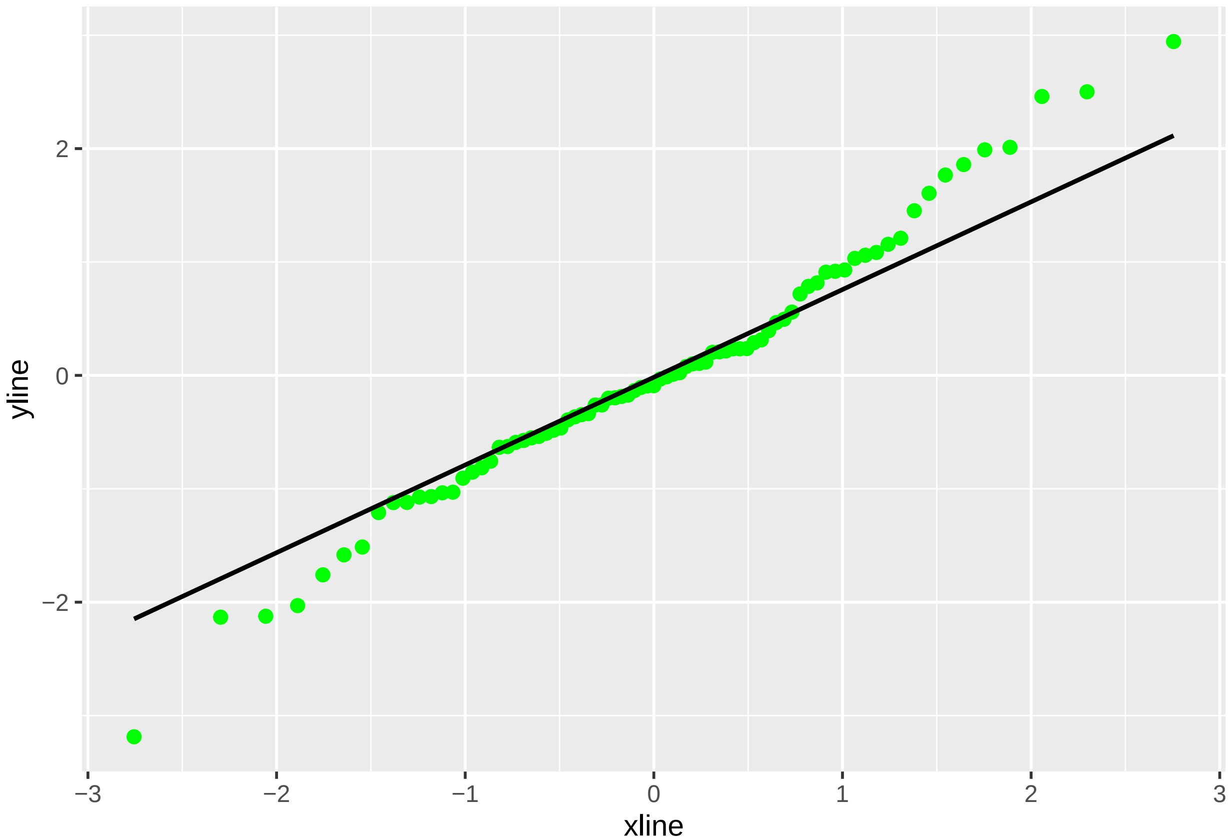
```
par(mfrow=c(2,3))
plot(x = data4$X, y = resid, xlab = "X", ylab = "Residuals", main="Residulas vs Y")
plot(x = data4$B, y = resid, xlab = "X1", ylab = "Residuals", main="Residulas vs B")
plot(x = data4$C, y = resid, xlab = "X2", ylab = "Residuals", main="Residulas vs C")
plot(x = data4$D, y = resid, xlab = "D", ylab = "Residuals", main="Residulas vs D")
plot(x = data4$F, y = resid, xlab = "F", ylab = "Residuals", main="Residulas vs F")
plot(lm4, which=2)
```



##

As we can see, both the Vacancy Rates Boxplot and the Residuals Boxplot contain outliers. Operating Expenses and taxes are left-skewed, while total square footage and vacancy rates are right-skewed. Additionally, the normal probability plot demonstrates the normal distribution of the residuals.

```
library(ggplot2)
library(qqplotr)
ggplot(mapping= aes(sample = resid)) + stat_qq_point(size=2, color="green") + stat_qq_line(color='black
```



f). Can you conduct a formal test for lack of fit here?

```
age2 <- factor(data4$B)
C <- factor(data4$C)
D <- factor(data4$D)
F <- factor(data4$F)
reduced_model1 <- lm(X ~ age2 *C*D*F, data=data4)
anova(reduced_model1)
```

```
## Analysis of Variance Table
##
## Response: X
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## age2       15 63.644   4.243   9.0013 5.767e-05 ***
## C           1 84.825  84.825 179.9541 9.292e-10 ***
## D           1  2.114   2.114   4.4854 0.0512986 .
## F           1 15.523  15.523  32.9327 3.918e-05 ***
## age2:C     11  9.623   0.875   1.8560 0.1317408
## age2:D      8 19.497   2.437   5.1702 0.0031010 **
## C:D         1  0.314   0.314   0.6659 0.4272387
## age2:F      6  1.385   0.231   0.4897 0.8061558
## C:F         1  0.082   0.082   0.1730 0.6833613
## D:F         1  2.839   2.839   6.0235 0.0268147 *
## age2:C:D    6  3.658   0.610   1.2935 0.3183362
## age2:C:F    4 20.780   5.195  11.0208 0.0002259 ***
## age2:D:F    4  3.640   0.910   1.9305 0.1575451
## C:D:F       1  0.642   0.642   1.3616 0.2614762
```

```
## age2:C:D:F  4  0.920   0.230   0.4881 0.7444585
## Residuals  15  7.071   0.471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above anova table we can say that No, we can not do the lack of fit test.

---

g). Divide the 81 cases into two groups. placing the 40 cases with the smallest fitted values $\hat{Y}_i$ into group I and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .05$. State the decision rule and conclusion

```
commercial_data <- data4[, -6]
commercial_data$fitted_values = as.numeric(lm4$fitted.values)
head(commercial_data)
```

```
##      X  B     C    D      F fitted_values
## 1 13.5  1  5.02 0.14 123000      14.53567
## 2 12.0 14  8.19 0.27 104079      13.51381
## 3 10.5 16  3.00 0.00  39998      11.09105
## 4 15.0  4 10.70 0.05  57112      15.13357
## 5 14.0 11  8.97 0.07  60000      13.68672
## 6 10.5 15  9.45 0.24 101385      13.68719
```

```
sorted_data = commercial_data[order(commercial_data$fitted_values),]
```

```
sorted_data$group <- "Group2"
sorted_data$group[1:40] <- "Group1"
head(sorted_data)
```

```
##        X  B     C    D     F fitted_values  group
## 3  10.50 16  3.00 0.00 39998      11.09105 Group1
## 78 13.50 18  8.60 0.08 59443      12.58991 Group1
## 40 11.50 15  8.20 0.00 30005      12.62039 Group1
## 42 15.50 15  8.32 0.00 73521      12.99906 Group1
## 34 13.00 16  8.43 0.04 96000      13.09095 Group1
## 44 14.25 15 10.10 0.00 50724      13.32040 Group1
```

```
library(onewaytests)
bf.test(fitted_values ~ group , data = sorted_data)
```

```
##
##   Brown-Forsythe Test (alpha = 0.05)
## -------------------------------------------------------------
##   data : fitted_values and group
##
##   statistic  : 135.8539
##   num df     : 1
##   denom df   : 74.89145
##   p.value    : 1.699122e-18
##
##   Result     : Difference is statistically significant.
```

```
## ------------------------------------------------------------
```

——————————Problem 5——————————————————

a). Test whether there is a regression relation; use $\alpha = .05$. State the alternatives decision rule, and conclusion. What does your test imply about $\beta_1, \beta_2, \beta_3, and \beta_4$? What is the P-value of the test?

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = X ~ B + C + D + F, data = data4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## B           -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## C            2.820e-01  6.317e-02   4.464 2.75e-05 ***
## D            6.193e-01  1.087e+00   0.570     0.57
## F            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
```

Null Hypothesis $H_0 : \beta_1 = beta_2 = \beta_3 = \beta_4 = 0$ Alternate Hypothesis $H_a : not\ all\ \beta_i = 0 (i = 1, 2, 3, 4)$ $F* = 26.756$ $F(.95; 4, 76) = 2.4920$ Decision : As F* <= 2.4920 conclude H_0 otherwise H_a Conclusion : Conclude H_a P -value for the test is + 7.272*10-14 positive

```r
reduced_rental <- lm(X ~ 1,data=data4)
anova(reduced_rental, lm4, test='F')
```

```
## Analysis of Variance Table
##
## Model 1: X ~ 1
## Model 2: X ~ B + C + D + F
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     80 236.558
## 2     76  98.231  4    138.33 26.756 7.272e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm4)$fstatistic
```

```
##    value    numdf    dendf
## 26.75553  4.00000 76.00000
```

```r
qf(p=.05, df1=4, df2=76, lower.tail=FALSE)
```

```
## [1] 2.492049
```

b). Estimate $\beta_1, \beta_2, \beta_3, and \beta_4$ jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results

```
SE <- sqrt(diag(vcov(lm4)))
SE
```

```
##  (Intercept)            B            C            D            F
## 5.779562e-01 2.134261e-02 6.317235e-02 1.086813e+00 1.384775e-06
```

c). Calculate R^2 and interpret this measure.

```
anova(reduced_rental,lm4, test='F')
```

```
## Analysis of Variance Table
##
## Model 1: X ~ 1
## Model 2: X ~ B + C + D + F
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     80 236.558
## 2     76  98.231  4    138.33 26.756 7.272e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm4)$r.squared
```

```
## [1] 0.5847496
```

$SSR = 138.33$ $SSTO = 236.558$ $R^2 = 0.5847$

##-----------------Problem 6-------------------------------------

a).Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the most efficient procedure.

```
data5<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%206%20Data%20Sets/CH06PR2

colnames(data5)[1] ="B"
colnames(data5)[2]="C"
colnames(data5)[3] ="D"
colnames(data5)[4]="F"

head(data5)
```

```
##    B     C    D      F
## 1  5  8.25 0.00 250000
## 2  6  8.50 0.23 270000
## 3 14 11.50 0.11 300000
## 4 12 10.25 0.00 310000
```

```
rental_pred_val = predict.lm(lm4,data5, se.fit=TRUE, level=0.95)
rental_pred_val
```

```
## $fit
##        1        2        3        4
## 15.79813 16.02754 15.90072 15.84339
##
## $se.fit
##         1         2         3         4
## 0.2780832 0.2359255 0.2221593 0.2591281
##
## $df
## [1] 76
##
## $residual.scale
## [1] 1.136885
```

```
alpha = 0.05
rental = rep(qt(1 - alpha/(2 *length(data5)),76), length(data5))
rental
```

```
## [1] 2.558541 2.558541 2.558541 2.558541
```

```
rbind(rental_pred_val$fit - rental * rental_pred_val$se.fit, rental_pred_val$fit + rental * rental_pred
```

```
##             1        2        3        4
## [1,] 15.08664 15.42391 15.33232 15.18040
## [2,] 16.50962 16.63116 16.46913 16.50638
```

```
F_rental = qf(.95,5,76)
F_rental
```

```
## [1] 2.33492
```

```r
W_rental = sqrt(2 * qf(.95,5,76))
W_rental
```

```
## [1] 2.160981
```

```r
rbind(rental_pred_val$fit - W_rental * rental_pred_val$se.fit, rental_pred_val$fit + W_rental * rental_p
```

```
##             1        2        3        4
## [1,] 15.19720 15.51770 15.42064 15.28341
## [2,] 16.39906 16.53737 16.38081 16.40336
```

$B = t(0.99375; 6) = 2.558541$ $W = 3.4168$ $F(.95; 5, 65) = 2.3349$

$15.79813 \pm 2.558541(0.2780832)$ $15.08664 \leq E\{Y_h\} \leq 16.50962$ $16.02754 \pm 2.558541(0.2359255)$ $15.42391 \leq E\{Y_h\} \leq 16.63116$ $15.90072 \pm 2.558541(0.2221593)$ $15.33232 \leq E\{Y_h\} \leq 16.46913$ $15.84339 \pm 2.558541(0.2591281)$ $15.18040 \leq E\{Y_h\} \leq 16.50638$

Problem 7

a).Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model

```
standarized_df = data.frame(scale(data4))
head(standarized_df)
```

```
##             X          B           C           D          F
## 1 -0.95307295 -1.0348893 -1.80714029  0.43858670 -0.3449462
## 2 -1.82537701  0.9250719 -0.57996529  1.40476201 -0.5183759
## 3 -2.69768106  1.2266044 -2.58912563 -0.60190978 -1.1057417
## 4 -0.08076889 -0.5825906  0.39170956 -0.23030390 -0.9488750
## 5 -0.66230493  0.4727732 -0.27801055 -0.08166154 -0.9224036
## 6 -2.69768106  1.0758382 -0.09219226  1.18179848 -0.5430691
```

```
stmodel = lm(X ~ B + C+ D+ F, data=standarized_df)
summary(stmodel)
```

```
##
## Call:
## lm(formula = X ~ B + C + D + F, data = standarized_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85346 -0.34372 -0.05289  0.32446  1.71213
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.365e-16  7.346e-02   0.000     1.00
## B           -5.479e-01  8.232e-02  -6.655 3.89e-09 ***
## C            4.236e-01  9.490e-02   4.464 2.75e-05 ***
## D            4.846e-02  8.504e-02   0.570     0.57
## F            5.028e-01  8.786e-02   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6611 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
```

b).Interpret the standardized regression coefficient β_2

```
cor(data4$X, data4$C)
```

```
## [1] 0.4137872
```

```
vif(lm4)
```

```
##        B        C        D        F
## 1.240348 1.648225 1.323552 1.412722
```

c)Transform the estimated standardized regression coefficients by means of (7.S3) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.18c

```
sy <- sd(data4$X)
sy
```

```
## [1] 1.719584
```

```
beta_1 = (sy/sd(data4$B))*(stmodel$coefficients['B'])
beta_2 = (sy/sd(data4$C))*(stmodel$coefficients['C'])
beta_3 = (sy/sd(data4$D))*(stmodel$coefficients['D'])
beta_4 = (sy/sd(data4$F))*(stmodel$coefficients['F'])
beta_0 = (mean(data4$X) - beta_1*mean(data4$B) - beta_2*mean(data4$C) - beta_3*mean(data4$D) - beta_4*me

cat(" Age (b1) : ",beta_1)
```

```
##  Age (b1) :  -0.1420336
```

```
cat("\n Operating Expenses (b2) : ",beta_2)
```

```
##
##  Operating Expenses (b2) :  0.2820165
```

```
cat("\n Vacancy Rates (b3) : ",beta_3)
```

```
##
##  Vacancy Rates (b3) :  0.6193435
```

```
cat("\n Total square footage(b4) : ",beta_4)
```

```
##
##  Total square footage(b4) :  7.924302e-06
```

```
cat("\n Intercept : ",as.numeric(beta_0))
```

```
##
##  Intercept :  12.20059
```

$b0 = 12.20059$ $b1 = -0.1420336$ $b2 = 0.2820165$ $b3 = 0.6193435$ $b4 = 7.924302e - 06$

```
summary(lm4)$coefficients
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  1.220059e+01 5.779562e-01 21.1098807 1.601720e-33
## B           -1.420336e-01 2.134261e-02 -6.6549332 3.894322e-09
## C            2.820165e-01 6.317235e-02  4.4642400 2.747396e-05
## D            6.193435e-01 1.086813e+00  0.5698714 5.704457e-01
## F            7.924302e-06 1.384775e-06  5.7224457 1.975990e-07
```

Problem 8

a). Fit first-order simple linear regression model (2.1) for relating brand liking (Y) to moisture content (X,). State the fitted regression function.

```
brandpref_model = lm(brand ~ moisture_content, data=data3)
summary(brandpref_model)
```

```
##
## Call:
## lm(formula = brand ~ moisture_content, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.475 -4.688 -0.100  4.638  7.525
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        50.775      4.395  11.554 1.52e-08 ***
## moisture_content    4.425      0.598   7.399 3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.349 on 14 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7818
## F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

The estimated function is $\hat{Y}_i = 50.775 + 4.425X_{1i}$

---

b). Compare the models

```
summary(lm3)
```

```
##
## Call:
## lm(formula = brand ~ moisture_content + sweetness, data = data3)
##
## 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 ***
## moisture_content   4.4250     0.3011  14.695 1.78e-09 ***
## sweetness          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
```

Both the coefficients in brand_model and brandpref_model are same.

c)

```r
swtnesbrandpref_model <- lm(brand ~ sweetness,data=data3)
```

```r
SSR_X1X2 <- deviance(swtnesbrandpref_model) - deviance(lm3)
SSR_X1X2
```

```
## [1] 1566.45
```

```r
SSR_X1 <- sum(anova(brandpref_model)$'Sum Sq') - deviance(brandpref_model)
SSR_X1
```

```
## [1] 1566.45
```

$SSR(X_1|X_2)\ and\ SSR(X_1)$ both are same

---

d)

```r
cor(data3)
```

```
##                      brand moisture_content sweetness
## brand            1.0000000        0.8923929 0.3945807
## moisture_content 0.8923929        1.0000000 0.0000000
## sweetness        0.3945807        0.0000000 1.0000000
```

Since the sweetness and moisture content are not correlated, as shown in the correlation matrix, the calculated coefficient in section (b) is the same for both models. $X1$ and $X2$ are uncorrelated, $SSR(X1|X2) : and : SSR(X1)$ are the same. As a result, the existence of $X2$ does not provide any information for $X1$ because it is unrelated.

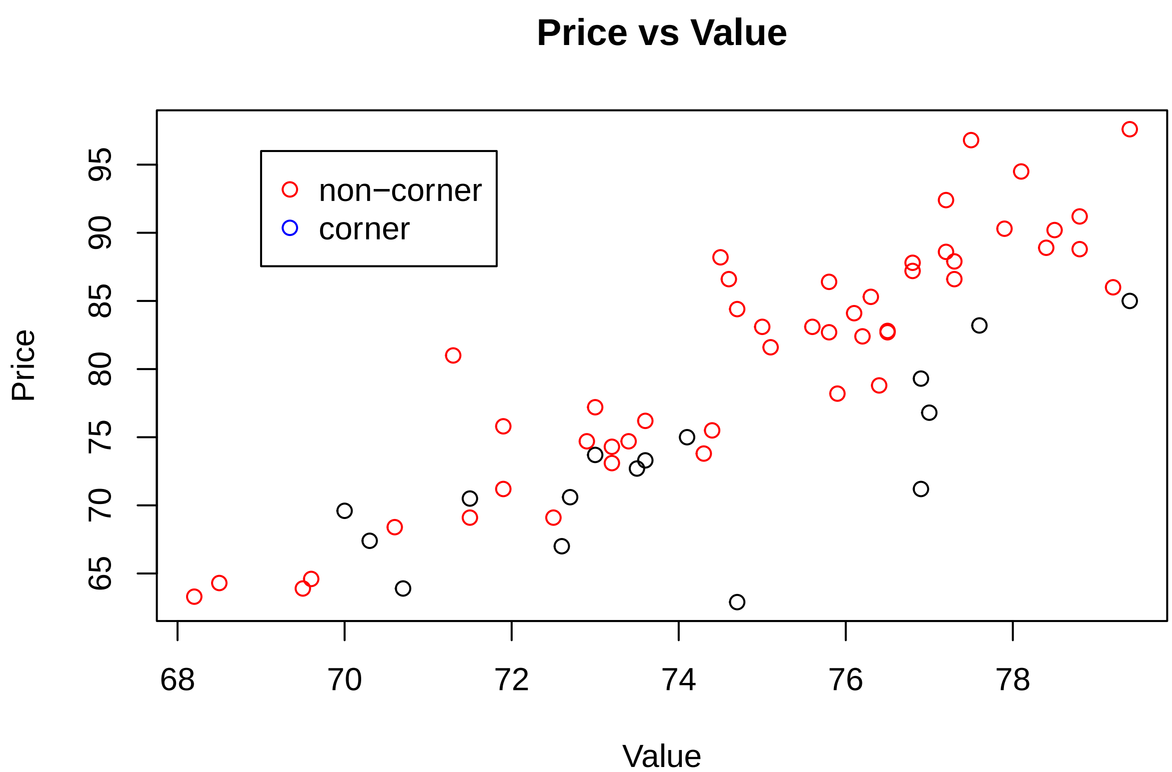
—————————Problem 9———————————————————

a). Plot the sample data for the two populations as a symbolic scatterplot.Does the regression relation appear to be the same for the two populations

```
assessed_val<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%208%20Data%20Sets/C
colnames(assessed_val)[1]="selling"
colnames(assessed_val)[2]="value"
colnames(assessed_val)[3]="location"
head(assessed_val)
```

```
##   selling value location
## 1    78.8  76.4        0
## 2    73.8  74.3        0
## 3    64.6  69.6        0
## 4    76.2  73.6        0
## 5    87.2  76.8        0
## 6    70.6  72.7        1
```

```
plot(assessed_val$value,assessed_val$selling, col=ifelse(assessed_val$location == 1, "black", "red"), xl
legend(69,96, col=c("red", "blue" ), pch=c(1,1),c("non-corner", "corner"))
```



The relationship between price and value is the same for both populations because both have a growing tendency in their numbers.

———————————————————

b). Test for identity of the regression functions for dwellings on comer lots and dwellings in other locations; control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
cornerplot <- assessed_val %>% filter(location==1)
head(cornerplot)
```

```
##   selling value location
## 1    70.6  72.7        1
## 2    71.2  76.9        1
## 3    76.8  77.0        1
## 4    73.7  73.0        1
## 5    85.0  79.4        1
## 6    69.6  70.0        1
```

```
noncornerplot <- assessed_val %>% filter(location==0)
head(noncornerplot)
```

```
##   selling value location
## 1    78.8  76.4        0
## 2    73.8  74.3        0
## 3    64.6  69.6        0
## 4    76.2  73.6        0
## 5    87.2  76.8        0
## 6    86.0  79.2        0
```

```
t_test <- t.test(cornerplot$value, noncornerplot$value, var.equal = T)
t_test
```

```
##
##  Two Sample t-test
##
## data:  cornerplot$value and noncornerplot$value
## t = -1.1083, df = 62, p-value = 0.272
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.5816824  0.7400158
```

```
## sample estimates:
## mean of x mean of y
##  74.03125  74.95208
```

Null Hypothesis $H_0 : \beta_1 = \beta_2$ Alternate Hypothesis $H_1 : \beta_1 \neq \beta_2$

Decision : Reject the null hypothesis, if the p value < 0.05 Level of Significance is give 0.05 Using T-test

From the following output for the t-test we can see that p-value is $0.272 > 0.05$, Conclusion : Hence the null Hypothesis is not rejected , and we can conclude that regression function for dwelling on corner lots and non corner lots are same, that is $\beta_1 \neq \beta_2$

---

c) Plot the estimated regression functions for the two populations and describe the nature of the differences between them

```
cornorlot_model <- lm(selling ~ value, data=cornerplot)
summary(cornorlot_model)
```

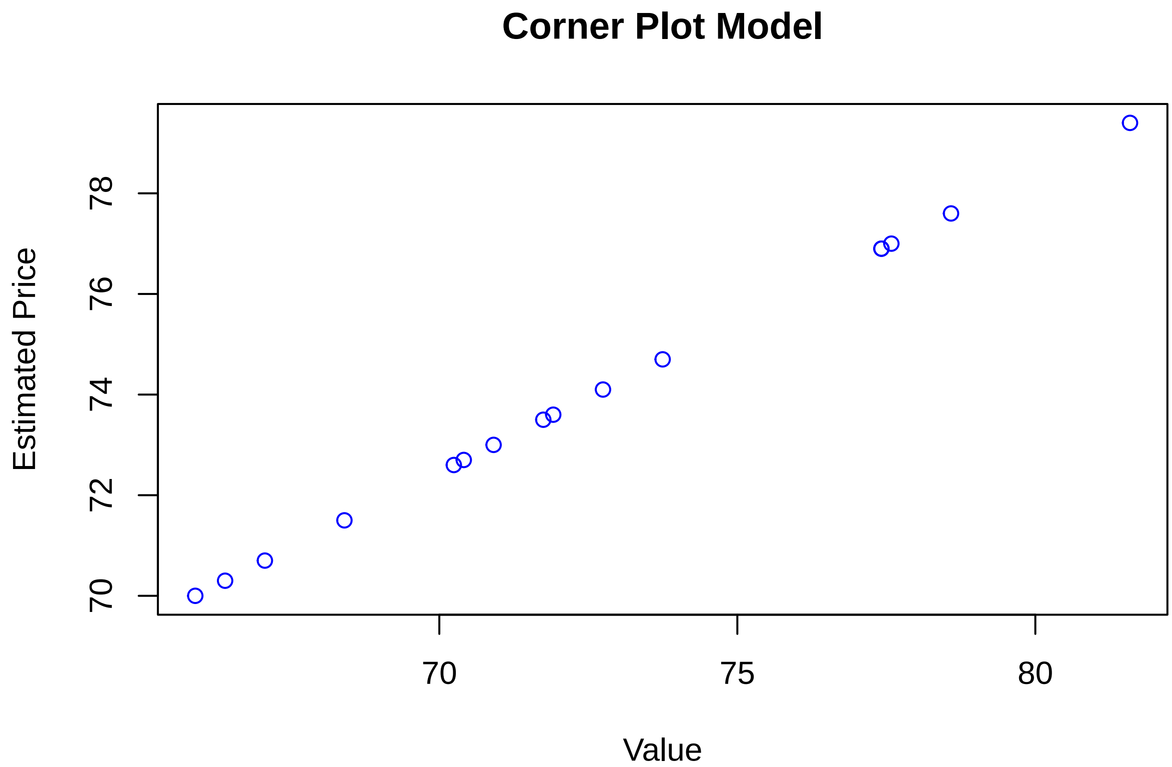
```
##
## Call:
## lm(formula = selling ~ value, data = cornerplot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.847  -1.382   1.191   2.388   4.615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.8836    28.4687  -1.787 0.095541 .
## value         1.6684     0.3843   4.342 0.000677 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.215 on 14 degrees of freedom
## Multiple R-squared:  0.5738, Adjusted R-squared:  0.5434
## F-statistic: 18.85 on 1 and 14 DF,  p-value: 0.0006769
```

```
noncornerlot_model <- lm(selling ~ value, data=noncornerplot)
summary(noncornerlot_model)
```

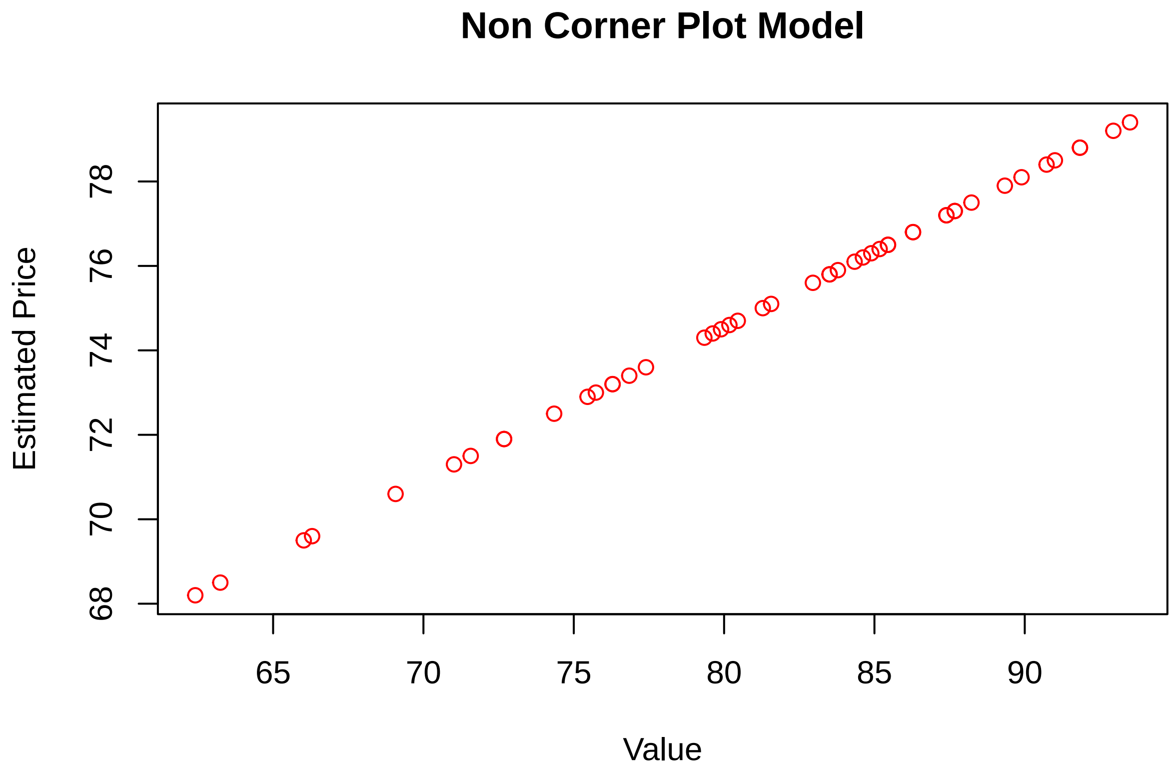
```
##
## Call:
## lm(formula = selling ~ value, data = noncornerplot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9460 -2.1639 -0.6544  1.4775  9.9836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.9052    14.3305  -8.856 1.68e-11 ***
## value          2.7759     0.1911  14.529  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.789 on 46 degrees of freedom
## Multiple R-squared:  0.8211, Adjusted R-squared:  0.8172
## F-statistic: 211.1 on 1 and 46 DF,  p-value: < 2.2e-16
```

```
plot(predict(cornorlot_model), cornerplot$value, xlab="Value", ylab="Estimated Price", main="Corner Plo
```




```
plot(predict(noncornerlot_model), noncornerplot$value, xlab="Value", ylab="Estimated Price", main="Non C
```


Non Corner Plot Model
Estimated Price
68
70
72
74
76
78
65
70
75
80
85
90
Value

10) Consider the equation

$$Y = \tilde{X}\beta + u$$

$$\hat{\beta} = (\tilde{X}'X)^{-1}\tilde{X}'Y$$

$$\hat{\beta} = (\tilde{X}'X)^{-1}\tilde{X}'y + Dy$$

1) $\hat{\beta}$ is unbiased

2) Variance of $\hat{\beta}$

Taking $\mathbb{E}[\bar{\beta}] = (\tilde{X}'X)^{-1}\tilde{X}'X\beta + (\tilde{X}'X)^{-1}X^{-1}u + DX\beta + Du$

By Substituting we get

$$\mathbb{E}[\bar{\beta}] = \beta + DX\beta \quad \text{—— (1)}$$

$\hat{\beta} = (\tilde{X}'X)^{-1}X'y + DY$ can be written as

$= Cy$

$DX = 0$

$Var(\hat{\beta}) = C\, Var(Y) C'$

WKT $Var(Y) = \sigma^2 I$

$= C \sigma^2 I C'$

$= \sigma^2 C C'$

$C = [(\tilde{X}'X)^{-1} X' + D]$

$\Downarrow$

By Simplifying above eq^n

$C' = X(\tilde{X}'X)^{-1} + D$

$Var(\hat{\beta}) = \sigma^2 [(\tilde{X}'X)^{-1} X' + D] \Rightarrow C$

$[X(\tilde{X}'X)^{-1} + D'] \Rightarrow C'$

By Simplifying above eq^n

$= \sigma^2 [(\tilde{X}'X)^{-1} X'X (X'X)^{-1} +$

$(\tilde{X}'X)^{-1} X'D' +$

$DX(\tilde{X}'X)^{-1} + DD']$

$$\boxed{\begin{aligned} &Var(AX) \\ &= A\, Var(X) A \\ &(AB)' = B'A' \\ &(A^{-1})' = (A')^{-1} \end{aligned}}$$

$$\boxed{\begin{aligned} \because X'D' &= (DX)' \\ &= 0 \end{aligned}}$$

$$\text{Var}(\hat{\beta}) = \underbrace{\sigma^2 (X'X)^{-1}}_{\text{Least Square.}} + \sigma^2 DD'$$

$$\text{Var}(\hat{\beta}) \geq \text{Var}(\hat{\beta}_{\text{Least Square}})$$

$$\therefore \text{Var}(\hat{\beta}_K) \geq \sigma^2 (X_K' X_K)^{-1}$$