Assignment - II

Mohammad Wasiq

2022-06-06

Table of Contents

Name -: **Mohammad Wasiq**

E-mail -: **mohammadwasiq0786@gmail.com**

1   First Problem

1. **The data on features given in the table below are collected to estimate the published relative performance (PRP) of a centralised processing unit. The data is given in the CPU_Data file.**

Feature Name	Description	Unit
MYCT	Machine Cycle Time	Nanoseconds
MMIN	Minimum Main Memory	Kilobytes
MMAX	Maximum Main Memory	Kilobytes
CACH	Cache Memory	Kilobytes
CHMIN	Minimum Channels	Channels
CHMAX	Maximum Channels	Channels

Read the Data

```
library(readxl)
df <- read_excel("F:/ISI R Course/Assignments/Assignment-II/CPU_Data.xlsx")
head(df)
```

```
## # A tibble: 6 × 7
##    MYCT  MMIN  MMAX  CACH CHMIN CHMAX   PRP
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   125   256  6000   256    16   128   198
## 2    29  8000 32000    32     8    32   269
## 3    29  8000 32000    32     8    32   220
## 4    29  8000 32000    32     8    32   172
## 5    29  8000 16000    32     8    16   132
## 6    26  8000 32000    64     8    32   318

attach(df)
```

a.   Provide a descriptive summary of fields in the dataset

```
summary(df)

##       MYCT             MMIN             MMAX             CACH           CHMIN
##  Min.   :  17.0   Min.   :   64   Min.   :   64   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  50.0   1st Qu.:  768   1st Qu.: 4000   1st Qu.:  0.00   1st Qu.: 1.000
##  Median : 110.0   Median : 2000   Median : 8000   Median :  8.00   Median : 2.000
##  Mean   : 203.8   Mean   : 2868   Mean   :11796   Mean   : 25.21   Mean   : 4.699
##  3rd Qu.: 225.0   3rd Qu.: 4000   3rd Qu.:16000   3rd Qu.: 32.00   3rd Qu.: 6.000
##  Max.   :1500.0   Max.   :32000   Max.   :64000   Max.   :256.00   Max.   :52.000
##      CHMAX            PRP
##  Min.   :  0.00   Min.   :   6.0
##  1st Qu.:  5.00   1st Qu.:  27.0
##  Median :  8.00   Median :  50.0
##  Mean   : 18.27   Mean   : 105.6
##  3rd Qu.: 24.00   3rd Qu.: 113.0
##  Max.   :176.00   Max.   :1150.0
```

- **MYCT** The *mean* of MYCT is 203.822 and the *standard deviation* is 260.26 .

The *minimum* and *maximum* values are 17.00 and 1500.00 respectively.

25 % of data of MYCT is below 50.00 and *75 %* data is above 50.00 .

50 % of data of MYCT is below 110.00 and rest data is above 110.00 .

75 % of data of MYCT is below 225.00 and *25 %* data is above 225.00 .

- **MMIN**

The *mean* of MMIN is 287.98 and the *standard deviation* is 3878.74 .

The *minimum* and *maximum* values are 64.00 and 32000.00 respectively.

25 % of data of MMIN is below 768.00 and *75 %* data is above 768.00 .

50 % of data of MMIN is below 2000.00 and rest data is above 2000.00 .

75 % of data of MMIN is below 4000.00 and *25 %* data is above 4000.00 .

- **MMAX**

The *mean* of MMAX is 11796.15 and the *standard deviation* is 11726.56 .

The *minimum* and *maximum* values are 64.00 and 64000.00 respectively.

25 % of data of MMAX is below 4000.00 and *75 %* data is above 4000.00 .

50 % of data of MMAX is below 8000.00 and rest data is above 8000.00 .

75 % of data of MMAX is below 16000.00 and *25 %* data is above 16000.00 .

- **CACH**

The *mean* of MMAX is 25.20 and the *standard deviation* is 40.62 .

The *minimum* and *maximum* values are 00.00 and 256.00 respectively.

25 % of data of MMAX is below 00.00 and *75 %* data is above 00.00 .

50 % of data of MMAX is below 8.00 and rest data is above 8.00 .

75 % of data of MMAX is below 32.00 and *25 %* data is above 32.00 .

- **CHMIN**

The *mean* of CHMIN is 4.67 and the *standard deviation* is 6.81 .

The *minimum* and *maximum* values are 00.00 and 52.00 respectively.

25 % of data of CHMIN is below 1.00 and *75 %* data is above 1.00 .

50 % of data of CHMIN is below 2.00 and rest data is above 2.00 .

75 % of data of CHMIN is below 6.00 and *25 %* data is above 6.00 .

- **CHMAX**

The *mean* of CHMAX is 18.26 and the *standard deviation* is 26.00 .

The *minimum* and *maximum* values are 00.00 and 176.00 respectively.

25 % of data of CHMAX is below 5.00 and *75 %* data is above 5.00 .

50 % of data of CHMAX is below 8.00 and rest data is above 8.00 .

75 % of data of CHMAX is below 24.00 and *25 %* data is above 24.00 .

- **PRP**

The *mean* of PRP is 105.62 and the *standard deviation* is 160.83.

The *minimum* and *maximum* values are 6.00 and 1150.00 respectively.

25 % of data of PRP is below 27.00 and *75 %* data is above 27.00 .

50 % of data of PRP is below 50.00 and rest data is above 50.00 .

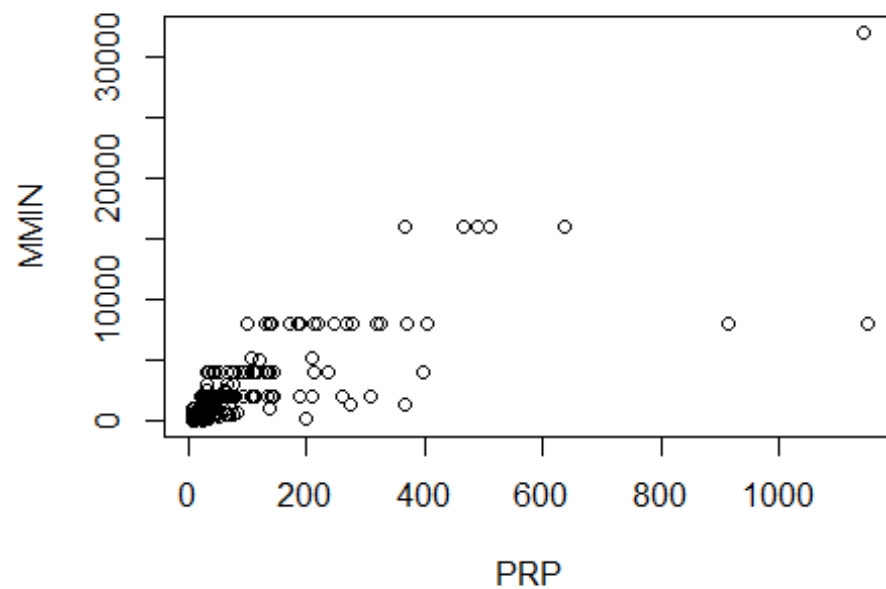*75 %* of data of PRP is below 113.00 and *25 %* data is above 113.00 .

b. Explore the relationship between the response variable PRP and explanatory features using scatter plots and correlation matrix. Give your interpretation of the relationship between response and explanatory features.
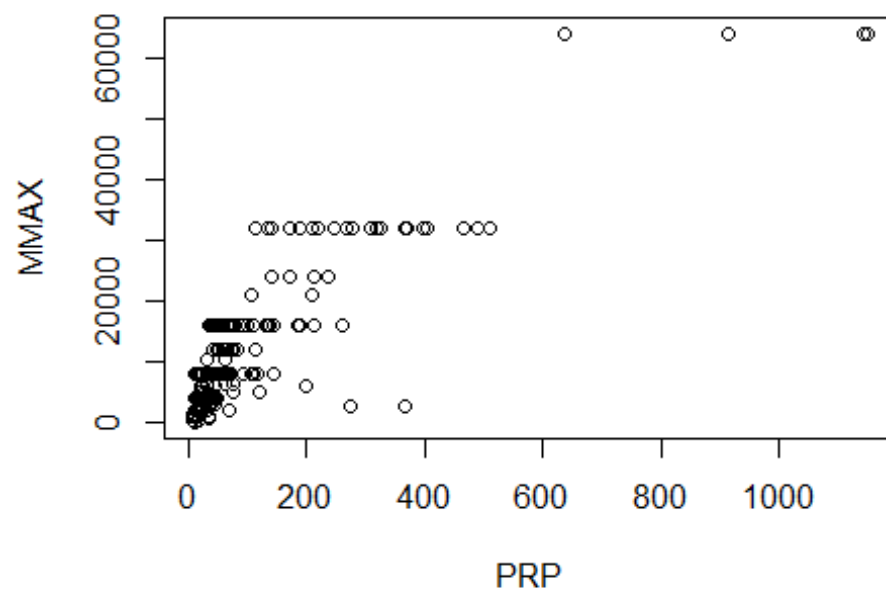
```
plot(PRP, MYCT, main = "PRP Vs MYCT")
```


PRP Vs MYCT
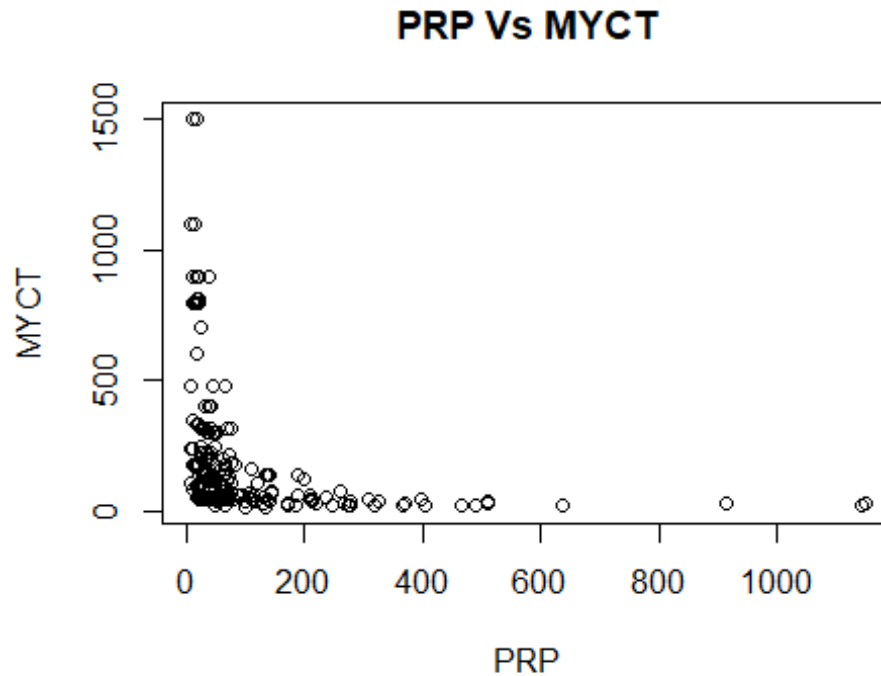
```
plot(PRP, MMIN, main = "PRP Vs MMIN")
```
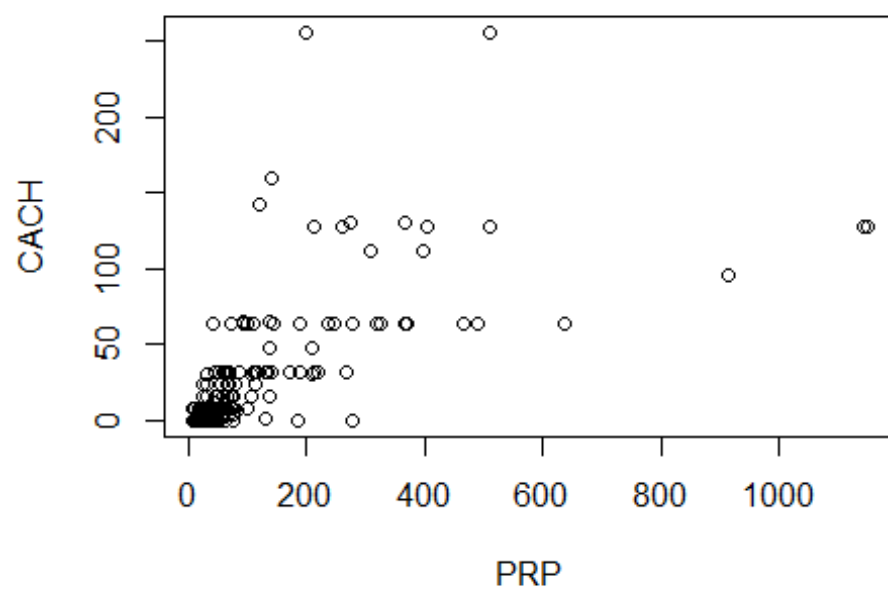
PRP Vs MMIN


```
plot(PRP, MMAX, main = "PRP Vs MMAX")
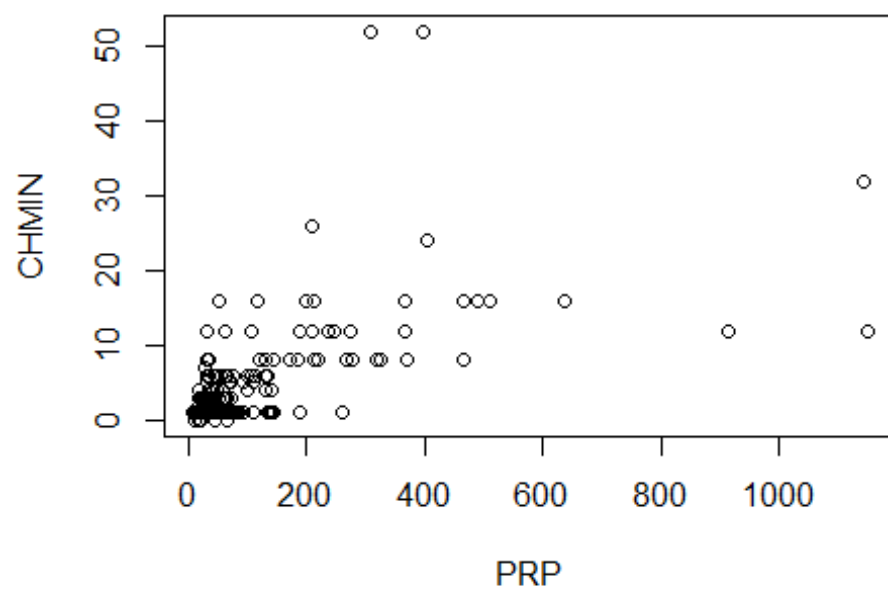```

## PRP Vs MMAX



```
plot(PRP, CACH, main = "PRP Vs CACH")
```
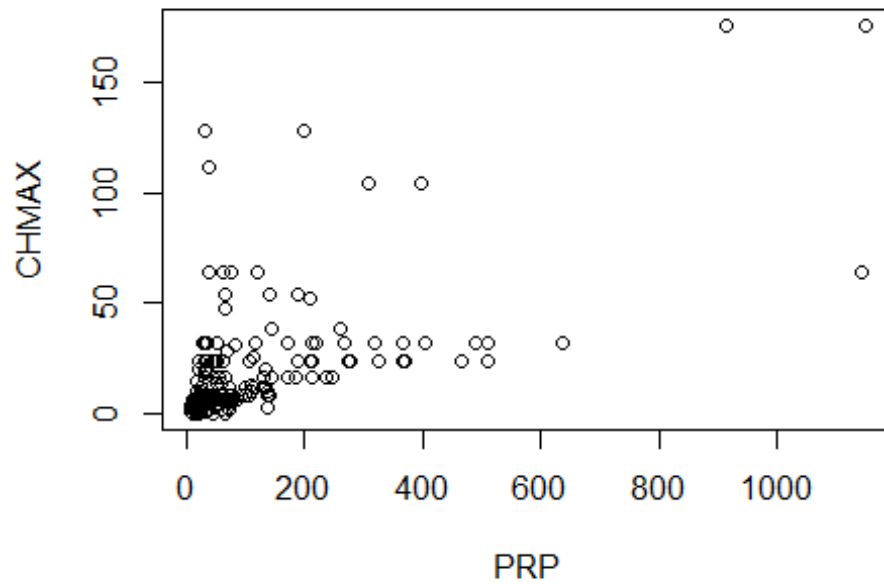
## PRP Vs CACH



```
plot(PRP, CHMIN, main = "PRP Vs MHMIN")
```
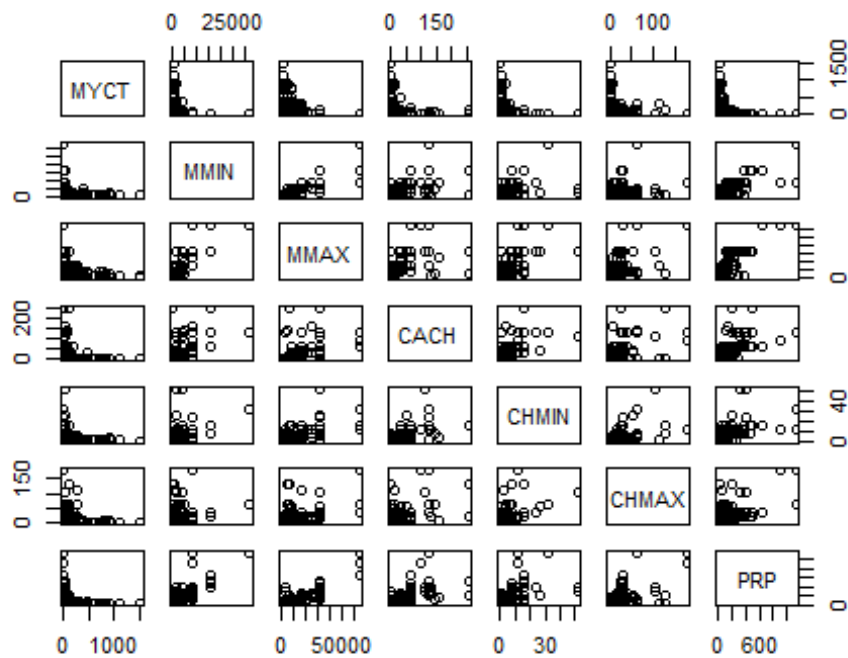
## PRP Vs MHMIN



```
plot(PRP, CHMAX, main = "PRP Vs MHMAX")
```

PRP Vs MHMAX


```
pairs(df)
```

There is a high correlation between *PRP VS MMAX, PRP VS MMIN* whereas there is a low correlation between *PRP* VS rest variables.

> c. Develop a model to predict the PRP using explanatory features. Provide the model coefficient table and interpret. Are all features have a significant impact on the response? Provide the mathematical expression of your model.

```
l_model<- lm(PRP~. , data = df)

summary(l_model)

##
## Call:
## lm(formula = PRP ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -195.82  -25.17    5.40   26.52  385.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.589e+01  8.045e+00  -6.948 5.00e-11 ***
## MYCT         4.885e-02  1.752e-02   2.789   0.0058 **
## MMIN         1.529e-02  1.827e-03   8.371 9.42e-15 ***
## MMAX         5.571e-03  6.418e-04   8.681 1.32e-15 ***
## CACH         6.414e-01  1.396e-01   4.596 7.59e-06 ***
## CHMIN       -2.704e-01  8.557e-01  -0.316   0.7524
## CHMAX        1.482e+00  2.200e-01   6.737 1.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.99 on 202 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8609
## F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

MOdel :

$$prp = \beta_0 + \beta_1(myct) + \beta_2(mmin) + \beta_3(mmax) + \beta_4(cach) + \beta_5(chmin) + \beta_6(chmax) + \epsilon$$

Fitted Model :

$\widehat{prp}$
$$= -55.8939 + 0.0489(\boldsymbol{myct}) + 0.0153(\boldsymbol{mmin}) + 0.0056(\boldsymbol{mmax}) + 0.6414(\boldsymbol{cach}) + (-0.2704)(\boldsymbol{chmin}) + 1.4825(\boldsymbol{chmax})$$

> d. Provide F statistic value and corresponding p_value? Give your comments on model significance?

$H_0\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6$ VS

H_1: At least one of $b_i's$ is not equal ; $i = 1, \cdots, 6$

The *F-statistic* of Model is 215.5 with 6 and 202 degree of freedom and the corresponding p-value is $2.2e^{-16}$ which is less than 0.05, so we reject our null hypothesis H_0 that implies that there is enough evidence to claim that *At least one of b_i's is not equal ; $i = 1, \cdots, 6$*

    e.   Provide R2 and adjusted R2 values. Comment on model accuracy.

The value of *R-squared* is 0.8649, and the value of *Adjusted R-squared* is 0.8609 this implies that 86% variation in *PRP* is explained by the model.
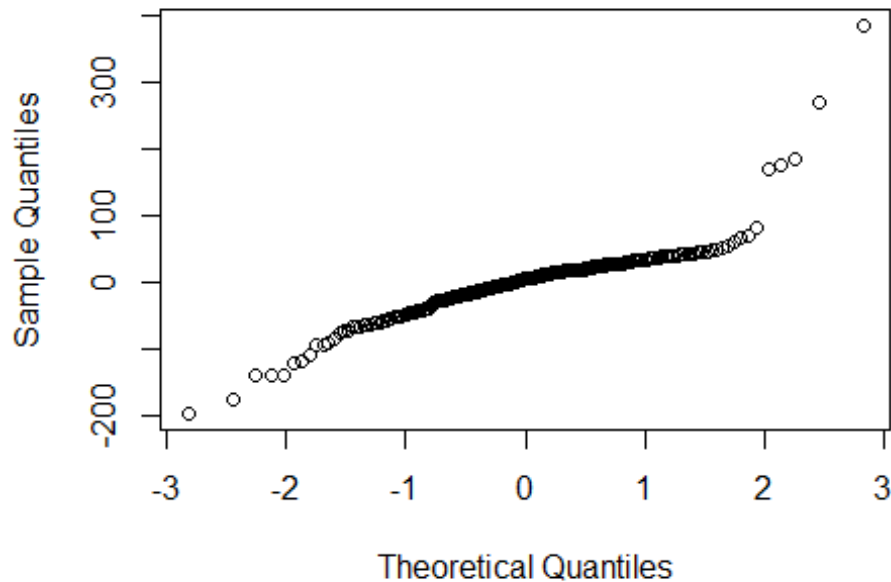
    f.   Comment on model adequacy based on residual analysis and plots. Provide normal probability plot of residuals and normality test result (test statistic and p_value) and comment on the normality of residuals.

```
plot(PRP ~ l_model$residuals)
```
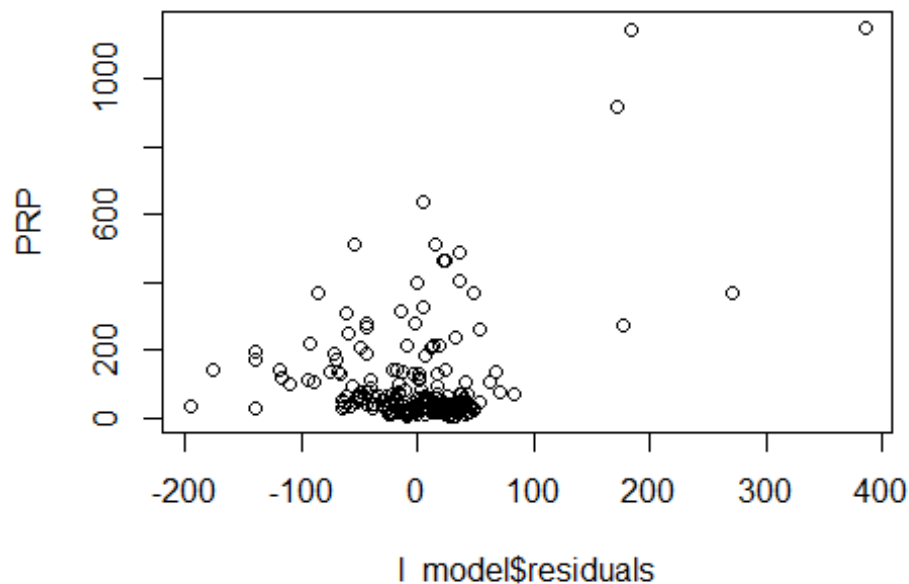


```
qqnorm(l_model$residuals)
```

Normal Q-Q Plot


```
shapiro.test(l_model$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  l_model$residuals
## W = 0.83127, p-value = 2.597e-14
```
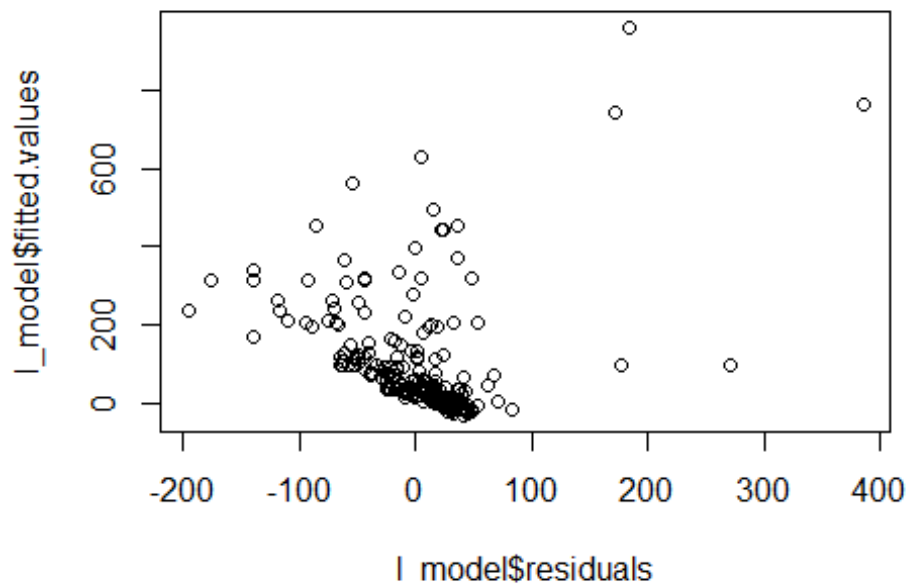
The p-value is $2.597e^{-14}$ which is lesser than 0.05 , now we conclude that *residuals* are not normal or not follows normality.

   g.   Provide Residuals versus Predicted values plot and given your interpretation.

```
plot(l_model$residuals, l_model$fitted.values)
```

We can easily see that there is no correlation between residuals and predicted values.

h. Perform k fold (k = 10) cross-validation. Provide the mean square error and root mean square error obtained for original data and during cross-validation. Give your comments on model generalizability.

```r
# package to compute
# cross - validation methods
library(caret)

# setting seed to generate a
# reproducible random sampling
set.seed(123)

# defining training control
# as cross-validation and
# value of K equal to 10
train_control <- trainControl(method = "cv",
                              number = 10)

# training the model by assigning sales column
# as target variable and rest other column
# as independent variable
model <- train(PRP ~., data = df,
               method = "lm",
               trControl = train_control)
```

```
print(model)

## Linear Regression
##
## 209 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 188, 188, 189, 188, 190, 187, ...
## Resampling results:
##
##   RMSE       Rsquared    MAE
##   72.77571   0.8063966   44.81945
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

- The *Root Mean Square Error* is 64.76 which is the square root of Mean Squared error. It measures the standard deviation of residuals.

- The value of *Mean Absolute Error* is 41.59 which represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

- The value of *Rsquared* is 0.792 which represents the proportion of the variance in the dependent variable which is explained by the linear regression model.

2   Second Problem

**2. The data has collected on health profile parameters of people shown symptoms of heart disease and their diagnostic results are given in the Heart_Disease_Data file. The list of health profile features on which data is collected is given in the table below.

SL No	Feature Name	Description
1	Age	Age
2	Sex	Sex
3	CP	Chest pain type
4	RestBP	Resting blood pressure
5	Cholesterol	Serum cholesterol in mg/dl
6	FBP	Fasting blood sugar > 120 mg/dl
7	RestECG	Resting electrocardiographic results
8	Max_HR	Maximum heart rate achieved
9	ExAngina	Exercise-induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	CA	Number of major vessels (0-3) colored by flourosopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect

```
library(readxl)
df <- read_excel("Heart_Disease_Data.xlsx")
dim(df)

## [1] 303  14

attach(df)

head(df)

## # A tibble: 6 × 14
##     Age   Sex    CP RestBP Cholesteral   FBP RestECG Max_HR ExAngina Oldpeak Slope    CA
##   <dbl> <dbl> <dbl>  <dbl>       <dbl> <dbl>   <dbl>  <dbl>    <dbl>   <dbl> <dbl> <dbl>
## 1    63     1     3    145         233     1       0    150        0     2.3     0     0
## 2    37     1     2    130         250     0       1    187        0     3.5     0     0
## 3    41     0     1    130         204     0       0    172        0     1.4     2     0
## 4    56     1     1    120         236     0       1    178        0     0.8     2     0
## 5    57     0     0    120         354     0       1    163        1     0.6     2     0
## 6    57     1     0    140         192     0       1    148        0     0.4     1     0
## # … with 2 more variables: Thal <dbl>, Result <dbl>
```

a. Provide a descriptive summary of the features

```
summary(df)
```

```
##       Age             Sex              CP             RestBP         Cholesteral
## Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0
## 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0
## Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0
## Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3
## 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5
## Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0
##      FBP            RestECG          Max_HR          ExAngina         Oldpeak
## Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00
## 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00
## Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000   Median :0.80
## Mean   :0.1485   Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04
## 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60
## Max.   :1.0000   Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20
##      Slope            CA              Thal            Result
## Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
## 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000
## Median :1.000   Median :0.0000   Median :2.000   Median :1.0000
## Mean   :1.399   Mean   :0.7294   Mean   :2.314   Mean   :0.5446
## 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000
## Max.   :2.000   Max.   :4.0000   Max.   :3.000   Max.   :1.0000
```

- **AGE**

The *mean* of AGE is 54.36 and the *standard deviation* is 9.08.

The *minimum* and *maximum* values are 29.00 and 77.00 respectively.

25 % of data of AGE is below 47.50 and *75 %* data is above 47.00.

50 % of data of AGE is below 55.00 and rest data is above 55.00.

75 % of data of AGE is below 61.00 and *25 %* data is above 61.00.

- **RestBP**

The *mean* of RestBP is 131.62 and the *standard deviation* is 17.53.

The *minimum* and *maximum* values are 94.00 and 200.00 respectively.

25 % of data of RestBP is below 120.00 and *75 %* data is above 120.00. *50 %* of data of RestBP is below 130.00 and rest data is above 130.00. *75 %* of data of RestBP is below 140.00 and *25 %* data is above 140.00.

- **Cholesteral**

The *mean* of Cholesteral is 246.26 and the *standard deviation* is 51.83.

The *minimum* and *maximum* values are 126.00 and 564.00 respectively.

25 % of data of Cholesteral is below 211.00 and *75 %* data is above 211.00.

50 % of data of Cholesteral is below 240.00 and rest data is above 240.00.

75 % of data of Cholesteral is below 274.00 and *25 %* data is above 274.00.

- **Max_HR**

The *mean* of MaxHR is 149.64 and the *standard deviation* is 22.9.

The *minimum* and *maximum* values are 71.00 and 202.00 respectively.

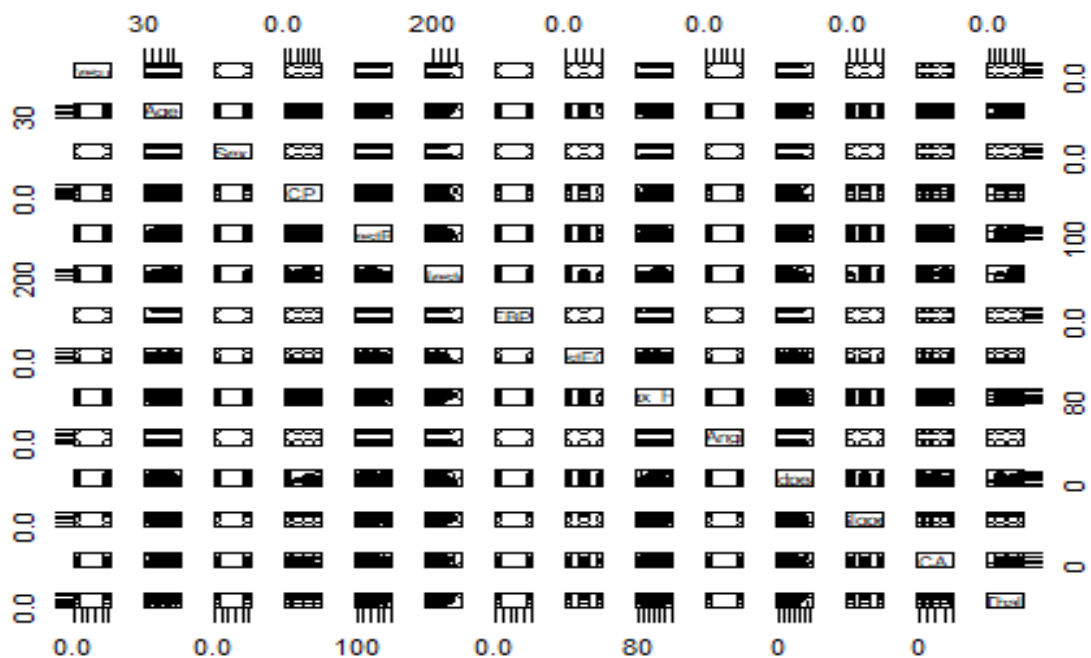*25 %* of data of MaxHR is below 133.50 and *75 %* data is above 133.50.

50 % of data of MaxHR is below 153.00 and rest data is above 153.00.

75 % of data of MaxHR is below 166.00 and *25 %* data is above 166.00.

and so on…

b. Explore the relationship between response "Result" and the features using graphically.

```
pairs(Result ~., data = df)
```


c. Develop a logistic regression model to classify whether a patient is having heart disease or not using the feature values. Give the model coefficient table and the mathematical expression of the model.

```
log_m<- glm(Result~., data = df, family = "binomial")

summary(log_m)
```

```
## 
## Call:
## glm(formula = Result ~ ., family = "binomial", data = df)
## 
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max  
## -2.5849  -0.3872   0.1551   0.5863   2.6249  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  3.450472   2.571479   1.342 0.179653    
## Age         -0.004908   0.023175  -0.212 0.832266    
## Sex         -1.758181   0.468774  -3.751 0.000176 ***
## CP           0.859851   0.185397   4.638 3.52e-06 ***
## RestBP      -0.019477   0.010339  -1.884 0.059582 .  
## Cholesteral -0.004630   0.003782  -1.224 0.220873    
## FBP          0.034888   0.529465   0.066 0.947464    
## RestECG      0.466282   0.348269   1.339 0.180618    
## Max_HR       0.023211   0.010460   2.219 0.026485 *  
## ExAngina    -0.979981   0.409784  -2.391 0.016782 *  
## Oldpeak     -0.540274   0.213849  -2.526 0.011523 *  
## Slope        0.579288   0.349807   1.656 0.097717 .  
## CA          -0.773349   0.190885  -4.051 5.09e-05 ***
## Thal        -0.900432   0.290098  -3.104 0.001910 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 211.44  on 289  degrees of freedom
## AIC: 239.44
## 
## Number of Fisher Scoring iterations: 6
```

$$Result$$
$$= \frac{e^{\beta_0+\beta_1\times Age+\beta_2\times Sex+\beta_3\times CP+\beta_4\times RestBP+\beta_5\times Cholesteral+\beta_6\times FBP+\beta_7\times RestECG+\beta_8\times Max_H R+\beta_9\times ExAngina+\beta_{10}\times Oldpeak+\beta_{11}\times Slope+\beta_{12}\times CA+\beta_{13}\times Thal}}{1+e^{\beta_0+\beta_1\times Age+\beta_2\times Sex+\beta_3\times CP+\beta_4\times RestBP+\beta_5\times Cholesteral+\beta_6\times FBP+\beta_7\times RestECG+\beta_8\times Max_H R+\beta_9\times ExAngina+\beta_{10}\times Oldpeak+\beta_{11}\times Slope+\beta_{12}\times CA+\beta_{13}\times Thal}}$$

$$Result$$
$$= \frac{e^{3.45+(-0.005)\times Age+(-1.76)\times Sex+0.86\times CP+(-0.019)\times RestBP+(-0.004)\times Cholesteral+0.034\times FBP+0.466\times RestECG+0.023\times Max_H R+(-98)\times ExAngina+(-0.54)\times Oldpeak+0.58\times Slope+(-0.77)\times CA+(-0.90)\times Thal}}{1+e^{3.45+(-0.005)\times Age+(-1.76)\times Sex+0.86\times CP+(-0.019)\times RestBP+(-0.004)\times Cholesteral+0.034\times FBP+0.466\times RestECG+0.023\times Max_H R+(-98)\times ExAngina+(-0.54)\times Oldpeak+0.58\times Slope+(-0.77)\times CA+(-0.90)\times Thal}}$$

d. Compare the model with a null model and provide corresponding p_value. Is the model significant?

Only **Sex, CP, Max_HR, ExAngina, Oldpeak, CA, Thal** are significant and rest variables are not-significant because their *p-values* are lesser than 0.05.

   e.   Compute and give actual versus predicted matrix. Compute accuracy & misclassification %. Comment on model accuracy.

```r
#calculate probability of default for each individual in test dataset
predicted <- predict(log_m, type="response")
ap<- data.frame(Result, predicted)
head(ap)

##   Result predicted
## 1      1 0.8073874
## 2      1 0.6583087
## 3      1 0.9618069
## 4      1 0.9145625
## 5      1 0.8172200
## 6      1 0.7615189

library(InformationValue)

#convert defaults from "Yes" and "No" to 1's and 0's
df$Result <- ifelse(df$Result == "Yes", 1, 0)

#find optimal cutoff probability to use to maximize accuracy
optimal <- optimalCutoff(df$Result, predicted)[1]
optimal

## [1] 0.9979853

library(pscl)
pscl::pR2(log_m)["McFadden"]

## fitting null model for pseudo-r2

##   McFadden
## 0.4937339
```

A value of 0.494 is quite high for McFadden's R^2, which indicates that our model fits the data very well and has high predictive power.

  f.   Check model generalizability using k -fold (k = 10) cross-validation. Provide accuracy & misclassification % obtained for data used for developing the model and during cross-validation. Is there a deterioration in model performance? Comment on model generalizability.

```r
library(boot)

set.seed(123)
cv = cv.glm(df, log_m, K = 10)

cv$delta

## [1] 0.5445545 0.1086771
```

The first component of delta is 0.544 which is the average mean-squared error that we obtain from doing K-fold CV.

The second component of delta is 0.108 which is the average mean-squared error that we obtain from doing K-fold CV, but with a bias correction.

    g.   Compute sensitivity, specificity, precision & f-measure. Comment on the aforementioned performance measures. Is the model equally good at predicting having heart disease (positive) and not having heart disease (negative) cases correctly?

```
# Confusion Matrix
cm<- confusionMatrix(Result, predicted)
cm

##     0   1
## 0 106  13
## 1  32 152
```

Calculate Sensitivity

```
recall<- sensitivity(Result, predicted)
recall

## [1] 0.9212121
```

Sensitivity of a classifier is the ratio between how much were correctly identified as positive to how much were actually positive.

It indicate that 92% are correctly identified as positive to how much were actually positive.

Calculate Precision

```
tp<- cm[1,1]
fp<- cm[1,2]
precision_ <- tp/(tp+fp)
precision_

## [1] 0.8907563
```

How much were correctly classified as positive out of all positives.

This indicate that 89% are correctly classified as positive out of all positives.

Calculate Specificity

```
specificity(Result, predicted)

## [1] 0.7681159
```

Specificity of a classifier is the ratio between how much were correctly classified as negative to how much was actually negative.

This indicate that 76% are correctly classified as negative to how much was actually negative.

Calculate Total Misclassification Error Rate

```
misClassError(Result, predicted, threshold=optimal)

## [1] 0.5413
```

F-measure

```
f_measure<- (2*(recall*precision_)) / (recall + precision_)
f_measure

## [1] 0.9057283
```

f-measure is considered a better indicator of the classifier's performance than the regular accuracy measure.

This indicates that 90% are better indicator of the classifier's performance than the regular accuracy measure.

ROC curve

```
#plot the ROC curve
plotROC(Result, predicted)
```

ROC Curve

AUROC: 0.9234