Introduction Statistical Linear Model

THE STATE UNIVERSITY
OF NEW JERSEY

DECEMBER 20, 2017

BY:
PRABHAT JOHL

Final Exam

Instructor: -
Mert Gurbuzbaalaban

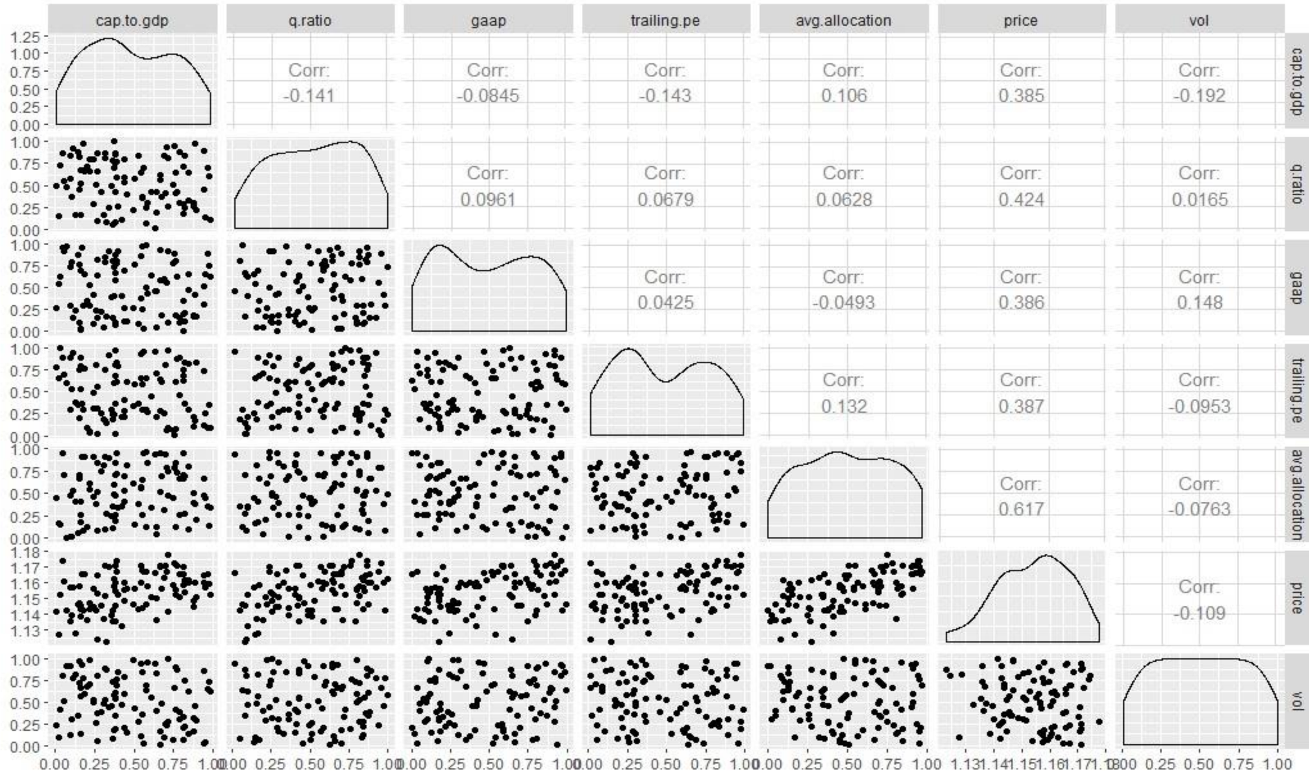
Exercise 1.

```
library(readr)
library(GGally)
library(lmPerm)
library(faraway)
library(car)
library(ggplot2)
library(gridExtra)
library(e1071)
library(zoo)
library(caret)
library(MASS)
library(ROCR)
library(pscl)
library(readxl)


stockdata <- read_csv("C:/Users/PRABHATJOHL/Desktop/Final/stockdata.csv")

summary(stockdata)

##       days          cap.to.gdp          q.ratio             gaap
##  Min.   :  1.00   Min.   :0.008325   Min.   :0.01581   Min.   :0.002853
##  1st Qu.: 25.75   1st Qu.:0.257293   1st Qu.:0.29298   1st Qu.:0.184894
##  Median : 50.50   Median :0.436999   Median :0.54062   Median :0.473457
##  Mean   : 50.50   Mean   :0.479715   Mean   :0.52838   Mean   :0.484240
##  3rd Qu.: 75.25   3rd Qu.:0.715682   3rd Qu.:0.77409   3rd Qu.:0.765638
##  Max.   :100.00   Max.   :0.980869   Max.   :0.99876   Max.   :0.991848

##   trailing.pe      avg.allocation         price            vol
##  Min.   :0.01565   Min.   :0.001411   Min.   :1.123   Min.   :0.0100
##  1st Qu.:0.24465   1st Qu.:0.264229   1st Qu.:1.145   1st Qu.:0.2575
##  Median :0.45378   Median :0.486100   Median :1.156   Median :0.5050
##  Mean   :0.48442   Mean   :0.502492   Mean   :1.154   Mean   :0.5050
##  3rd Qu.:0.75006   3rd Qu.:0.726507   3rd Qu.:1.165   3rd Qu.:0.7525
##  Max.   :0.99022   Max.   :0.964295   Max.   :1.178   Max.   :1.0000



data.frame(Variables = c("days", "cap.to.gdp", "q.ratio", "gaap", "trailing.p
e","avg.allocation","price","vol"),MissingCount = as.vector(colSums(is.na(sto
ckdata))))



##        Variables Missing Count
## 1           days             0
## 2     cap.to.gdp             0
## 3        q.ratio             0
```

```
## 4            gaap            0
## 5     trailing.pe            0
## 6  avg.allocation            0
## 7           price            0
## 8             vol            0
```

there are no missing values in the dataset.

```
ggpairs(stockdata,columns = c(2:8), lower=list(combo=wrap("facethist", binwid
th=0.8)))
```



here ggpairs is used to find out the structure of the variable, their correlation & distribution.

(a) Fit a model to explain price in terms of the predictors. Which variables are important, can any of the variables be removed ? do use F-test justify ?

```
A <- lm(price~.-days,data = stockdata)
summary(A)

##
## Call:
## lm(formula = price ~ . - days, data = stockdata)
##
## Residuals:
##        Min          1Q     Median          3Q         Max
```

```
## -0.0067787 -0.0015687  0.0002342  0.0019888  0.0075661
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.1087154  0.0012722 871.527   <2e-16 ***
## cap.to.gdp      0.0209002  0.0010535  19.839   <2e-16 ***
## q.ratio         0.0181111  0.0010414  17.391   <2e-16 ***
## gaap            0.0163251  0.0009298  17.557   <2e-16 ***
## trailing.pe     0.0143780  0.0009750  14.747   <2e-16 ***
## avg.allocation  0.0225869  0.0009978  22.637   <2e-16 ***
## vol            -0.0005667  0.0009918  -0.571    0.569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002758 on 93 degrees of freedom
## Multiple R-squared:  0.9535, Adjusted R-squared:  0.9505
## F-statistic:   318 on 6 and 93 DF, p-value: < 2.2e-16
```

Variable – Volume is not a significant predictor as its p-value is high, so we will drop this variable in our New model

```
B <- lm(price~.-days-vol,data = stockdata)
summary(B)
```

```
##
## Call:
## lm(formula = price ~ . - days - vol, data = stockdata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0066732 -0.0015245  0.0003056  0.0019045  0.0073869
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1083616  0.0011074 1000.89   <2e-16 ***
## cap.to.gdp     0.0210170  0.0010297   20.41   <2e-16 ***
## q.ratio        0.0181196  0.0010375   17.46   <2e-16 ***
## gaap           0.0162510  0.0009174   17.71   <2e-16 ***
## trailing.pe    0.0144476  0.0009639   14.99   <2e-16 ***
## avg.allocation 0.0226051  0.0009937   22.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002748 on 94 degrees of freedom
## Multiple R-squared:  0.9534, Adjusted R-squared:  0.9509
## F-statistic: 384.3 on 5 and 94 DF,  p-value: < 2.2e-16
```

From the above, summary it's clear that all the predictors are significant at low p-value.

```
anova(B,A)

## Analysis of Variance Table
##
## Model 1: price ~ (days + cap.to.gdp + q.ratio + gaap + trailing.pe + avg.a
llocation + vol) - days

## Model 2: price ~ (days + cap.to.gdp + q.ratio + gaap + trailing.pe + avg.a
llocation + vol) - days - vol

##   Res.Df        RSS Df   Sum of Sq      F Pr(>F)
## 1     93 0.00070738
## 2     94 0.00070986 -1 -2.4832e-06 0.3265 0.5691
```

Here we are using Anova Function for conducting F- utility test of the above Model's, it's clear the p-value = 0.56 is high & f-ratio = 0.32 is drastically low. which states that model B is better & stable than model A. here we are accepting null hypothesis (small model B) against alternative hypothesis (big model A).

```
var.test(stockdata$price, stockdata$vol, alternative = "two.sided")

##  F test to compare two variances
##
## data: price and volume
## F = 0.0018266, num df = 99, denom df = 99, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.001228979 0.002714681
## sample estimates:
## ratio of variances
##        0.001826551
```

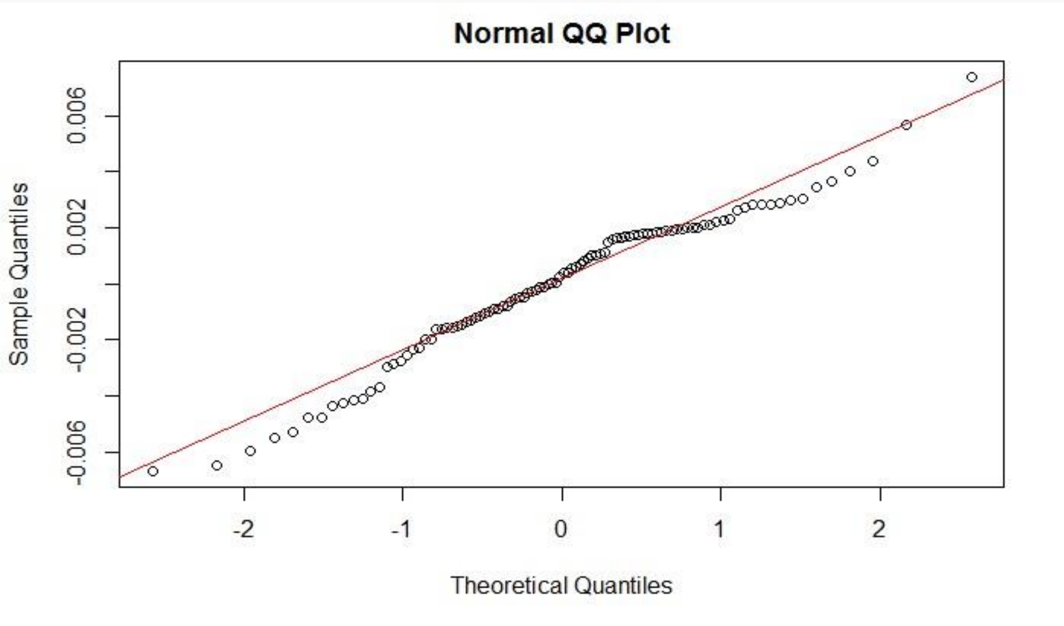
the p-value < 2.2e-16, which states that there is a significant difference between the two variances of price & volume variable.

(b) Construct confidence intervals using permutation tests?

```
C<-aovp(price~.-vol-days,data=stockdata)

## [1] "Settings:  unique SS : numeric variables centered"

summary(C)

## Component 1 :
##                 Df   R Sum Sq R Mean Sq Iter  Pr(Prob)
## cap.to.gdp       1 0.0031459 0.0031459 5000 < 2.2e-16 ***
## q.ratio          1 0.0023032 0.0023032 5000 < 2.2e-16 ***
```

```
## gaap            1 0.0023695 0.0023695 5000 < 2.2e-16 ***
## trailing.pe     1 0.0016967 0.0016967 5000 < 2.2e-16 ***
## avg.allocation  1 0.0039080 0.0039080 5000 < 2.2e-16 ***
## Residuals      94 0.0007099 0.0000076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qqnorm(resid(C),main="Normal QQ Plot")
qqline(resid(C),col='red')
```



here we are using AOVP function to perform permutation test & check for normality behavior using q-q plot.

```
confint(C)
```

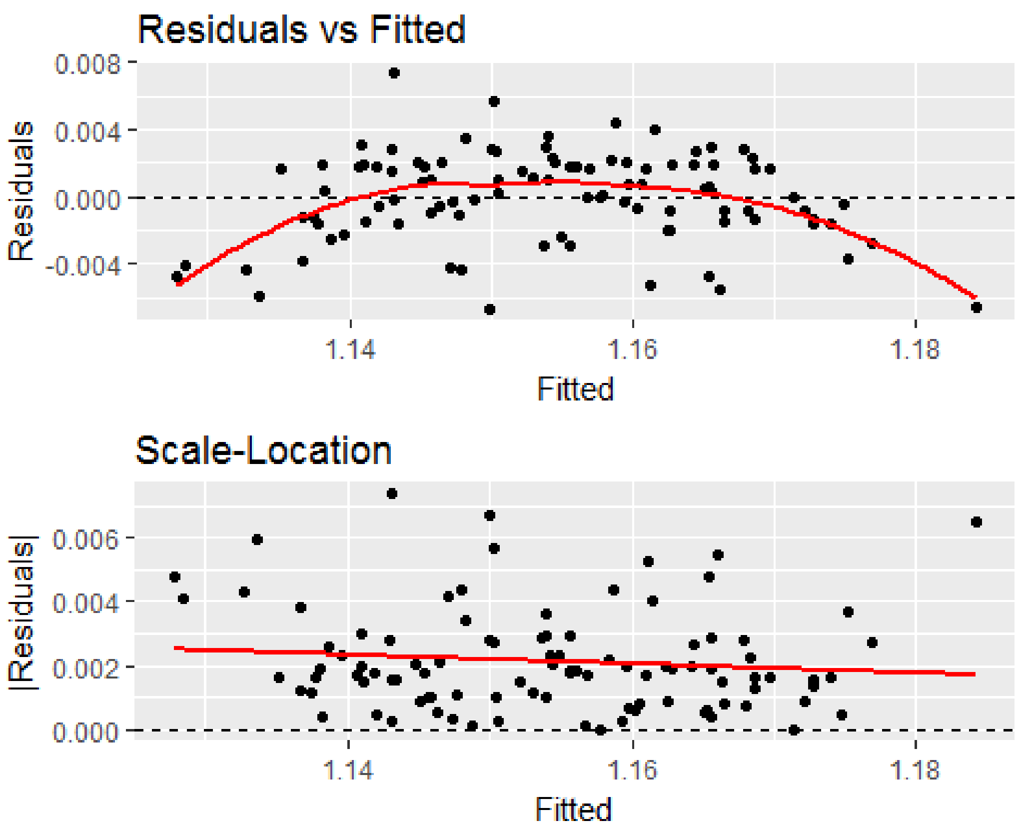
```
##                     2.5 %     97.5 %
## (Intercept)    1.15369917 1.15479043
## cap.to.gdp     0.01897251 0.02306158
## q.ratio        0.01605950 0.02017963
## gaap           0.01442942 0.01807256
## trailing.pe    0.01253386 0.01636137
## avg.allocation 0.02063209 0.02457811
```

above provides the confidence interval generated using permutation test.

(c) Check the constant variance assumption for the errors?

```
mod <- fortify(B)
```

```r
p1 <- qplot(.fitted, .resid, data = mod) + geom_hline(yintercept = 0, linetyp
e = "dashed") + labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residu
als") + geom_smooth(color = "red", se = F)

p2 <- qplot(.fitted, abs(.resid), data = mod) + geom_hline(yintercept = 0, li
netype = "dashed") +labs(title = "Scale-Location", x = "Fitted", y = "|Residu
als|") + geom_smooth(method = "lm", color = "red", se = F)

grid.arrange(p1, p2, nrow = 2)
```



From, the above Residual vs Fitted, it can be noted that the residuals are randomly scattered. The redline depicts the behavior of the model, from the trend line it's clear that residuals float along the trend line. Which states that there is no discrete pattern in the residuals, hence no presence of heteroskedasticity. So we don't need to undergo transform (log,sqrt,sqr) on response variable. A floor effect of residuals is depicted from scale-location graph.

```r
ncvTest(B)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
```

```
## Chisquare = 0.7231311    Df = 1     p = 0.3951188
```

here chi-square value is very low, depicting that is no abnormal pattern among the residuals showing no presence heteroskedasticity.

Approximate test of non-constant error variance.

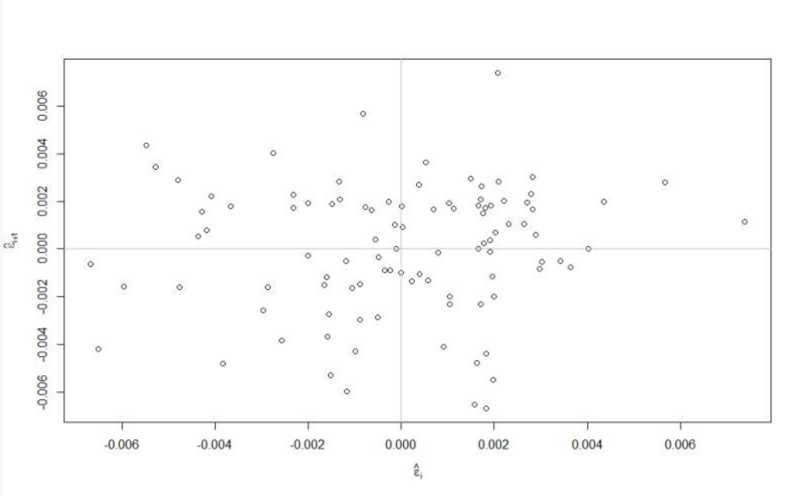
```
summary(lm(abs(residuals(B)) ~ fitted(B)))

##
## Call:
## lm(formula = abs(residuals(B)) ~ fitted(B))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0020725 -0.0011673 -0.0002912  0.0007120  0.0050963
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01837    0.01535   1.197    0.234
## fitted(B)   -0.01407    0.01329  -1.058    0.293
##
## Residual standard error: 0.001601 on 98 degrees of freedom
## Multiple R-squared:  0.0113, Adjusted R-squared:  0.001211
## F-statistic:  1.12 on 1 and 98 DF,  p-value: 0.2925
```

From, the above approximation test for non-constant variance. It can that t-value (-1.058) is very low, p-value is a bit high so we can’t reject the null hypothesis of fitted(B1) being zero.

(d) Check the independentness of the errors assumption?

```
res = residuals(B)
nres = length(res)
summary(lm (tail(res,nres-1) ~ head(res, nres-1)))

##
## Call:
## lm(formula = tail(res, nres - 1) ~ head(res, nres - 1))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0068179 -0.0013705  0.0002232  0.0018688  0.0072227
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.141e-06  2.711e-04   0.015    0.988
## head(res, nres - 1) 7.714e-02  1.016e-01   0.759    0.450
##
```

```
## Residual standard error: 0.002697 on 97 degrees of freedom
## Multiple R-squared:  0.005907,   Adjusted R-squared:  -0.004341
## F-statistic: 0.5764 on 1 and 97 DF,  p-value: 0.4496
```



Random Error co-Relation Pattern.

```
durbinWatsonTest(B)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.07656505      1.839522   0.422
##  Alternative hypothesis: rho != 0
```

Both the method, generate the same result with different interpretation.

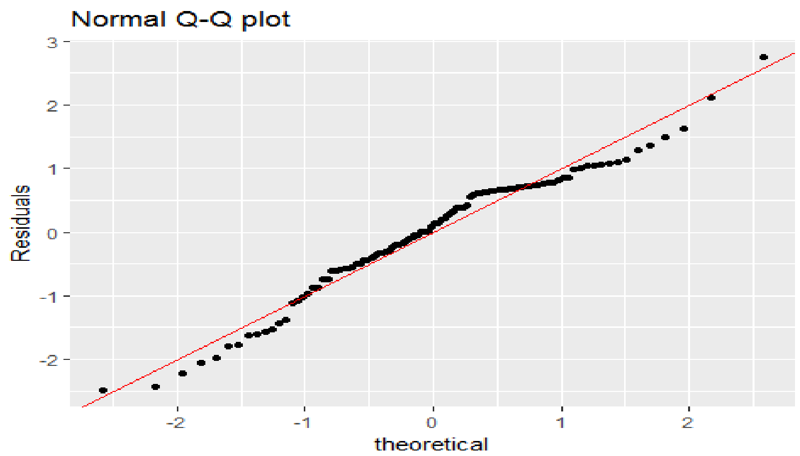
Durbin Watson test is used for finding out if the residuals of a regression model are correlated or not.

- **The null hypothesis (H0) is that there is no correlation among residuals, i.e., they are independent.**
- **The alternative hypothesis (Ha) is that residuals are autocorrelated.**

From our test on model B1, the P-value is a bit high (0.44), so we reject alternative hypothesis that residuals are correlated because the Auto-correlation factor is 0.076 very low & D-W statistic (1.85) is near to 2 However, the accuracy depends on the normality & unbiasedness Assumption of model.

(e) Check the normality assumption?

```
p3 <- qplot(sample = scale(.resid), data = mod) + geom_abline(intercept = 0,
slope = 1, color = "red") + labs(title = "Normal Q-Q plot", y = "Residuals")
p3
```

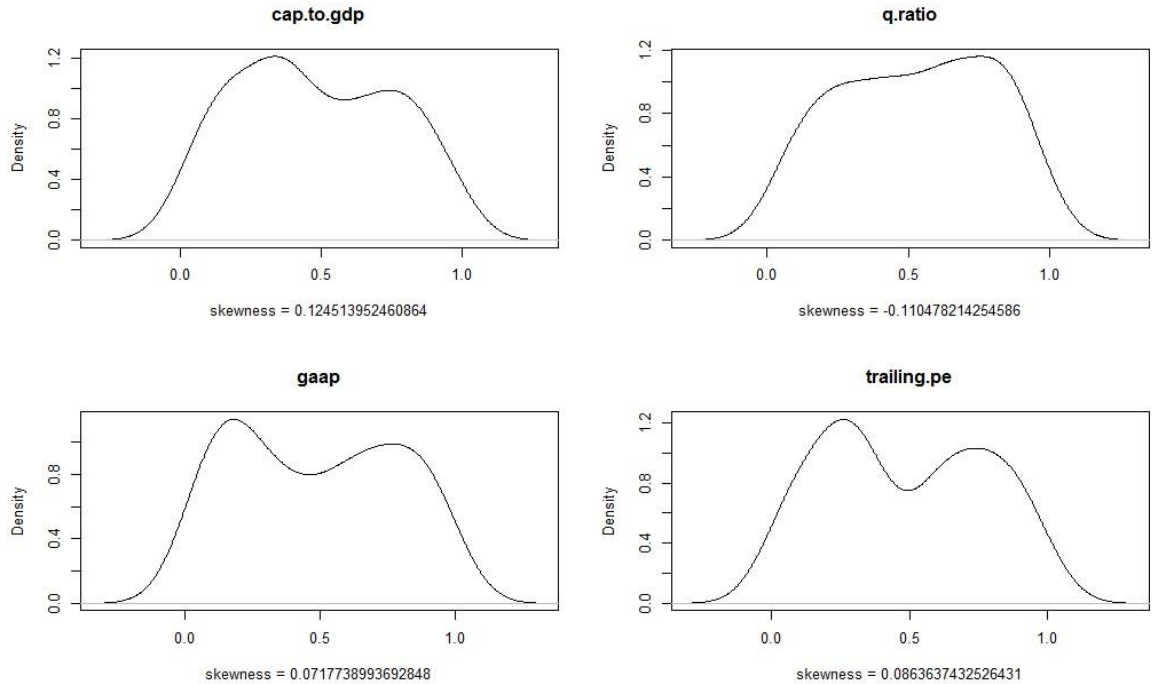



```
From the above Normal Q-Q Plot, our base model B follows a normal
distribution along the standardized residuals line. There are few amounts of
outlier & influential points in the model
```

```
shapiro.test(residuals(B))

##  Shapiro-Wilk normality test
## data:  residuals(B)
## W = 0.97164, p-value = 0.02955

After performing shapiro test on our model, it can be seen that p-value of
our model is optimal & test statistic w (0.97) is high near to 1, depicting a
strong evidence that a non-normal distribution behavior doesn’t exist. So,
normality assumption is true.
```

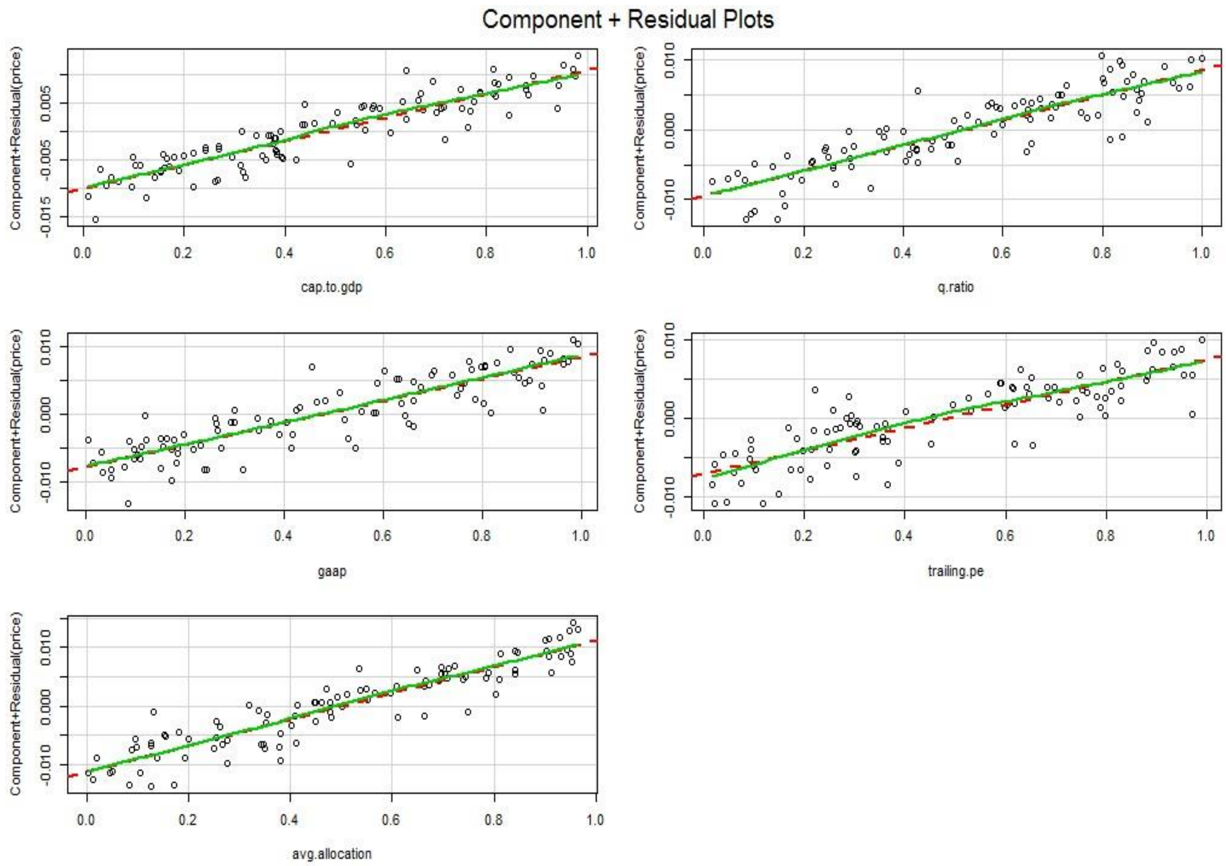
(f) Is non-linearity a problem?

```
par(mfrow=c(2,2))
plot(density(stockdata$cap.to.gdp),main = "cap.to.gdp",xlab = paste("skewness
=",skewness(stockdata$cap.to.gdp)))
plot(density(stockdata$q.ratio),main = "q.ratio",xlab = paste("skewness =",sk
ewness(stockdata$q.ratio)))
plot(density(stockdata$gaap),main = "gaap",xlab = paste("skewness =",skewness
(stockdata$gaap)))
plot(density(stockdata$trailing.pe),main = "trailing.pe",xlab = paste("skewne
ss =",skewness(stockdata$trailing.pe)))
```



The plot displays the distribution of various predictors, All the predictors are normally distributed with skewness of all predictors being very low & near to “0”. So, no requirement of any transform.

```
crPlots(B)
```



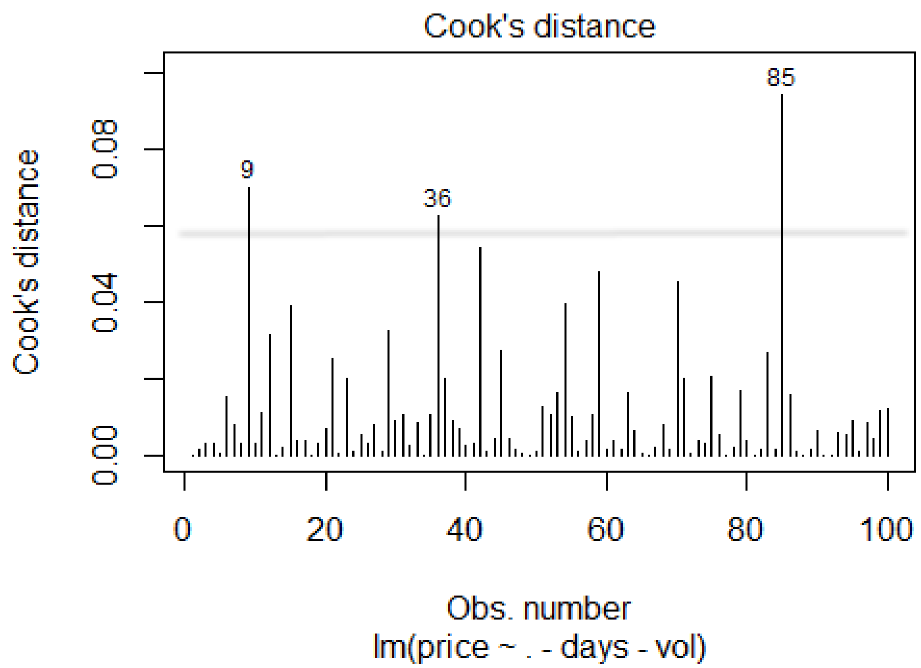
the structure of relationship between response & predictors is highly linear so nonlinearity won't be any problem.

(g) Check for outliers, compute and plot Cook's distance?

```
outlierTest(B)

## No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 59 2.840296          0.0055368      0.55368
```

Using Outlier test, observation number "59" is an outlier for our model B as per standardized residuals deviation.

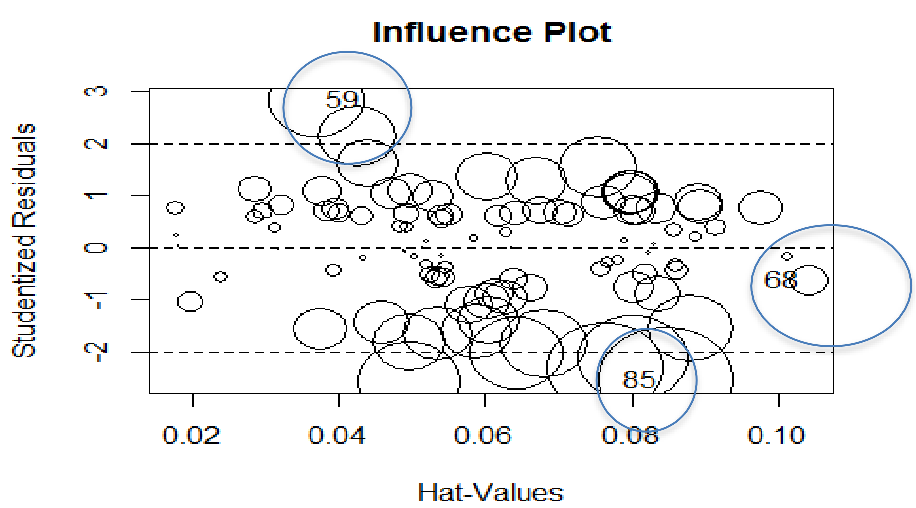
```
plot(B, which=4,cook.levels=0.06)
```



As per cook's distance residuals deviation, observation number: 9, 36, 85 are the outlier for our model.

(h) Check for influential points?

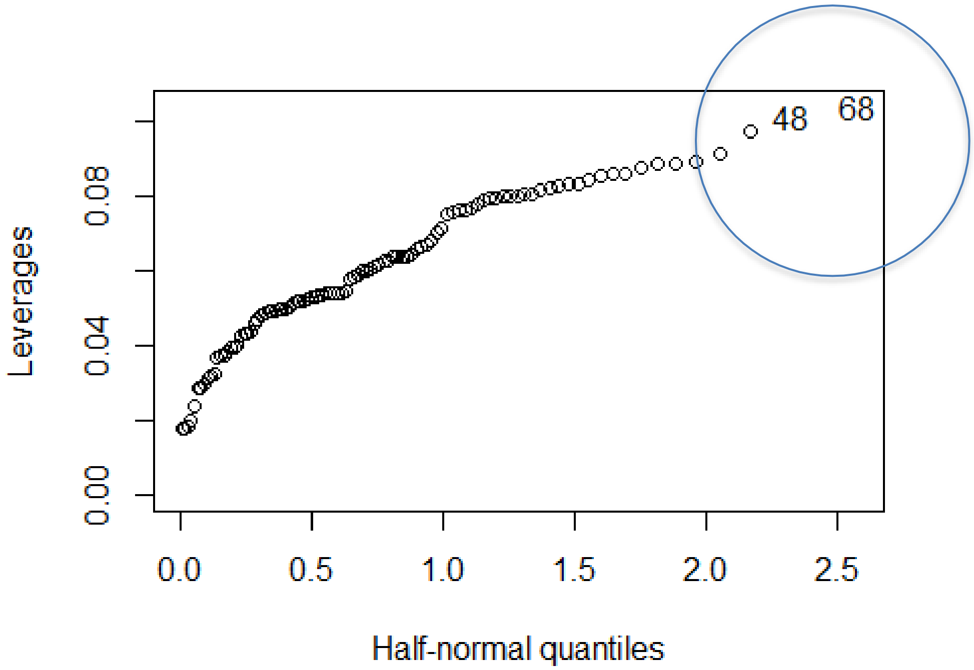
```
influencePlot(B,main="Influence Plot")
```



```
         StudRes        Hat      CookD
## 59  2.8402957 0.03698628 0.04802875
## 68 -0.6296949 0.10409891 0.00772845
## 85 -2.5435520 0.08469051 0.09428313
```

Above are the influential points that affect our model

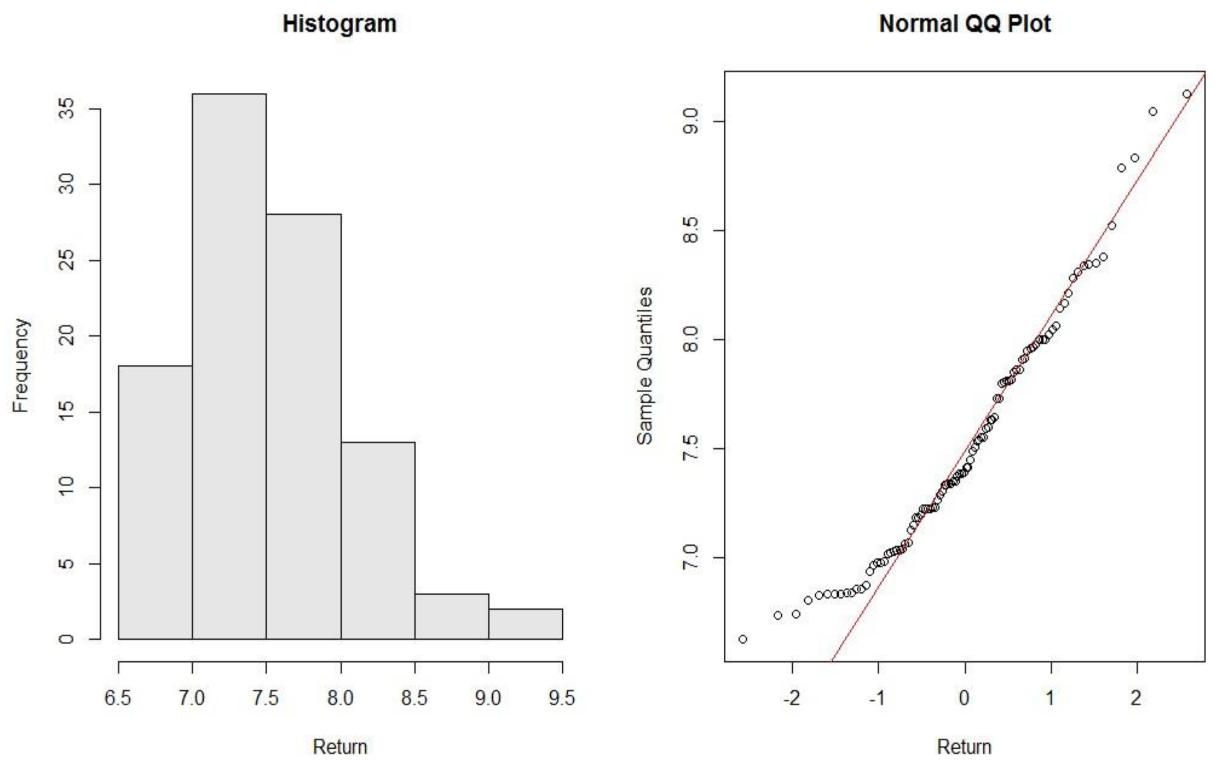
```
halfnorm(lm.influence(B)$hat, ylab = "Leverages")
```



observation: 48, 68 are the high Leverage points

(i) The return at time t is defined as r(t) = p(t + 1)/p(t)-1 where p is the price data for day t. Are the returns normally distributed? Please justify your answer using Q-Q plots.

```
E<- (stockdata$price*+1)/(stockdata$price-1)

par(mfrow=c(1,2))

hist(E, main="Histogram",xlab = "Return",col =gray(0.9))
qqnorm(E,main="Normal QQ Plot",xlab = "Return") + qqline(E,col='red')
```



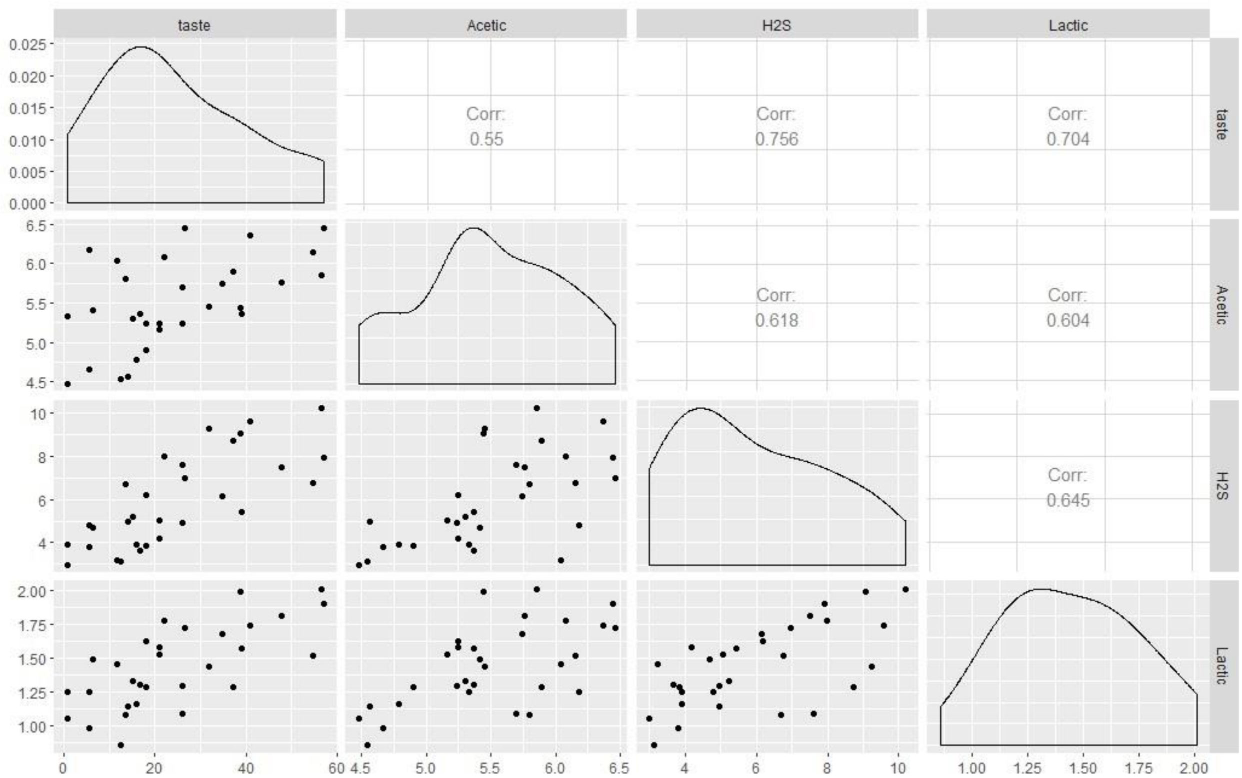
The histogram shows the distribution of various return, to future analysis the pattern we plot the normal Q-Q plot for the return. All the points are colligated along the red standardized line which depicts that return follows a normal distribution behavior.

Exercise 2.

```
data("cheddar")
summary(cheddar)

##      taste           Acetic           H2S             Lactic     
##  Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860  
##  1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250  
##  Median :20.95   Median :5.425   Median : 5.329   Median :1.450  
##  Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442  
##  3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667  
##  Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010  

data.frame(Variables = c("taste", "Acetic", "H2S", "Lactic"),
MissingCount = as.vector(colSums(is.na(cheddar))))

##   Variables MissingCount
## 1     taste            0
## 2    Acetic            0
## 3       H2S            0
## 4    Lactic            0
```

There is No Missing Value in the dataset.

```
ggpairs(cheddar,columns = c(1:4), lower=list(combo=wrap("facethist", binwidth
=0.8)))
```



The Above ggpairs displays the predictor's distribution & their correlation. H2S has the highest co-relation with Taste.

(a) Fit a model to explain taste in terms of the predictors. Which variables are important, can any of the variables be removed ?

```
F <- lm(taste~.,data = cheddar)
summary(F)

##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF, p-value: 3.81e-06
```

It can be seen from the summary that Acetic Predictor is not significant, so we will remove that variable in our new model.

```
F1 <- lm(taste~.-Acetic,data = cheddar)
summary(F1)

##
## Call:
## lm(formula = taste ~ . - Acetic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.343  -6.530  -1.164   4.844  25.618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.592      8.982  -3.072  0.00481 **
## H2S            3.946      1.136   3.475  0.00174 **
## Lactic        19.887      7.959   2.499  0.01885 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.942 on 27 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6259
## F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

The new model F1, has all the predictor's that are significant.

```
anova(F,F1)
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ Acetic + H2S + Lactic
## Model 2: taste ~ (Acetic + H2S + Lactic) - Acetic
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 2668.4
## 2     27 2669.0 -1  -0.55427 0.0054  0.942
```

As per Anova summary, the p-value obtained is very high & F-ratio is very low. thus as per F-utility test of model, we accept alternative hypothesis of model F1 to be better than F.

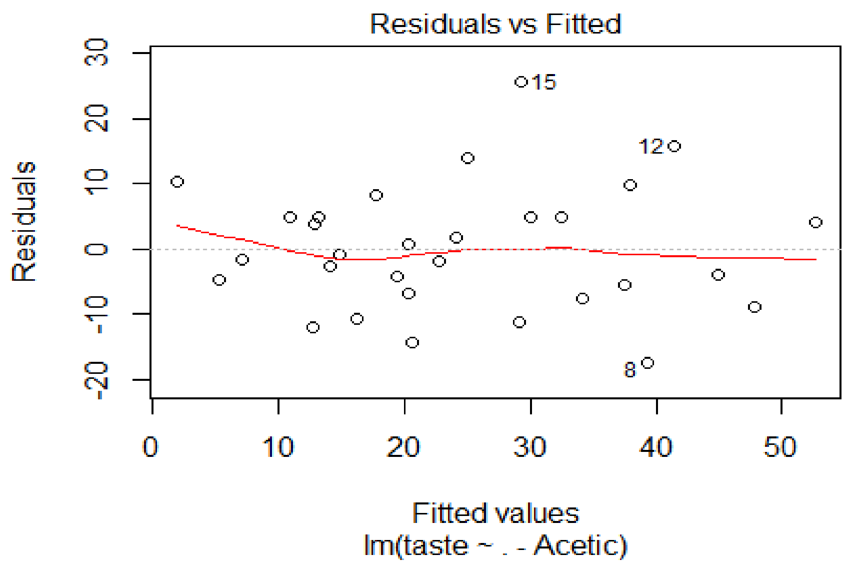
```
var.test(cheddar$taste, cheddar$Acetic, alternative = "two.sided")
```

```
##  F test to compare two variances

## data:  taste and Acetic
## F = 810.79, num df = 29, denom df = 29, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   385.9065 1703.4619
## sample estimates:
## ratio of variances
##           810.7879
```

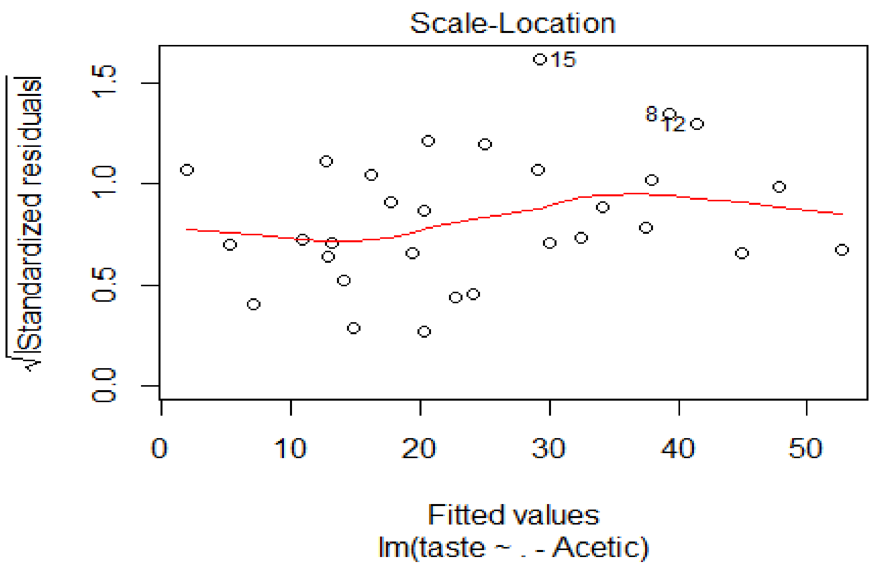
The above Variable – taste & Acetic have significantly differ in Variance as the P-value obtained from the taste is very low.

(b) Check the constant variance assumption for the errors?

```
plot(F1,which=1)
```




```
plot(F1,which = 3)
```



From the above Plots, all the residuals are randomly scattered showing absence of any abnormal pattern. Thus, there is no presence of heteroskedasticity. So, there is no requirement of any transform on Response.

```
ncvTest(F1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted. values
## Chi-square = 1.181682    Df = 1     p = 0.2770139
```

The Chi-square value obtained from ncvtest is very low, which aligns with our above plots that constant variance assumption among residuals is True.

Approximate test of non-constant error variance.

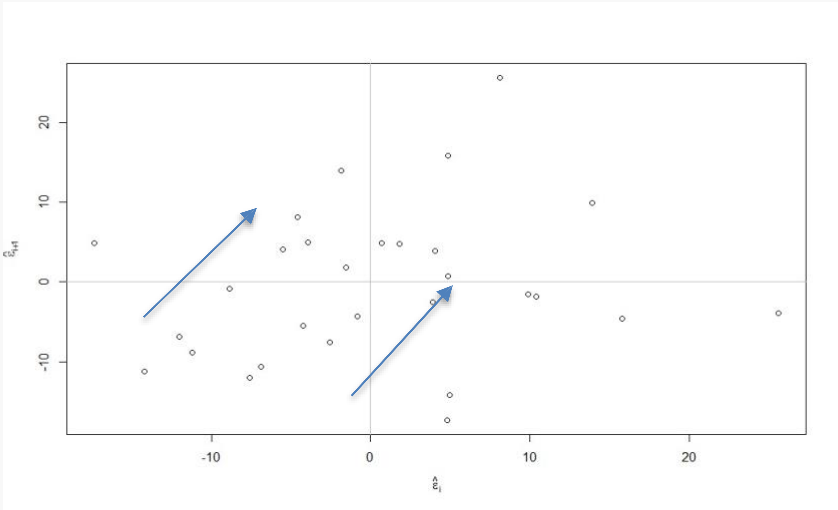
```
summary(lm(abs(residuals(F1)) ~ fitted(F1)))

## Call:
## lm(formula = abs(residuals(F1)) ~ fitted(F1))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.467 -3.900 -1.165  3.713 17.567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.16309    2.21835   2.327   0.0274 *
## fitted(F1)   0.09862    0.08003   1.232   0.2281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.656 on 28 degrees of freedom
## Multiple R-squared:  0.05144,    Adjusted R-squared:  0.01756
## F-statistic: 1.518 on 1 and 28 DF, p-value: 0.2281
```

The t-value is very low which depicts that null hypothesis is true for fitted F1 ~ 0.

(C) Check the independents of the errors assumption?

```
res = residuals(F1)
nres = length(res)
summary(lm (tail(res,nres-1) ~ head(res, nres-1)))

##
## Call:
## lm(formula = tail(res, nres - 1) ~ head(res, nres - 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.769  -6.713  -2.809   5.472  24.602
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.4243     1.7799  -0.238    0.813
## head(res, nres - 1)    0.1771     0.1896   0.934    0.359
##
## Residual standard error: 9.578 on 27 degrees of freedom
## Multiple R-squared:  0.03129,    Adjusted R-squared:  -0.004586
## F-statistic: 0.8722 on 1 and 27 DF,  p-value: 0.3586
```

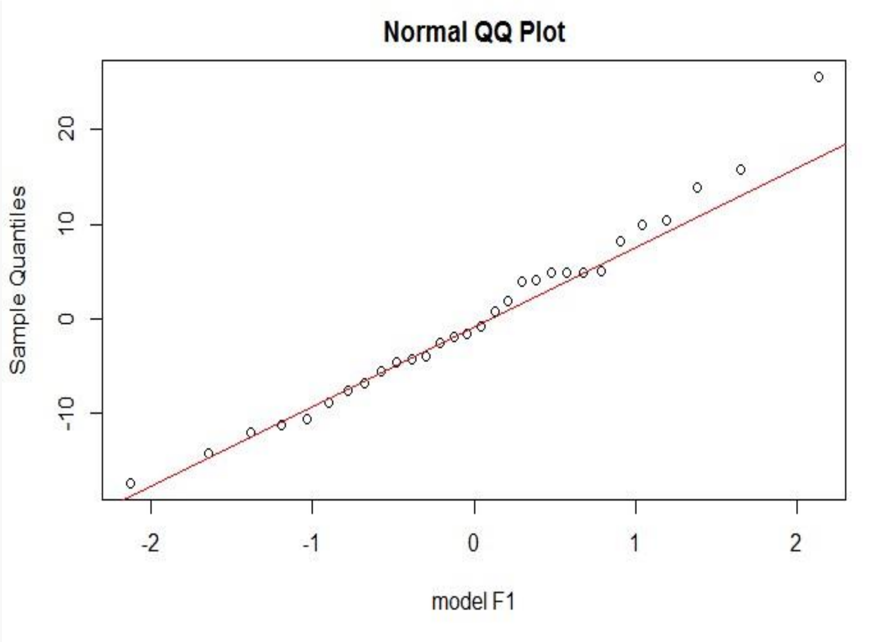


```
durbinWatsonTest(F1)

##  lag Autocorrelation D-W Statistic p-value
##    1        0.167847      1.581086   0.216
##  Alternative hypothesis: rho != 0
```

As per Durbin-Watson test, the D-W statistic is 1.58 which states that there is a weak positive Auto-correlation among the Residuals.

(d) Check the normality assumption?

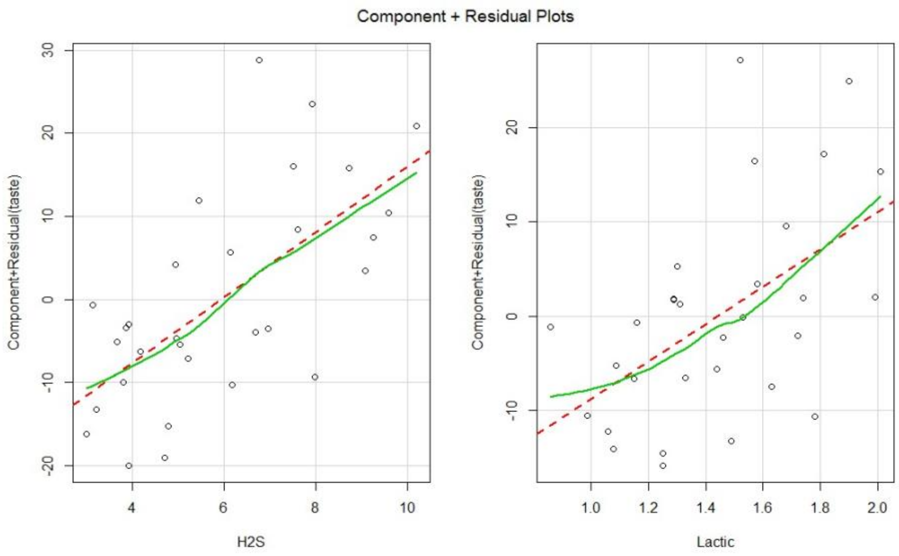
```
qqnorm(residuals(F1),main="Normal QQ Plot",xlab = "model F1")
qqline(residuals(F1),col='red')
```



From the above Plot, it’s clear that the residuals of model F1 are reasonably normal.

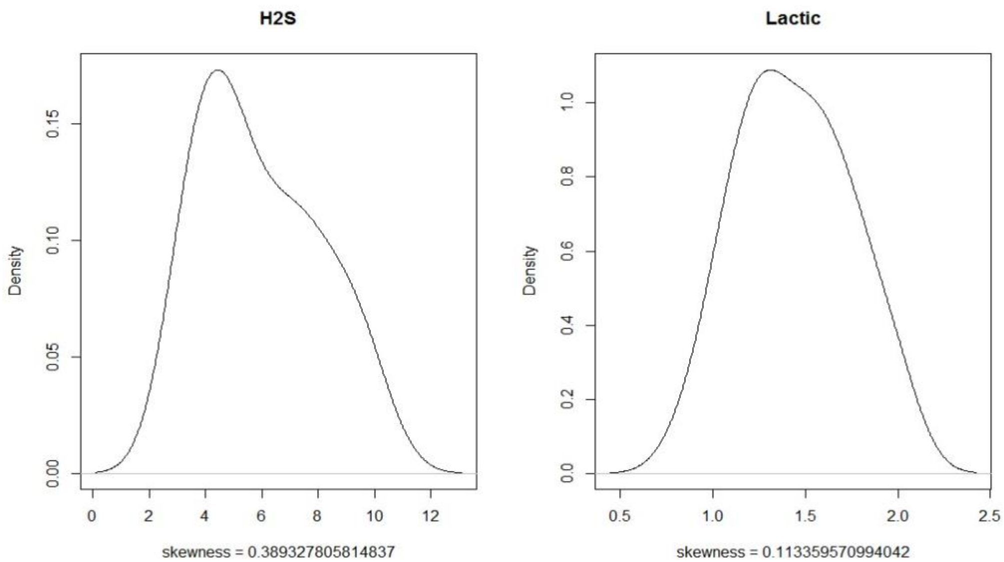
(e) Is nonlinearity a problem?

```
crPlots(F1)
```



The CR plot shows that all the predictors gets aligned along the standardized line with no deviation.

```
par(mfrow=c(1,2))

plot(density(cheddar$H2S),main = "H2S",xlab = paste("skewness =",skewness(che
ddar$H2S)))

plot(density(cheddar$Lactic),main = "Lactic",xlab = paste("skewness =",skewne
ss(cheddar$Lactic)))
```



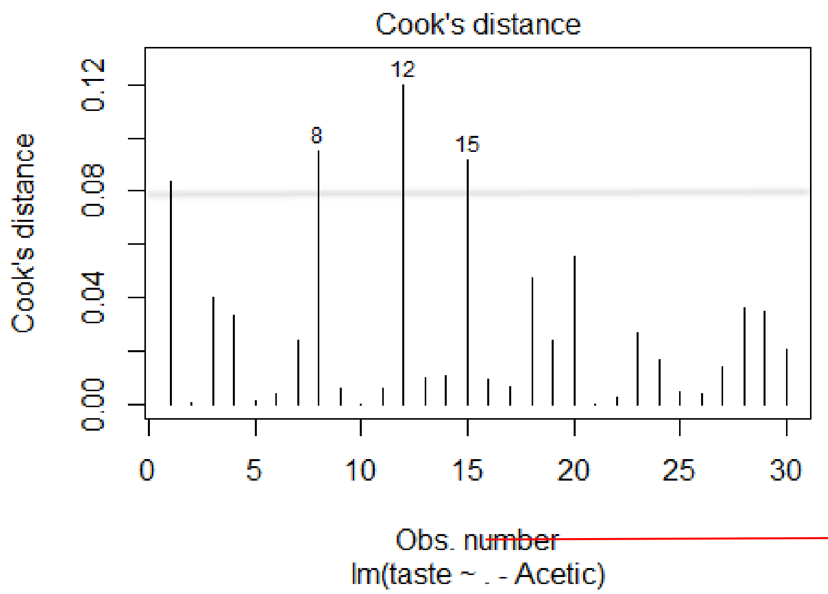
The predictors are normally distributed with low skewness, thus from above plots the structure of response & predictors is highly linear so no requirement of any transform on predictors.

(g) Check for outliers, compute and plot Cook's distance, Influencial points?

```
outlierTest(F1)

## No Studentized residuals with Bonferonni p < 0.05
   Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
  15  2.98867          0.0060495      0.18148
```

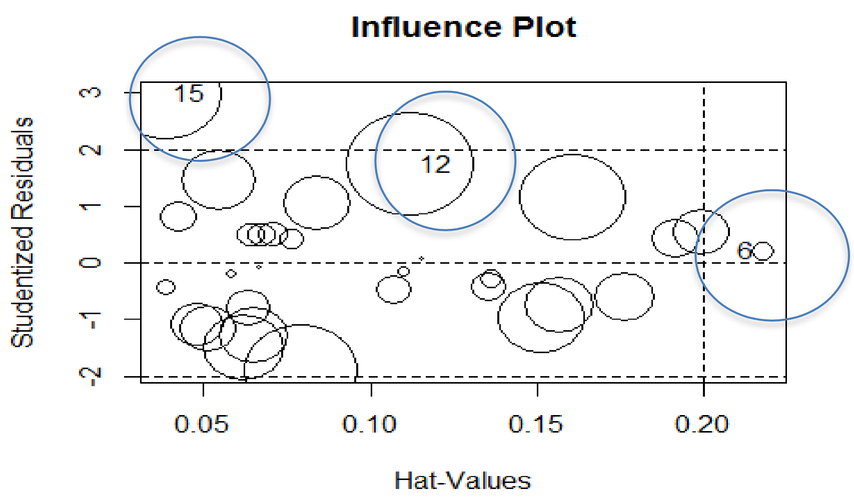
Observation 15 is an outlier for our model F1 , as per Studentized Residuals.

```r
plot(F1, which=4,cook.levels=0.08)
```



Observation: 8,12,15 are outlier's as per cook's distance.

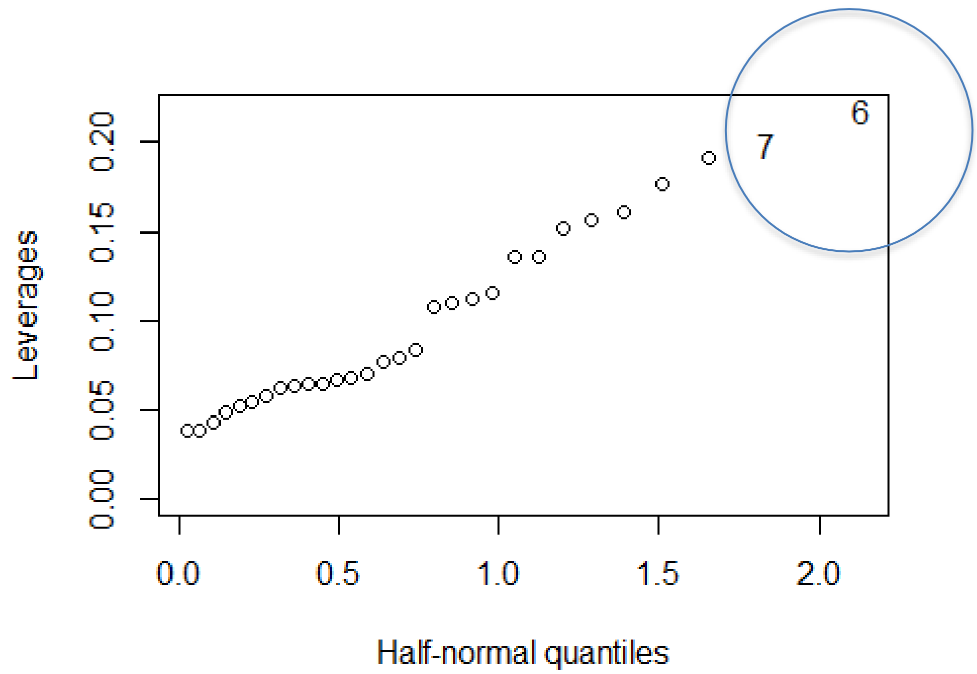
```r
influencePlot(F1,main="Influence Plot")
```




```
        StudRes        Hat       CookD
## 6  0.2031432 0.21752949 0.003964927
## 12 1.7494704 0.11200441 0.119556929
## 15 2.9886698 0.03834518 0.091762390
```

Above are the Influential points that can affect our model F1.

```
halfnorm(lm.influence(F1)$hat, ylab = "Leverages")
```



Exercise 3.

(a) Data preparation: combine all data into an R data frame object, and construct dummy or factor variable for 4 quarters. First model is HOUST ~ GDP + CPI + quarter?

```
CPI <- read_excel("C:/Users/PRABHATJOHL/Desktop/Final/House/CPI.xls")
GDP <- read_excel("C:/Users/PRABHATJOHL/Desktop/Final/House/GDP.xls")
HOUST <-
read_excel("C:/Users/PRABHATJOHL/Desktop/Final/House/HOUST.xls") POP <-
read_excel("C:/Users/PRABHATJOHL/Desktop/Final/House/POP.xls")

View(head(GDP))
```

DATE	GDP
1976-01-01	58.6
1976-04-01	32.4
1976-07-01	33.6
1976-10-01	47.9
1977-01-01	54.1
1977-04-01	67.7

```
View(head(CPI))
```

DATE	CPI
1976-01-01	0.633
1976-04-01	0.500
1976-07-01	0.900
1976-10-01	0.833
1977-01-01	1.067
1977-04-01	1.033

```
View(head(HOUST))
```

DATE	HOUST
1975-10-01	296.6
1976-01-01	280.8
1976-04-01	439.3
1976-07-01	434.3
1976-10-01	382.9
1977-01-01	367.4

```
View(head(POP))
```

DATE	POPULATION
1976-01-01	462
1976-04-01	562
1976-07-01	579
1976-10-01	510
1977-01-01	529
1977-04-01	617

```r
df= merge(x = CPI,y= GDP, by.x = "DATE" , by.y = "DATE" ,all="TRUE")
df= merge(x = df , y= HOUST , by.x = "DATE" , by.y= "DATE", all = "TRUE")
df = merge(x=df, y =POP , by.x="DATE" , by.y= "DATE", all="TRUE")

df = na.omit(df)
summary(df)
```

```
##       DATE                      CPI             GDP
##  Min.   :1976-01-01    Min.   :-5.012   Min.   :-293.10
##  1st Qu.:1985-12-09    1st Qu.: 0.833   1st Qu.:  62.27
##  Median :1995-11-16    Median : 1.125   Median : 101.20
##  Mean   :1995-11-15    Mean   : 1.143   Mean   : 102.86
##  3rd Qu.:2005-10-24    3rd Qu.: 1.500   3rd Qu.: 140.07
##  Max.   :2015-10-01    Max.   : 3.323   Max.   : 283.80

##      HOUST          POPULATION
##  Min.   :114.4   Min.   :441.0
##  1st Qu.:274.5   1st Qu.:574.0
##  Median :357.4   Median :650.5
##  Mean   :351.9   Mean   :662.0
##  3rd Qu.:440.4   3rd Qu.:746.8
##  Max.   :624.5   Max.   :947.0
```

```r
View(head(df))
```

DATE	CPI	GDP	HOUST	POPULATION
1976-01-01	0.633	58.6	280.8	462
1976-04-01	0.500	32.4	439.3	562
1976-07-01	0.900	33.6	434.3	579
1976-10-01	0.833	47.9	382.9	510
1977-01-01	1.067	54.1	367.4	529
1977-04-01	1.033	67.7	581.1	617

```r
df$QUARTER = as.yearqtr(df$DATE, format = "%Y-%m-%d")
df$QUARTER = as.numeric(format(df$QUARTER, format="%q"))
df$QUARTER = as.factor(df$QUARTER)
df$DATE = NULL

View(head (df))
```

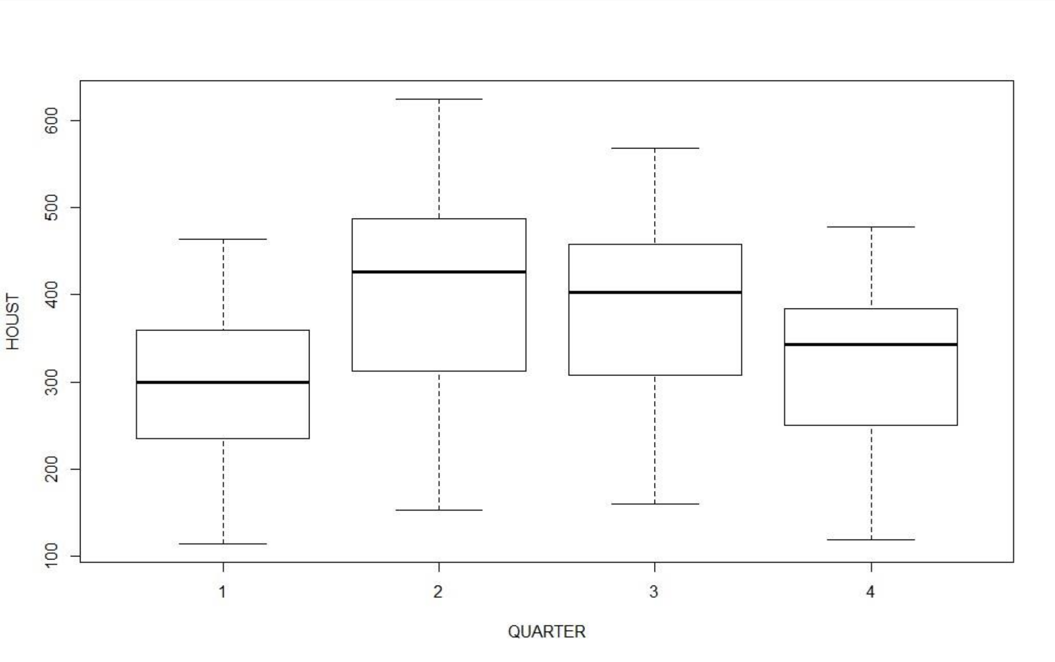
CPI	GDP	HOUST	POPULATION	QUARTER
0.633	58.6	280.8	462	1
0.500	32.4	439.3	562	2
0.900	33.6	434.3	579	3
0.833	47.9	382.9	510	4
1.067	54.1	367.4	529	1
1.033	67.7	581.1	617	2

```r
G <- lm(HOUST ~ GDP+CPI+QUARTER , data = df)
summary(G)
```

```
## Call:
## lm(formula = HOUST ~ GDP + CPI + QUARTER, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -266.68  -69.19   12.62   71.28  217.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 271.0686    20.9148  12.961  < 2e-16 ***
## GDP           0.2208     0.1207   1.829 0.069328 .
## CPI           1.8468     9.8302   0.188 0.851224
## QUARTER2    105.3363    23.5381   4.475 1.48e-05 ***
## QUARTER3     88.2852    23.4548   3.764 0.000237 ***
## QUARTER4     30.4973    23.4202   1.302 0.194801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.4 on 154 degrees of freedom
## Multiple R-squared:  0.1782, Adjusted R-squared:  0.1515
## F-statistic: 6.677 on 5 and 154 DF,  p-value: 1.173e-05
```

(b) Use one-way ANOVA to determine whether there's a seasonal effect. Show necessary steps and explanation?

```r
Boxplot(HOUST ~ QUARTER, df)
```



It can be clearly seen that during quarter 2 (April -June), there was a high unit housing price compare to other quarter. During the pike winter season in united states (October-march). the unit housing start price was low. So, seasonality does have an impact.

```
G2 <- lm(HOUST~QUARTER,data = df)
summary(G2)

##
## Call:
## lm(formula = HOUST ~ QUARTER, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250.32  -76.57   17.00   72.60  220.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   292.88      16.61  17.633  < 2e-16 ***
## QUARTER2      111.24      23.49   4.736 4.87e-06 ***
## QUARTER3       92.36      23.49   3.932 0.000126 ***
## QUARTER4       32.51      23.49   1.384 0.168265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 156 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1409
## F-statistic: 9.696 on 3 and 156 DF,  p-value: 6.625e-06
```

It can be clearly seen that Quarter 2 & 3 have high coefficient estimate. so for 1 unit increase in response(HOUST),there is a drastic increase for Quarter 2 & 3 Values in align with Intercept.

```
anova(G2)

Analysis of Variance Table
##
## Response: HOUST
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## QUARTER     3  320981  106994  9.6956 6.625e-06 ***
## Residuals 156 1721498   11035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


round(coef(G),1)

## (Intercept)         GDP         CPI                  QUARTER3    QUARTER4
##       271.1         0.2         1.8                      88.3        30.5
```

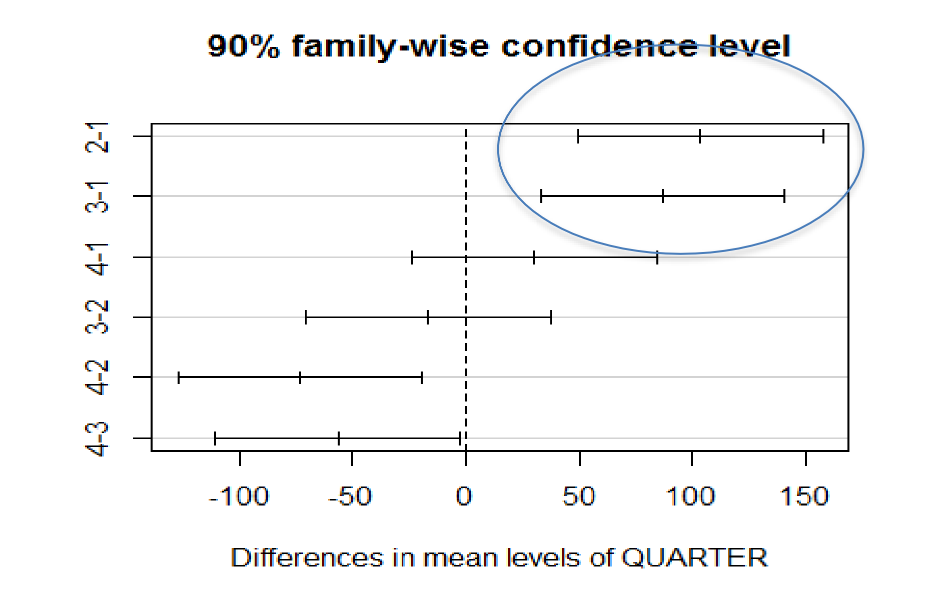
```r
anova(G)

## Analysis of Variance Table
##
## Response: HOUST
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP         1   76745   76745  7.0408  0.008803 **
## CPI         1    2025    2025  0.1858  0.667068
## QUARTER     3  285122   95041  8.7194 2.229e-05 ***
## Residuals 154 1678588   10900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that there is indeed a difference in the quarters and thus we state that there is seasonal effect

(c) Do pair-wise comparison for various levels. Construct %90 confidence intervals for the pairwise differences?

```r
#The function TukeyHD() takes the fitted ANOVA as an argument.

X = TukeyHSD(aov(G),conf.level = 0.90)
summary(X)

##   Tukey multiple comparisons of means,90% family-wise confidence level
## Fit: aov(formula = G)

$QUARTER
##          diff        lwr        upr     p adj
## 2-1 103.40958   49.46254 157.356616 0.0001042
## 3-1  86.73156   32.78452 140.678595 0.0016008
## 4-1  30.09146  -23.85557  84.038500 0.5713311
## 3-2 -16.67802  -70.62506  37.269016 0.8912549
## 4-2 -73.31812 -127.26515 -19.371079 0.0107807
## 4-3 -56.64009 -110.58713  -2.693057 0.0764098
```

```
plot(X)
```



From the above Plot, Quarter 2-1 & 3-1 have the largest confidence interval a long with p-value being significantly low.

(d) Add population to the first model, do the steps (b) and (c) again?

```
G3 <- lm(HOUST ~ GDP+CPI+QUARTER+POPULATION , data = df)
summary(G3)

## Call:
## lm(formula = HOUST ~ GDP + CPI + QUARTER + POPULATION, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -271.25   -64.11    15.78    70.93   213.90
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.00784   53.40285   5.599 9.73e-08 ***
## GDP           0.22720    0.12149   1.870  0.06337 .
## CPI           1.88789    9.85210   0.192  0.84829
## QUARTER2    109.61624   24.76083   4.427 1.81e-05 ***
## QUARTER3     91.91961   24.35938   3.773  0.00023 ***
## QUARTER4     28.98273   23.62236   1.227  0.22174
## POPULATION   -0.04569    0.08032  -0.569  0.57031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.6 on 153 degrees of freedom
## Multiple R-squared:  0.1799, Adjusted R-squared:  0.1477
```

```
## F-statistic: 5.594 on 6 and 153 DF,  p-value: 2.852e-05
```

Adding Population predictor to the model does not much of R^2 value.

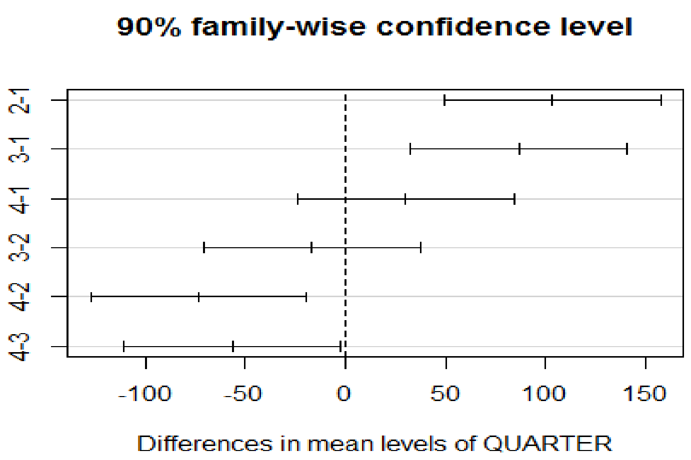
```r
compareCoefs(G3,G2,se=FALSE)
```

```
## Call:
## 1: lm(formula = HOUST ~ GDP + CPI + QUARTER + POPULATION, data = df)
## 2: lm(formula = HOUST ~ QUARTER, data = df)
##                Est. 1   Est. 2
## (Intercept) 299.0078 292.8850
## GDP           0.2272
## CPI           1.8879
## QUARTER2    109.6162 111.2400
## QUARTER3     91.9196  92.3625
## QUARTER4     28.9827  32.5150
## POPULATION   -0.0457
```

Above we are comparing Coefficients of the models G2 & G3. Population estimate is very small so has negligible impact on HOUST.

```r
anova(G3)
```

```
## Analysis of Variance Table

## Response: HOUST
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## GDP           1   76745   76745  7.0099 0.008955 **
## CPI           1    2025    2025  0.1849 0.667759
## QUARTER       3  285122   95041  8.6811 2.35e-05 ***
## POPULATION    1    3542    3542  0.3236 0.570308
## Residuals   153 1675046   10948
X1 = TukeyHSD(aov(G3),conf.level = 0.90)
plot(X1)
```



The P-value of Population Is high & Coefficient estimate is very low. So, adding Population to model G3 doesn't have any impact on Seasonality effect.

Exercise 4

```
test <- read_csv("C:/Users/PRABHATJOHL/Desktop/Final/test-default.csv")

Converting Character Variable to Factor
test$default <- factor(test$default)
test$student <- factor(test$student)

sapply(test,function(x) sum(is.na(x))) { Checking for missing values }

customer  default  student  balance   income
       0        0        0        0        0

View(head(test))
```

customer	default	student	balance	income
2	No	Yes	817.1804	12106.13
4	No	No	529.2506	35704.49
5	No	No	785.6559	38463.50
8	No	Yes	808.6675	17600.45
11	No	Yes	0.0000	21871.07
13	No	No	237.0451	28251.70

```
train <- read_csv("C:/Users/PRABHATJOHL/Desktop/Final/train-default.csv")

Converting Character Variable to Factor
train$default <- factor(train$default)
train$student <- factor(train$student)

sapply(train,function(x) sum(is.na(x))) { Checking for missing values }

customer  default  student  balance   income
       0        0        0        0        0

View(head(train))
```

customer	default	student	balance	income
1	No	No	729.5265	44361.625
3	No	No	1073.5492	31767.139
6	No	Yes	919.5885	7491.559
7	No	No	825.5133	24905.227
9	No	No	1161.0579	37468.529
10	No	No	0.0000	29275.268

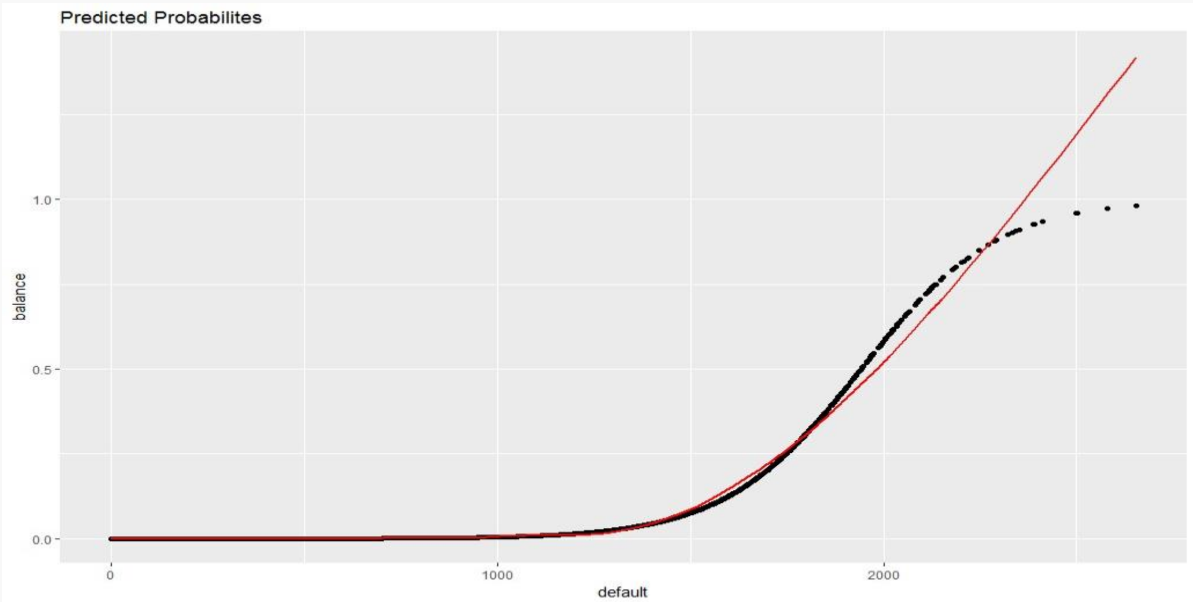
(a) Fit a logistic regression model with the default as the response and the variable balance as the predictor. Make sure that predictor variable in your model is significant?

```
A1 <- glm(default~balance,data = train, family =binomial (link='logit'),maxit
= 100)
summary(A1)

## Call:
## glm(formula = default ~ balance, family = binomial(link = "logit"),
##     data = train, maxit = 100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2905  -0.1395  -0.0528  -0.0189   3.3346
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.101e+01  4.887e-01  -22.52   <2e-16 ***
## balance      5.669e-03  2.949e-04   19.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1723.03  on 6046  degrees of freedom
## Residual deviance:  908.69  on 6045  degrees of freedom
## AIC: 912.69
##
## Number of Fisher Scoring iterations: 8


The Fisher scoring Integration are small & p-value of the predictor is low
which is good for our model. Only the AIC value is high.
```

(b) Why is your model a good/reasonable model? Check the AIC and pseudo-R2 values?

Model A1 behavior - it follows a sigmoidal curve along the xy plane.

```
p <- qplot(train$balance,fitted.values(A1)) + labs(title = "Predicted Probabi
lites", x = "balance", y = "default") + geom_smooth(color = "red", se = T)
```




```
round(pR2(A1),2)

        llh   llhNull      G2     McFadden      r2ML      r2CU
##  -454.34   -861.51   814.34       0.47       0.13      0.51
```

Here we are using McFadden Pseudo R^2 for Accessing the predictive power of the model A1. The values obtained here is 0.47 which states that the model A1 is "moderately" effective in predicting the likelihood of default on balance by customer. (it must be close to 1 for better Accuracy)

```
varImp(A1)
           Overall
## balance 19.21985

anova(A1,test = "Chisq")

## Analysis of Deviance Table
## Model: binomial, link: logit
## Response: default
## Terms added sequentially (first to last)

##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     6046    1723.03
## balance  1   814.34      6045     908.69 < 2.2e-16
```

```
 AIC for model A1 is high, as we have only 1 predictor, it might not be
reasonable to judge. As per Anova chi-square test, balance has low p-value
which depicts that its effective in providing outcome.
```

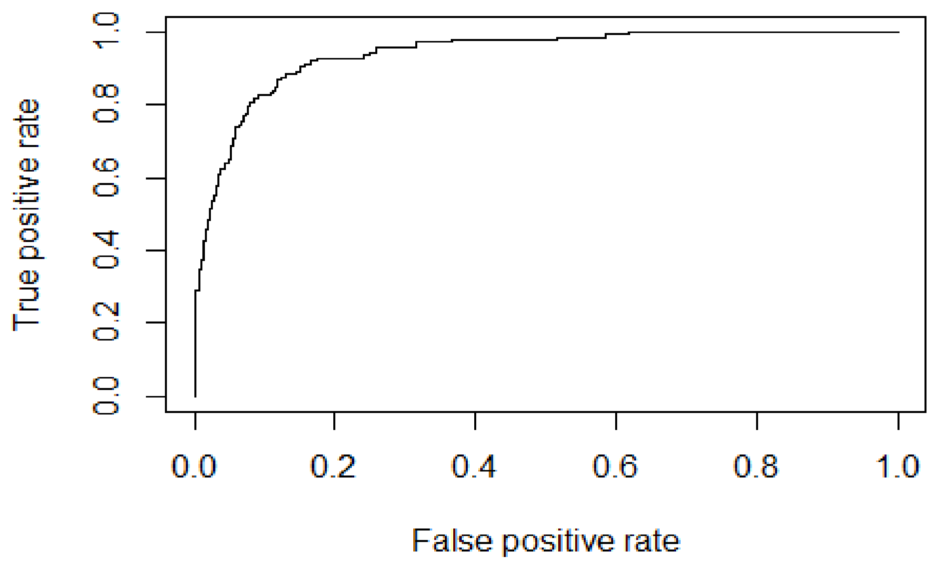
(c) Give an interpretation of the regression coefficients?

```
round(coef(A1),3)

## (Intercept)     balance
##     -11.006       0.006

Equation :- p = e^0.006[x]/1 + e^0.006[x]  { keeping Intercept Constant }

It can be seen from above equation that with 1 unit increase in balance of
customer, likelihood probability on default increases by 0.006 unit.
```

(d) Form the confusion matrix over the test data. What percentage of the time, are your predictions correct?

```
Assessing the predictive ability of the model A1

Test dataset

pred1 <- predict(A1, newdata =test,type = "response")
table(Actual=test$default,Predicted=pred1>0.5)

          Predicted
   Actual  FALSE    TRUE
     No    3803     12
     Yes     98     40

(Accuracy = ((3803+40)/3953 )*100)

97.2173 %

(Misclassification = 100 - Accuracy)

2.782697 %
```

```r
pr <- prediction(pred1,test$default)
prf <- performance(pr, measure = "tpr",x.measure = "fpr")
plot(prf)
```




```r
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
print(auc)
```

0.939932 -> is the Area Under the Curve for our Model.

(e) Now, let's add the variables income and student to the model. Fit a logistic regression model of the form "default balance + income + student", in other words, regress the variable default to all the other predictors with logistic regression?

```r
B1 <- glm(default~.-customer,data = train, family =binomial (link='logit'), maxit = 100)
summary(B1)

## Call:
## glm(formula = default ~ . - customer, family = binomial(link = "logit"),
##     data = train, maxit = 100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.4556  -0.1344  -0.0499  -0.0174   3.4155
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.091e+01  6.481e-01 -16.830  < 2e-16 ***
## studentYes  -8.095e-01  3.133e-01  -2.584  0.00978 **
## balance      5.907e-03  3.102e-04  19.040  < 2e-16 ***
## income      -5.013e-06  1.079e-05  -0.465  0.64212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1723.03  on 6046  degrees of freedom
## Residual deviance:  895.02  on 6043  degrees of freedom
## AIC: 903.02
##
## Number of Fisher Scoring iterations: 8
```

```
round(pR2(B1),2)
```

```
##      llh  llhNull       G2   McFadden     r2ML     r2CU
##  -447.51  -861.51   828.01       0.48     0.13     0.52
```

We have an optimal Value of McFadden Pseudo R^2.

```
varImp(B1)
```

```
                Overall
## studentYes  2.5835947
## balance    19.0403764
## income      0.4647343
```

Balance has the highest impact compared to other Variables.

```
anova(B1,test = "Chisq")
```

```
## Analysis of Deviance Table
## Model: binomial, link: logit
## Response: default
## Terms added sequentially (first to last)
```

```
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## student  1     8.44      6045    1714.59  0.003671 **
## balance  1   819.35      6044     895.23 < 2.2e-16 ***
## income   1     0.22      6043     895.02  0.642115
```

As per Anova Chi-square test, we can remove income variable for better prediction.

Assessing the predictive ability of the model B1

Test dataset

```
pred2 <- predict(B1, newdata =test,type = "response")
table(Actual=test$default,Predicted=pred1>0.5)
```

```
        Predicted
 Actual FALSE TRUE
   No    3803   12
   Yes     98   40
```

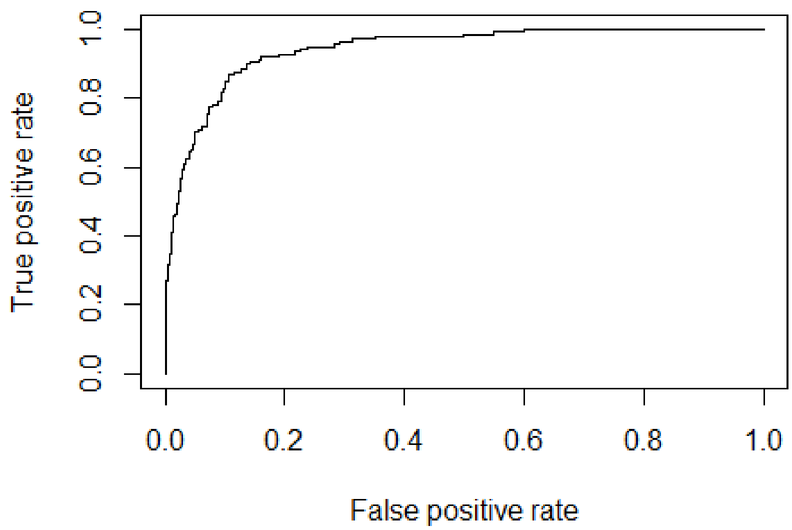
```
(Accuracy = ((3805+37)/3953 )*100)
```

97.19201 %

```
(Misclassification = 100 - Accuracy)
```

2.807994 %

```
pr <- prediction(pred2,test$default)
prf <- performance(pr, measure = "tpr",x.measure = "fpr")
plot(prf)
```




```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
print(auc)
```

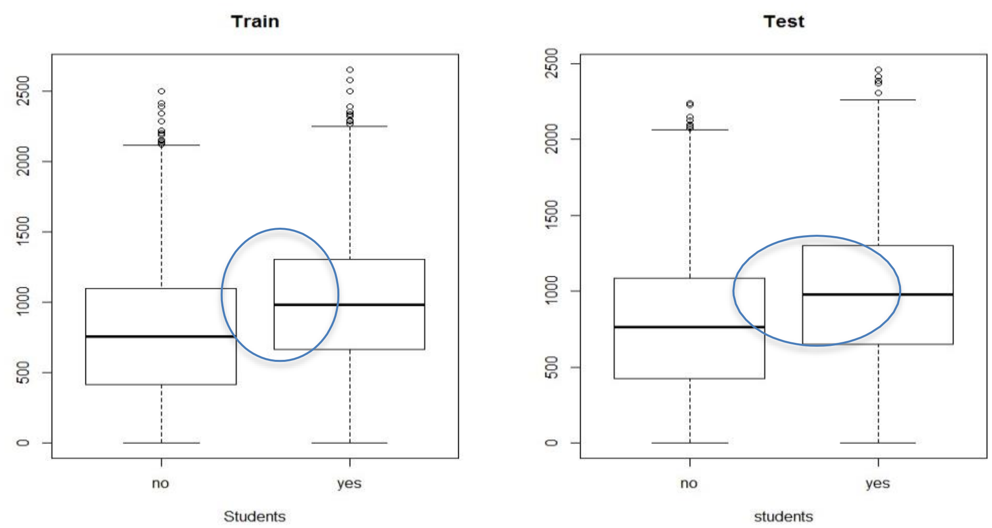
0.9419207 -> is the area under the curve

(f) In your model in question (e), what is the estimated probability of default for a student with a credit card balance of $2,000 and an income of $40,000? What is the probability of the default for a non-student with the same credit card balance and income?

```
new = data.frame(customer = c(1,2),balance = c(2000,2000),
                 income = c(40000,40000),student = c("No","Yes"))

new$student <- factor(new$student)

a<-predict(B1,new,type = "response")

table(students = new$student, Probability= round(a,2))

                      Probability
## students     0.47              0.66
        No       0                  1
        Yes      1                  0

It clear from the above table that student have a probability of 0.47 on
default in comparison to non-students who has probability of 0.66
```

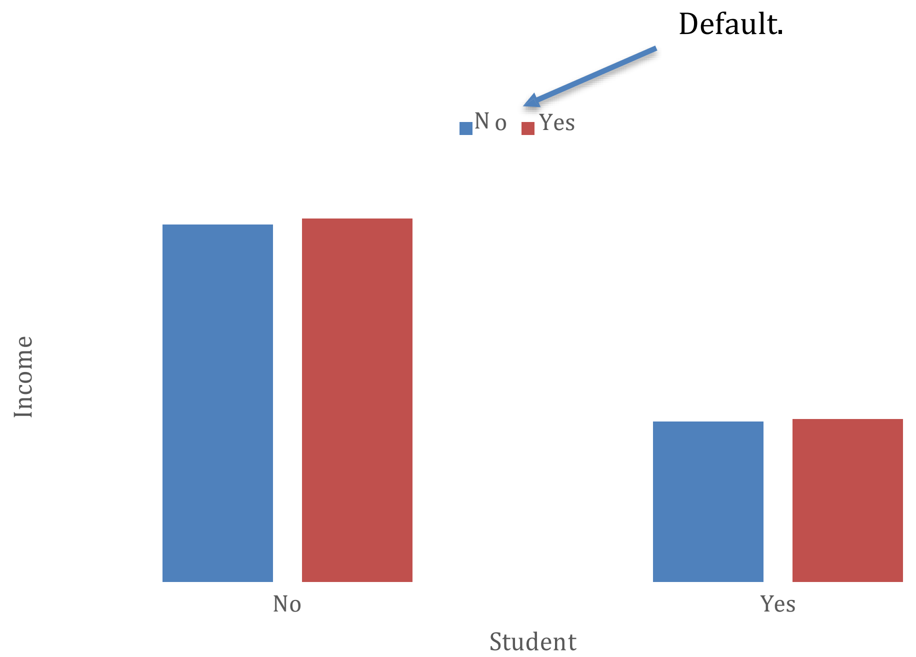
(g) Are the variables student and balance are correlated? If yes, why do you think this is the case? If no, please explain?

```
levels(train$student) <- c("no","yes")
levels(test$student) <- c("no","yes")

par(mfrow=c(1,2))
plot(train$student,train$balance, xlab="Students",main="Train")
plot(test$student,test$balance,xlab="students",main="Test")
```



From the Above Plots, it can be stated that Students & Balance are correlated because students poses high balance in their account as compared to non- students.

(h) Does the data say that it is more likely for a student to default compared to a non-student for different values of income level? Please comment. In other words, if you were the credit card company, would you prefer students as customers or non-students as customers with the same income level?



From the Above Analysis, it clear to target non-student customer as they have high income along with Default on “Yes” is high in comparison to students.