Stamatics_MLR

Group – 2
Prithwijit Ghosh
Amit Meena

Data - set : Student alcohol consumption

The data were obtained in a survey of students' math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. We will do EDA on this data set by using multiple regression analysis and analyze the student grades.

data-set description

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1.**school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) 2.**sex** - student's sex (binary: 'F' - female or 'M' - male) 3.**age** - student's age (numeric: from 15 to 22) 4.**address** - student's home address type (binary: 'U' - urban or 'R' - rural) 5.**famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) 6.**Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) 7.**Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) 8.**Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) 9.**Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') 10.**Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') 11.**reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') 12.**guardian** - student's guardian (nominal: 'mother', 'father' or 'other') 13.**traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) 14.**studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) 15.**failures** - number of past class failures (numeric: n if 1<=n<3, else 4) 16.**schoolsup** - extra educational support (binary: yes or no) 17.**famsup** - family educational support (binary: yes or no) 18.**paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) 19.**activities** - extra-curricular activities (binary: yes or no) 20.**nursery** - attended nursery school (binary: yes or no) 21.**higher** - wants to take higher education (binary: yes or no) 22.**internet** - Internet access at home (binary: yes or no) 23.**romantic** - with a romantic relationship (binary: yes or no) 24.**famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent) 25.**freetime -** free time after

school (numeric: from 1 - very low to 5 - very high) 26.**goout** - going out with friends (numeric: from 1 - very low to 5 - very high) 27.**Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) 28.**Walc -** weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) 29.**health** - current health status (numeric: from 1 - very bad to 5 - very good) 30.**absences** - number of school absences (numeric: from 0 to 93) These grades are related with the course subject, Math or Portuguese:

31.**G1** - first period grade (numeric: from 0 to 20) 32.**G2** - second period grade (numeric: from 0 to 20) 33.**G3** - final grade (numeric: from 0 to 20, output target)

Finally the link of the data - set. https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

```
getwd()
```

```
## [1] "A:/Regression Project/New Project"
```

```
library(readxl)
D = read.csv("A:/Regression Project/New Project/student-mat.csv")
```

Data - set visualization via the **"ColorDF"** package

```
library(dplyr)
```

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

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

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

```
library(colorDF)
```

```
## colorDF: for best results, use terminals which support 256 colors.
```

```
D %>% colorDF(theme = "bw")
```

Now we will see the summary our data-set.

```
summary(D)
```

```
##     school              sex                 age         address
##  Length:395         Length:395         Min.   :15.0   Length:395
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :17.0   Mode  :character
##                                        Mean   :16.7
##                                        3rd Qu.:18.0
##                                        Max.   :22.0
##    famsize            Pstatus               Medu            Fedu
```

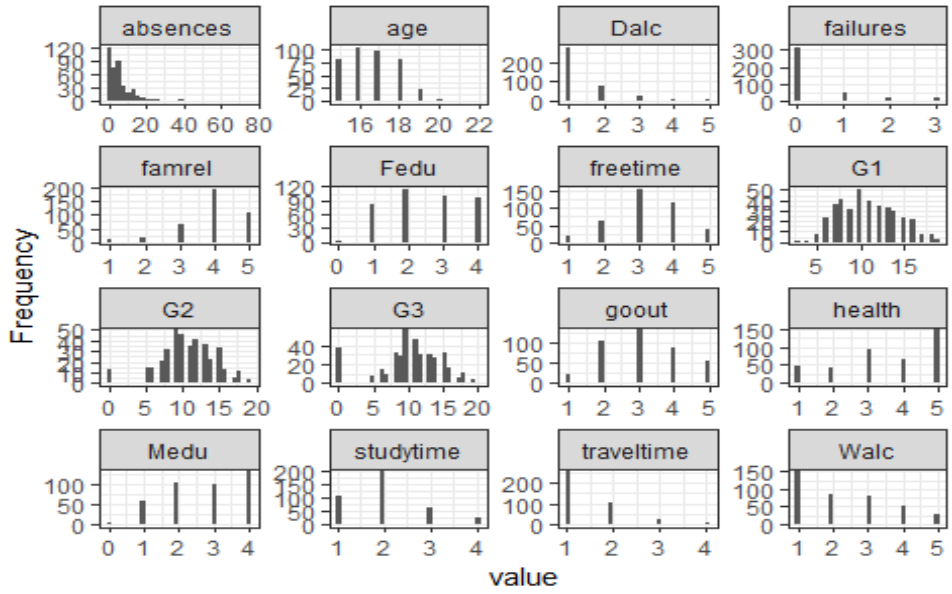
```
##  Length:395         Length:395         Min.   :0.000   Min.   :0.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :3.000   Median :2.000
##                                        Mean   :2.749   Mean   :2.522
##                                        3rd Qu.:4.000   3rd Qu.:3.000
##                                        Max.   :4.000   Max.   :4.000
##      Mjob              Fjob              reason            guardian
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    traveltime      studytime        failures       schoolsup
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :2.000   Median :0.0000   Mode  :character
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##     famsup              paid            activities          nursery
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     higher            internet           romantic            famrel
##  Length:395         Length:395         Length:395         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.944
##                                                           3rd Qu.:5.000
##                                                           Max.   :5.000
##     freetime         goout           Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health         absences            G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
```

```
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

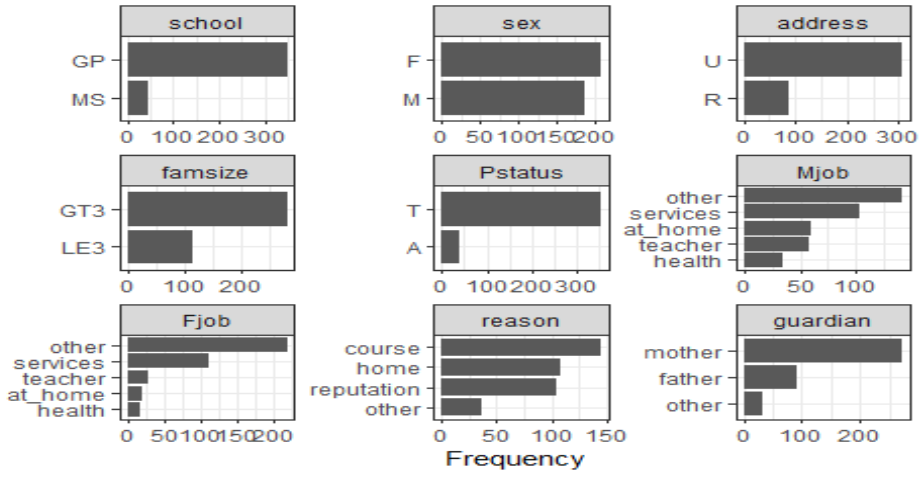
Exploratory Data Analysis:

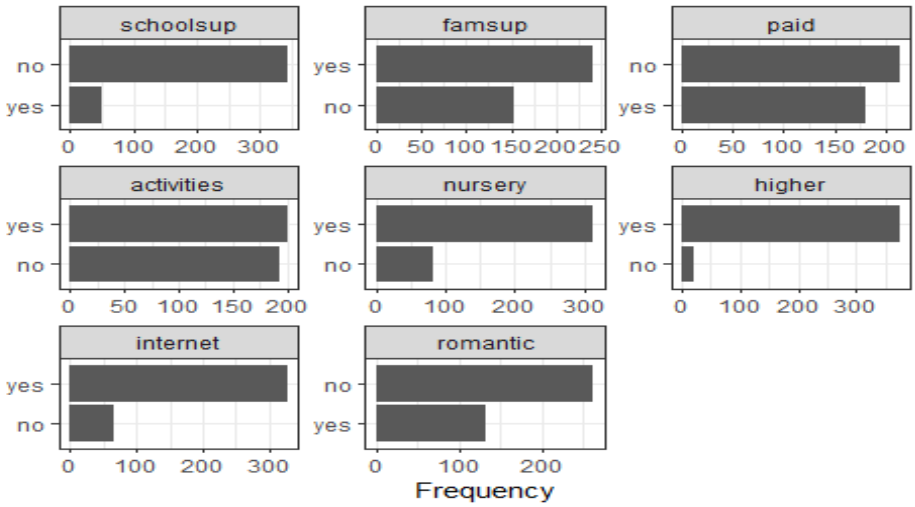
Here we use the "DataExplorer" package for the graphical visuallization of the different variable in our data-set.

```
library(DataExplorer)
library(ggplot2)
plot_histogram(D,ggtheme = theme_bw())
```




```
plot_bar(D,ggtheme = theme_bw())
```





Conversion the columns of binary category into dummy variable taking the numerical variable “0” and “1”.

```
Binary = function(y,x)
{
  z=NULL
  for ( i in 1:length(x))
  {
    if (x[i]==y[1])
      z[i]=1
    else
      z[i]=0
  }
  return(z)
}
```

Encoding the categorical columns with more than two category into a group of dummy variable where the number of dummy variables are exactly equal to the number of categories - 1.

```
Encoding = function(x,y)
{
  D1 = data.frame()
  for (i in 1:length(y))
  {
    for (j in 1:(length(x)-1))
    {
      if (y[i]==x[j])
        D1[i,j]=1
      else
        D1[i,j]=0
      if (y[i]==x[length(x)])
        D1[i,j]=0
    }
```

```
  }
  return(D1)
}
```

Converting the categorical columns into dummy varibles.

```
school = Binary(unique(D$school),D$school)
sex = Binary(unique(D$sex),D$sex)
address = Binary(unique(D$address),D$address)
famsize = Binary(unique(D$famsize),D$famsize)
Pstatus = Binary(unique(D$Pstatus),D$Pstatus)

Mjob = Encoding(unique(D$Mjob),D$Mjob)
colnames(Mjob) <- c("Mjob1","Mjob2","Mjob3","Mjob4")
Fjob = Encoding(unique(D$Fjob),D$Fjob)
colnames(Fjob) <- c("Fjob1","Fjob2","Fjob3","Fjob4")
reason = Encoding(unique(D$reason),D$reason)
colnames(reason) <- c("reason1","reason2","reason3")
guardian = Encoding(unique(D$guardian),D$guardian)
colnames(guardian) <- c("guardian1","guardian2")
schoolsup = Binary(unique(D$schoolsup),D$schoolsup)
famsup = Binary(unique(D$famsup),D$famsup)
paid = Binary(unique(D$paid),D$paid)
activities = Binary(unique(D$activities),D$activities)
nursery = Binary(unique(D$nursery),D$nursery)
higher = Binary(unique(D$higher),D$higher)
internet = Binary(unique(D$internet),D$internet)
romantic = Binary(unique(D$romantic),D$romantic)
D$school = school
D$sex = sex
D$address =address
D$famsize = famsize
D$Pstatus = Pstatus
D$schoolsup = schoolsup
D$famsup = famsup
D$paid = paid
D$activities =activities
D$nursery = nursery
D$higher = higher
D$internet = internet
D$romantic = romantic
```

Extracting the response variable.

```
Raw_Y = D$G3
Raw_Y
```

```
##   [1]  6  6 10 15 10 15 11  6 19 15  9 12 14 11 16 14 14 10  5 10 15 15 16
12  8
##  [26]  8 11 15 11 11 12 17 16 12 15  6 18 15 11 13 11 12 18 11  9  6 11 20
14  7
```

```
##  [51] 13 13 10 11 13 10 15 15  9 16 11 11  9  9 10 15 12  6  8 16 15 10  5
14 11
##  [76] 10 10 11 10  5 12 11  6 15 10  8  6 14 10  7  8 18  6 10 14 10 15 10
14  8
## [101]  5 17 14  6 18 11  8 18 13 16 19 10 13 19  9 16 14 13  8 13 15 15 13
13  8
## [126] 12 11  9  0 18  0  0 12 11  0  0  0  0 12 15  0  9 11 13  0 11  0 11
0 10
## [151]  0 14 10  0 12  8 13 10 15 12  0  7  0 10  7 12 10 16  0 14  0 16 10
0  9
## [176]  9 11  6  9 11  8 12 17  8 12 11 11 15  9 10 13  9  8 10 14 15 16 10
18 10
## [201] 16 10 10  6 11  9  7 13 10  7  8 13 14  8 10 15  4  8  8 10  6  0 17
13 14
## [226]  7 15 12  9 12 14 11  9 13  6 10 13 12 11  0 12 12  0 12  0 18 13  8
5 15
## [251]  8 10  8  8 12  8 13 11 14  0 18  8 12  9  0 17 10 11 10  0  9 14 11
14 10
## [276] 12  9  9  8 10  8 10 12 10 11 11 19 12 14 15 11 15 13 18 14 11  0  8
14 16
## [301] 11 10 14 18 13 12 18  8 12 10  0 13 11 11 13 11  0  9 10 11 13  9 11
15 15
## [326] 11 16 10  9 14  8 14  0  0  0 15 13  0 17 10 11  0 15  0 10 14 16  9
15 13
## [351]  8 13  8  8 11  9 13 11 10 16 13 12 10 15 12 10 13  0 10 11  9 12 11
5 19
## [376] 10 15 10 15 10 14  7 10  0  5 10  6  0  8  0  9 16  7 10  9
```

Making the design matrix after converting the categorical columns into binary ones.

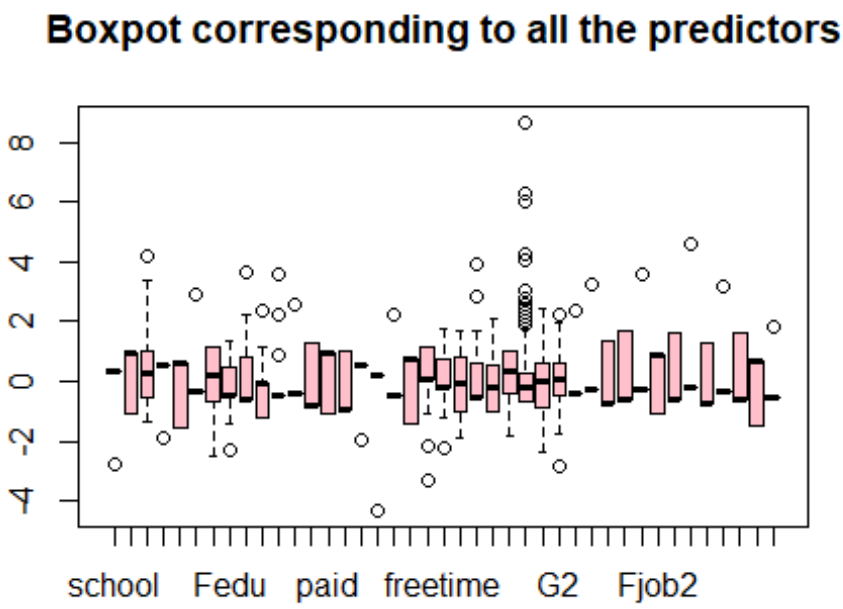
```
D1=cbind(D[-ncol(D)],Mjob,Fjob,reason,guardian)
D1 = D1[-c(9:12)]
str(D1)
```

Standardizing the response variable and all the regressor columns.

```
Y = scale(Raw_Y,center = TRUE,scale = TRUE)#Response variable
str(D1)

dim(D1)

## [1] 395  41

Raw_data = D1

D1 = scale(as.matrix(D1),center = TRUE,scale = TRUE)
```

Visualizing different columns via boxplot.

```
par(mfrow = c(1,1))
boxplot(D1,col = "pink",main = paste("Boxpot corresponding to all the predict
ors"))
```

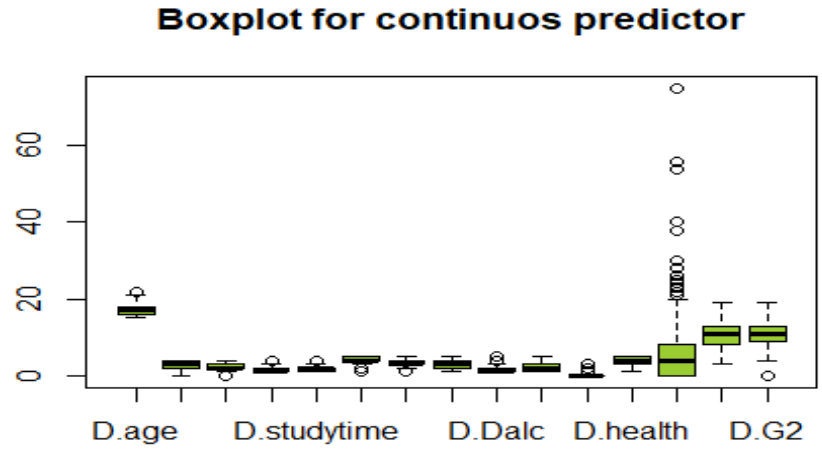


Extracting the continuous columns.

```
Data_cont = data.frame(D$age,D$Medu,D$Fedu,D$traveltime,D$studytime,D$famrel,
D$freetime,D$goout,D$Dalc,D$Walc,D$failures,D$health,D$absences,D$G1,D$G2)
Data_cont
```

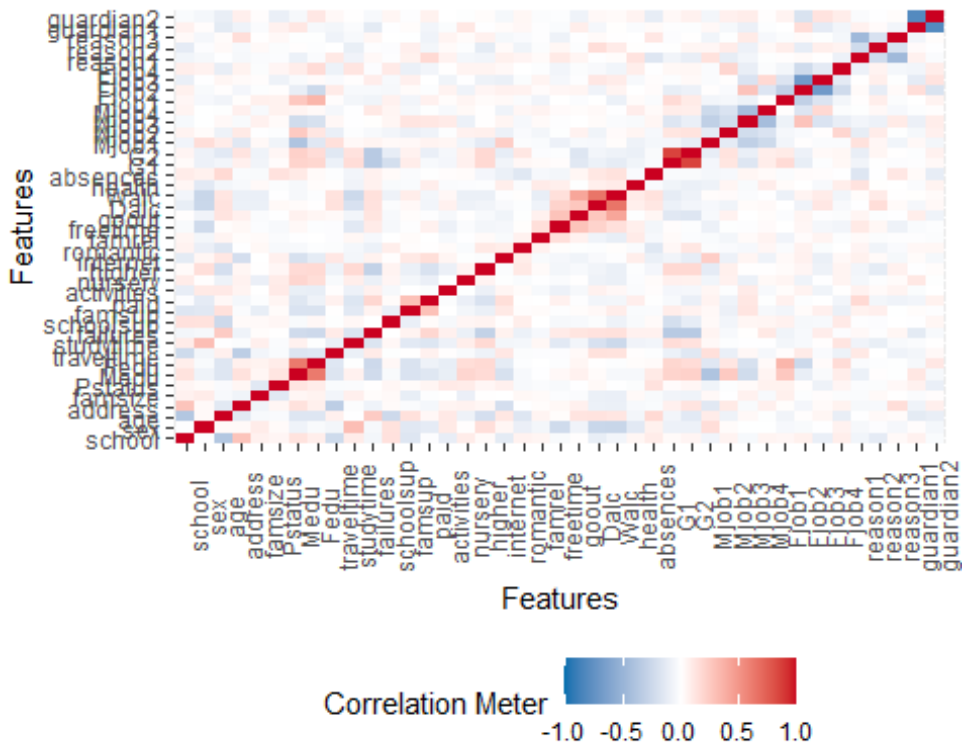
Plotting the box-plot for the continuous variable.

```
boxplot(Data_cont,col = "yellowgreen",main = paste("Boxplot for continuos pre
dictor"))
```



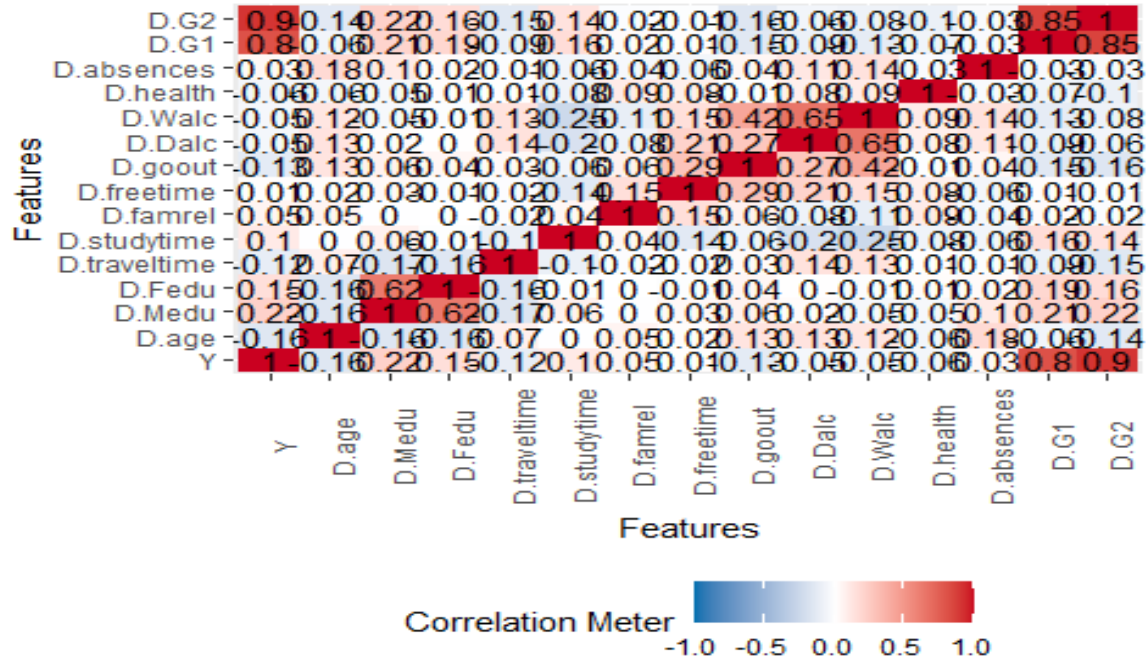
Checking the interrelation among the variables via the correlaion heatmap.

```
plot_correlation(D1)
```



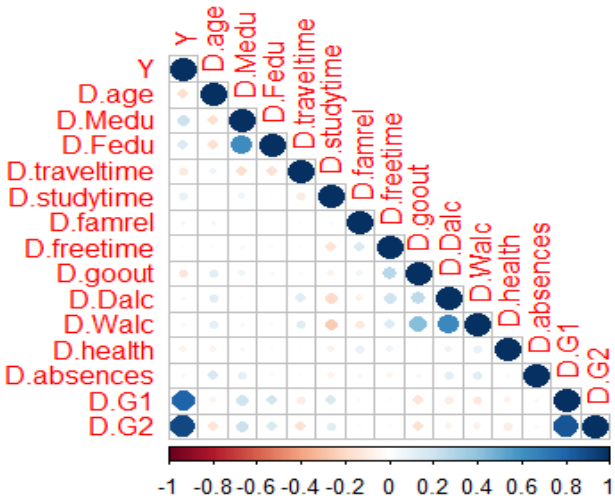
Correlation heatmap among the continuous regressors and the response.

```
plot_correlation(data.frame(Y,Data_cont[,-11]))
```



Most effectively obseving the correlation heatmap.

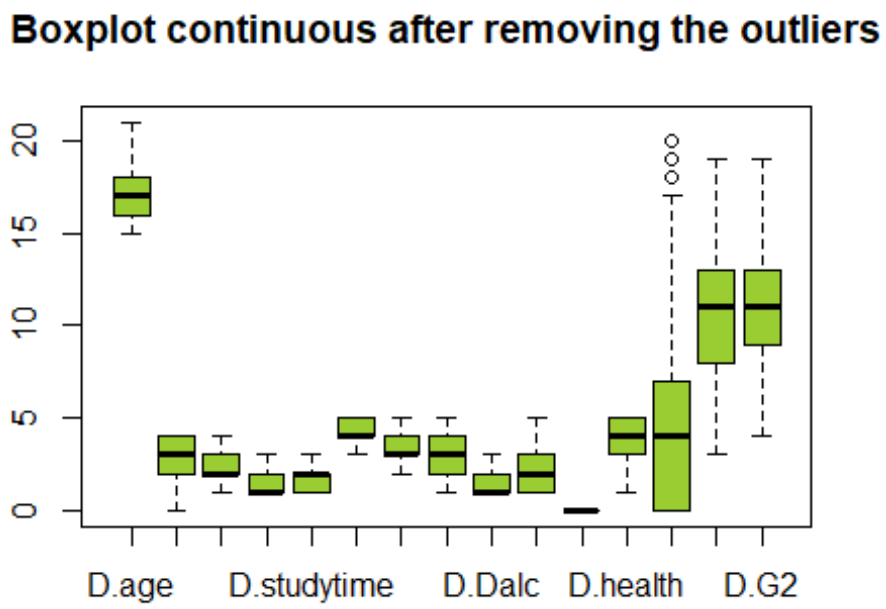
```
library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(as.matrix(data.frame(Y,Data_cont[,-11]))),method = "circle",type
= "lower",title = paste("Cicle Correlogram"))
```



All the points outside the box-whisker are named as outlier for any variables.

So we replace all the outliers by the median of the corresponding variable and below is their box-plot.

```
par(mfrow = c(1,1))
i=1
while(i<=ncol(Data_cont))
{
  if (length(boxplot.stats(Data_cont[,i])$out) > 0)
  {
    out = boxplot.stats(Data_cont[,i])$out
    outlier = which(Data_cont[,i] %in% c(out))
    for (j in 1:length(outlier))
    {
      Data_cont[outlier[j],i] = median(Data_cont[,i])
    }
  }
  i = i + 1
}
boxplot (Data_cont,col = "yellowgreen",main = paste("Boxplot continuous after
removing the outliers"))
```




```
D1 = data.frame(D1)
Dataclass = D1[,-which(names(D1) %in% c("age","Medu","Fedu","traveltime","stu
dytime","famrel","freetime","goout","Dalc","Walc","failures","health","absenc
es","G1","G2"))]
dim(Dataclass)
```

```
## [1] 395  26
```

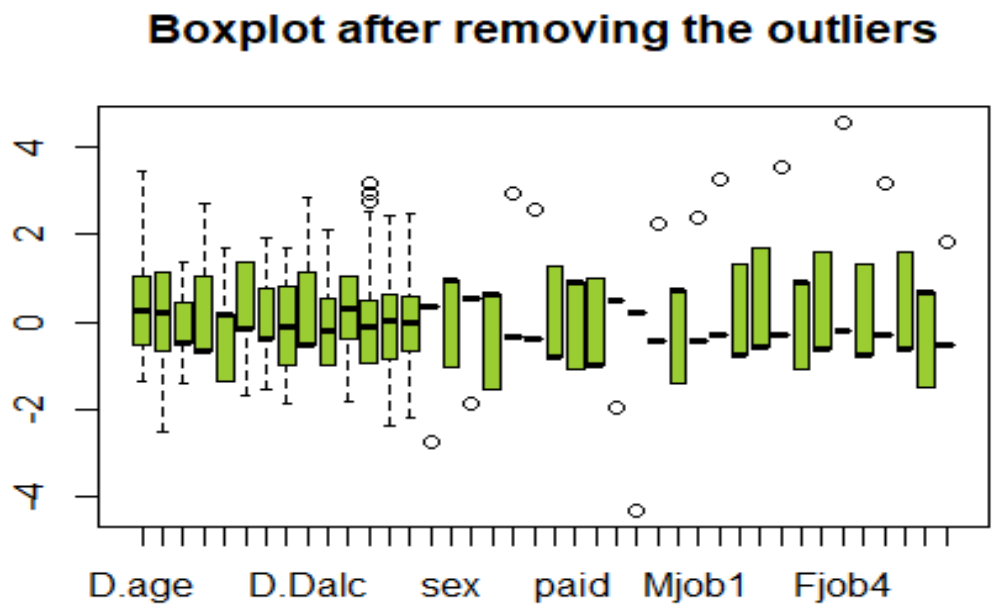
```
D1 = data.frame(as.data.frame(scale(Data_cont,center = TRUE,scale = TRUE)),Da
taclass)
colnames(D1[,11])
```

```
## NULL
```

```
D1 = D1[,-11]
dim(D1)
```

```
## [1] 395  40
```

```
boxplot(D1,col = "yellowgreen",main = paste("Boxplot after removing the outli
ers"))
```



We didn't replace the outlier for the binary dummy variables because there exists a possibility of variable deletion as an effect of outlier removal.

We now fit the regression model with the regressors as Dalc(Workday alcohol consumption habit),Walc(Weekend alcohol consumption habit) and romantic(Whether the student fall in relationship or not). And te summary of this regresssion is –

```
summary(lm(Y~D.Dalc+D.Walc+romantic,data = D1))

##
## Call:
## lm(formula = Y ~ D.Dalc + D.Walc + romantic, data = D1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.42777 -0.43089  0.06118  0.64966  1.93766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.759e-16  4.988e-02   0.000  1.00000
## D.Dalc      -8.107e-02  5.734e-02  -1.414  0.15820
## D.Walc      -1.393e-02  5.717e-02  -0.244  0.80768
## romantic     1.363e-01  5.011e-02   2.720  0.00682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9913 on 391 degrees of freedom
```

```
## Multiple R-squared:  0.02472,    Adjusted R-squared:  0.01723
## F-statistic: 3.303 on 3 and 391 DF,  p-value: 0.02037
```

Since the R-squared and the Adjusted R-squared values are really small hence it is cleared that the model is not explained properly by the regressors.

Finally the F-statistic value is 3.303 and the corresponding p-value is 0.02037 which is greater than 0.01, so we fail to reject the null hypothesis at 1% level of significance. So the regression is not valid or more elegantly these regressors have not significant effect on the response i.e. student's grade is not significanly affected by the alcohol consumption habit or falling in relationship.

let us now see the other variables effect on the response variable.

```
Model = lm(Y~.,as.data.frame(D1))
summary(Model)

##
## Call:
## lm(formula = Y ~ ., data = as.data.frame(D1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77821 -0.16485  0.07631  0.30952  1.13110
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.477e-16  2.638e-02   0.000  1.00000
## D.age        -4.461e-02  3.577e-02  -1.247  0.21309
## D.Medu        2.340e-02  4.481e-02   0.522  0.60182
## D.Fedu       -2.846e-02  3.795e-02  -0.750  0.45388
## D.traveltime -5.021e-02  2.984e-02  -1.683  0.09332 .
## D.studytime   2.406e-02  3.130e-02   0.769  0.44268
## D.famrel      1.542e-02  2.811e-02   0.549  0.58360
## D.freetime    5.704e-03  2.915e-02   0.196  0.84495
## D.goout      -5.435e-02  3.184e-02  -1.707  0.08872 .
## D.Dalc       -1.929e-02  3.240e-02  -0.595  0.55207
## D.Walc        6.017e-02  3.713e-02   1.621  0.10598
## D.health      5.509e-03  2.850e-02   0.193  0.84684
## D.absences    1.624e-01  3.035e-02   5.350 1.59e-07 ***
## D.G1          3.711e-01  6.395e-02   5.804 1.44e-08 ***
## D.G2          5.030e-01  6.406e-02   7.852 4.94e-14 ***
## school       -7.828e-02  3.257e-02  -2.404  0.01674 *
## sex          -2.260e-02  3.223e-02  -0.701  0.48354
## address       3.143e-02  3.080e-02   1.020  0.30821
## famsize      -1.405e-02  2.861e-02  -0.491  0.62373
## Pstatus       2.015e-02  2.819e-02   0.715  0.47537
## schoolsup     8.457e-02  2.946e-02   2.871  0.00434 **
## famsup       -5.344e-03  3.002e-02  -0.178  0.85882
## paid         -6.320e-02  3.029e-02  -2.086  0.03767 *
```

```
## activities    3.411e-02  2.846e-02   1.199  0.23142
## nursery      -1.715e-02  2.851e-02  -0.602  0.54786
## higher        5.129e-03  2.982e-02   0.172  0.86353
## internet      8.266e-03  2.964e-02   0.279  0.78052
## romantic      7.904e-02  2.851e-02   2.773  0.00585 **
## Mjob1        -1.096e-02  4.734e-02  -0.232  0.81699
## Mjob2        -6.863e-03  3.455e-02  -0.199  0.84267
## Mjob3         1.976e-02  5.169e-02   0.382  0.70252
## Mjob4        -1.960e-02  4.381e-02  -0.447  0.65489
## Fjob1         1.620e-02  4.327e-02   0.374  0.70826
## Fjob2         1.051e-01  6.534e-02   1.608  0.10876
## Fjob3         9.895e-02  6.114e-02   1.618  0.10648
## Fjob4         4.839e-02  3.806e-02   1.271  0.20442
## reason1      -3.535e-02  3.550e-02  -0.996  0.32012
## reason2       1.544e-02  3.151e-02   0.490  0.62458
## reason3      -3.717e-02  3.399e-02  -1.094  0.27484
## guardian1     7.066e-02  5.399e-02   1.309  0.19146
## guardian2     6.659e-02  5.325e-02   1.250  0.21196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5243 on 354 degrees of freedom
## Multiple R-squared:  0.753,  Adjusted R-squared:  0.7251
## F-statistic: 26.98 on 40 and 354 DF,  p-value: < 2.2e-16
```

Now the regression is significant and but the R-squared value is 0.7561 and the adjusted R-squared value is 0.7285, which are quite small. So we proceed further.

Residual analysis 1.different variables and residuals

```
library(olsrr)

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

ols_plot_comp_plus_resid(Model)
```

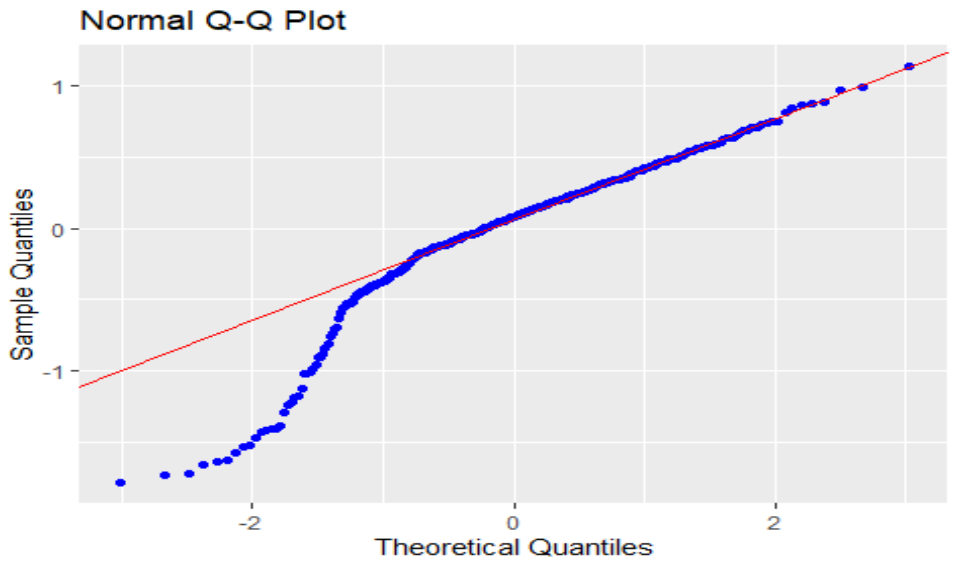
2.Homoscadasticity checking

```
ols_test_breusch_pagan(Model)

##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##             Data
```

```
##  ----------------------------
##  Response : Y
##  Variables: fitted values of Y
##
##          Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    108.1665
##  Prob > Chi2   =    2.471174e-25
```

Since the p-value is less than 0.01, so the null hypothesis is rejected and the error distribution is not homoscedastic. 3.a.Normality Checking : Q-Q Plot

```
ols_plot_resid_qq(Model)
```



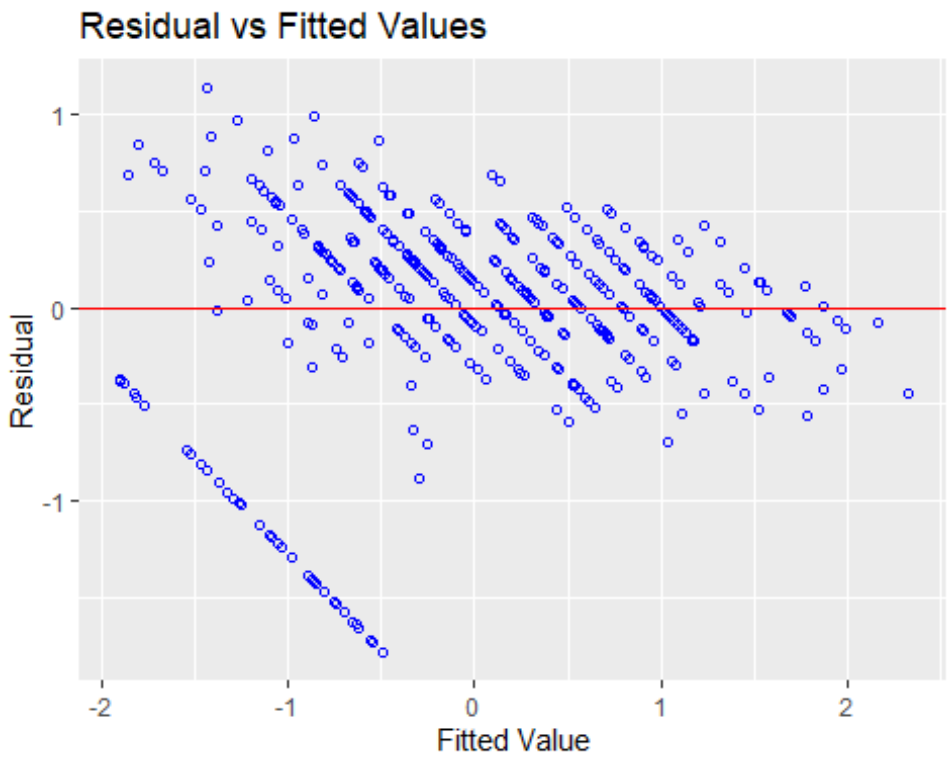
3.b.Normality Checking : Theoretiical vai hypthesis

```
ols_test_normality(Model)

## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9042         0.0000
## Kolmogorov-Smirnov        0.127          0.0000
## Cramer-von Mises         48.4784         0.0000
## Anderson-Darling         10.4358         0.0000
## -----------------------------------------------
```

From the Q-Q plot and also from the theoretical checking we can safely conclude that the error distribution is not normal.

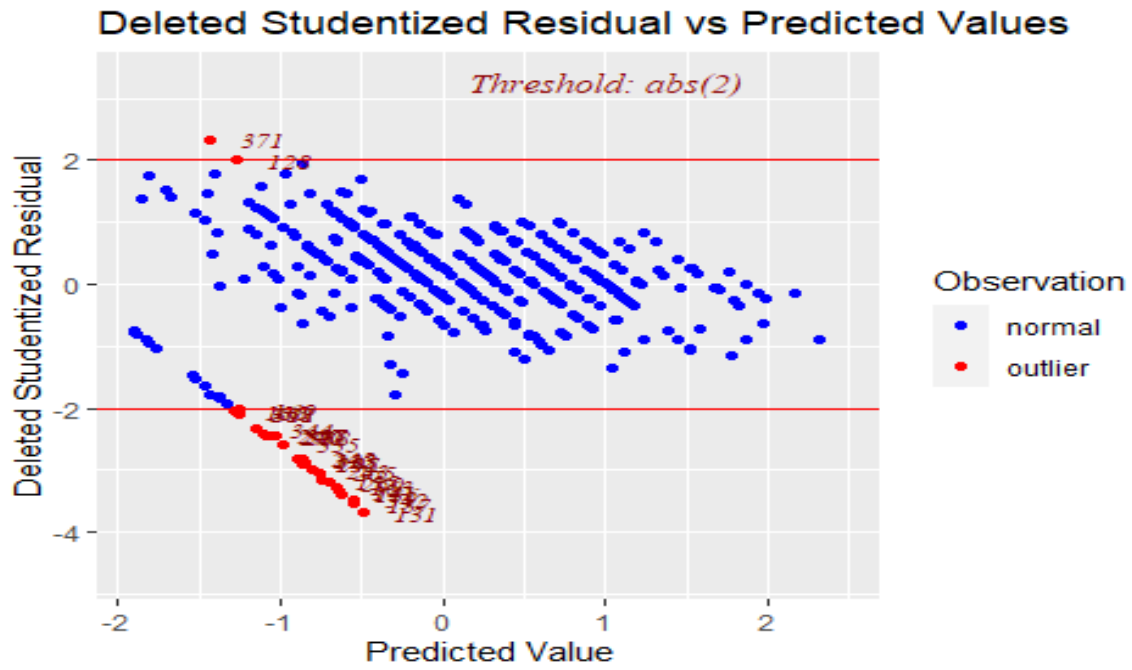
4.Residual vs predicted value plot

```
ols_plot_resid_fit(Model)
```



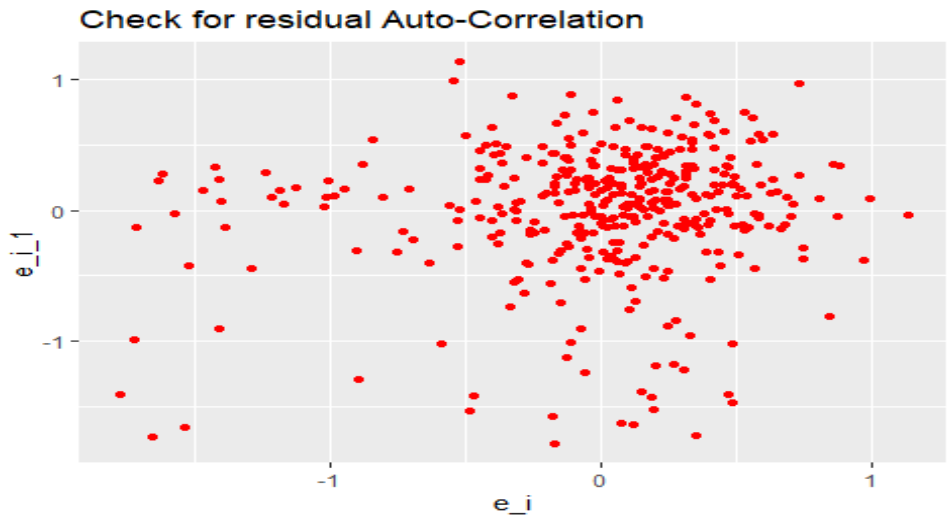
5.Studentized residual vs Predicted value plot

```
ols_plot_resid_stud_fit(Model)
```



Both the above two plots does not say more about the problem in the model. 6. Plot for autocorrelation.

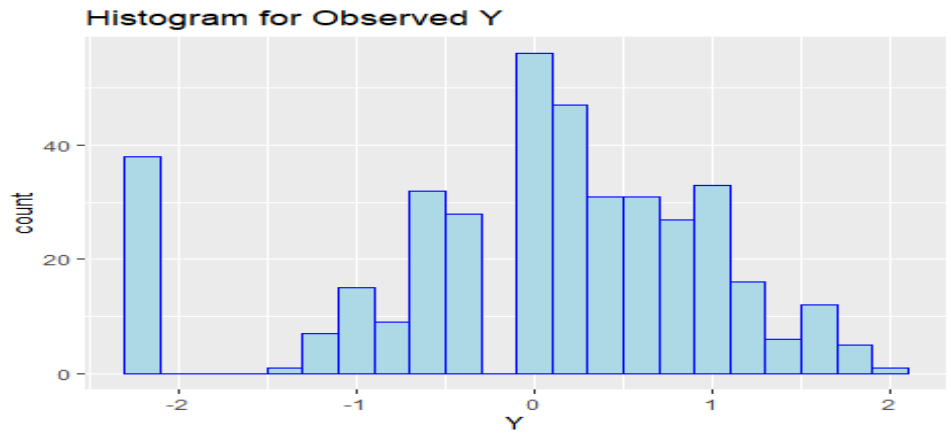
```
res_hat = residuals.lm(Model)
e_i=res_hat[-length(res_hat)]
e_i_1 = res_hat[-1]
residual = data.frame(e_i,e_i_1)
ggplot(data = residual,mapping = aes(e_i,e_i_1)) + geom_point(fill = "red",co
lor = "red") + ggtitle("Check for residual Auto-Correlation")
```



Since there is no pattern in the plot, so can say that there is no autocorrelation in the data.

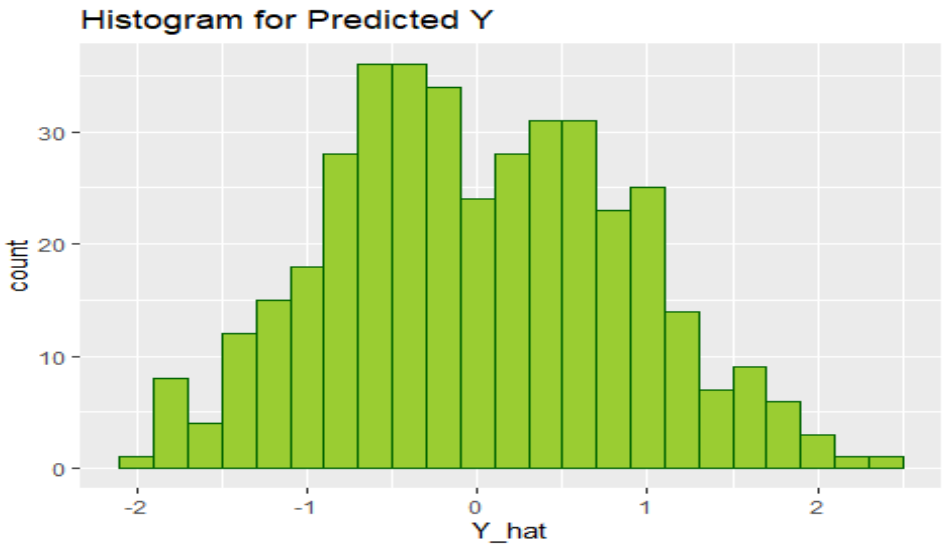
let us see the histogram of the 1.response variable.

```
Y_hat = predict.lm(Model)  #Predicted Y
res_hat = residuals.lm(Model)                #residuals "observed error"
Predicted = data.frame(Y,Y_hat,res_hat)
ggplot(data = Predicted,mapping = aes(Y)) + geom_histogram(color = "blue",fil
l = "lightblue",binwidth = 0.2) + ggtitle("Histogram for Observed Y")
```



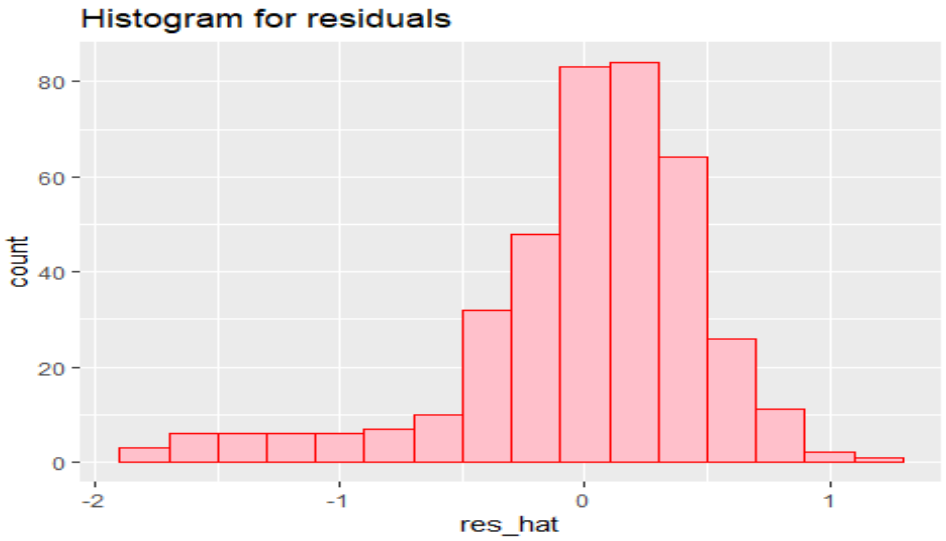
2.Predicted Response variable.

```
ggplot(data = Predicted,mapping = aes(Y_hat)) + geom_histogram(color = "darkg
reen",fill = "yellowgreen",binwidth = 0.2) + ggtitle("Histogram for Predicted
Y")
```



3.error variable.

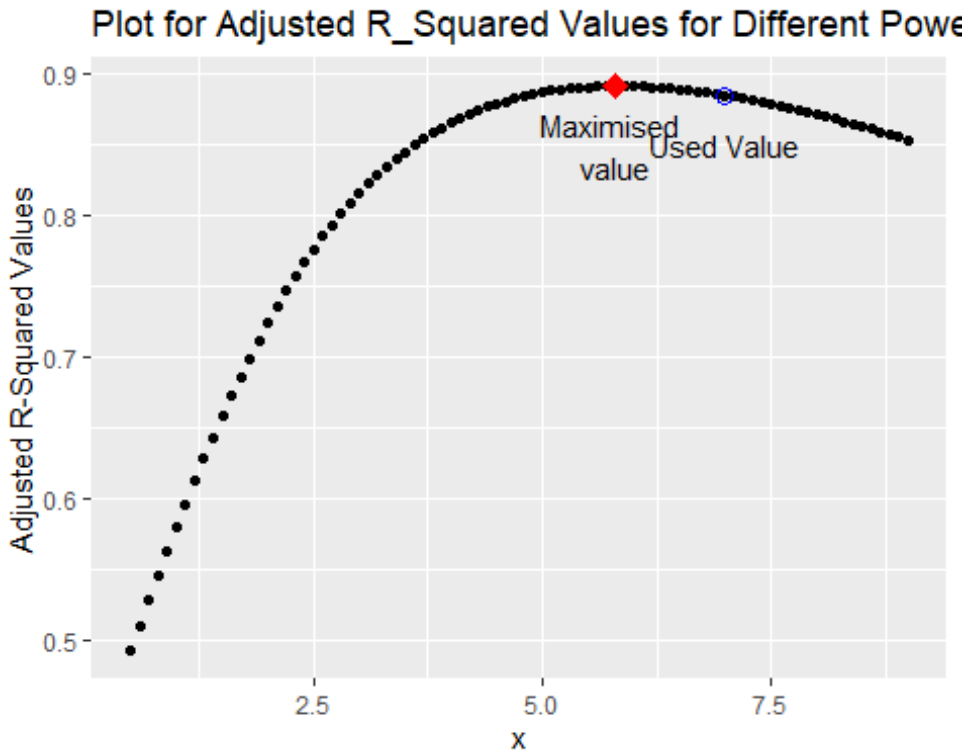
```
ggplot(data = Predicted,mapping = aes(res_hat)) + geom_histogram(color = "red
",fill = "pink",binwidth = 0.2) + ggtitle("Histogram for residuals")
```



So we would transform the response by the sinh_inverse method i,e, we write sinh_inverse(response)^x and the value of x is determined by choosing that value by whih we can have the maximum adjusted R_squared value.

```
De = data.frame(Raw_Y,Raw_data)
De = De[,-12]
```

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

De= De [,-1]


x = seq(0.5,9,by = 0.1)
adj = NULL
for (i in 1:length(x) )
{
  new_Y = asinh(Raw_Y)^(x[i])
  MM = lm(new_Y~.,data = De)
  adj[i] = summary(MM)$adj.r.squared
}
library(ggplot2)
R_sq = data.frame(x,adj)
p = ggplot(data = R_sq,mapping = aes(x,adj)) + geom_point() + ggtitle("Plot f
or Adjusted R_Squared Values for Different Powers") + labs(y= "Adjusted R-Squ
ared Values")
data1 = data.frame(5.8,adj[54])
p1 = p + geom_point(data1,mapping = aes(5.8,adj[54]),shape = 23,fill = "red",
color = "red",size = 3) + annotate ("text",x=5.8,y=0.85,label = "Maximised \n
value")
data2 = data.frame(7.0,adj[66])
p1 + geom_point(data2,mapping = aes(7.0,adj[66]),shape = 1,fill = "blue",colo
r = "blue",size = 3) + annotate ("text",x=7.0,y=0.85,label = "Used Value")
```



```
for (i in 1:length(x))
{
  if (x[i]==7.0)
  {
    print(i)
    print(x[i])
    print(adj[i])
  }
}
```

```
## [1] 66
## [1] 7
## [1] 0.885374
```

The point indicates the maximum adjusted R_squared value and the blue circle indicates the model that best fits the Q-Q plot. Since the curve is flat the the point of maximization implies that the difference in adjusted R_squared value is very small between these two points. So we can safely choose the value with blue circle i.e. x = 7.0.

So the new response varible is – sinh_inverse(response)^7 So the new model summary is –

```
new_Y = asinh(Raw_Y)^(7.0)
MM = lm(new_Y~.,data = De)
summary(MM)
```

```
##
## Call:
## lm(formula = new_Y ~ ., data = De)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2247.70  -402.09   -19.34   415.77  1864.98
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2159.778    884.491  -2.442  0.01510 *
## school        -18.703    133.475  -0.140  0.88864
## sex           -36.867     85.085  -0.433  0.66506
## age           -80.669     36.668  -2.200  0.02845 *
## address        17.917     98.384   0.182  0.85560
## famsize         4.497     82.552   0.054  0.95658
## Pstatus       114.631    122.251   0.938  0.34905
## Medu          101.273     54.632   1.854  0.06461 .
## Fedu          -69.744     46.433  -1.502  0.13398
## traveltime     23.535     57.475   0.409  0.68243
## studytime     -59.467     48.939  -1.215  0.22513
## schoolsup    -141.442    116.367  -1.215  0.22499
## famsup       -118.221     81.669  -1.448  0.14863
## paid          178.375     80.336   2.220  0.02703 *
## activities    100.534     75.009   1.340  0.18101
## nursery       -80.823     92.576  -0.873  0.38323
```

```
## higher       -130.154    179.431  -0.725  0.46870
## internet      -95.593    104.723  -0.913  0.36196
## romantic       93.286     79.994   1.166  0.24433
## famrel        127.814     41.494   3.080  0.00223 **
## freetime      -12.281     40.103  -0.306  0.75959
## goout          24.580     38.292   0.642  0.52135
## Dalc          -58.999     55.764  -1.058  0.29077
## Walc           -3.963     41.843  -0.095  0.92459
## health        -12.926     27.243  -0.474  0.63547
## absences       -6.782      4.884  -1.389  0.16584
## G1            220.360     22.518   9.786  < 2e-16 ***
## G2            329.470     19.460  16.930  < 2e-16 ***
## Mjob1          91.546    175.401   0.522  0.60205
## Mjob2          71.901    162.624   0.442  0.65866
## Mjob3         127.200    141.839   0.897  0.37044
## Mjob4         192.952    130.289   1.481  0.13951
## Fjob1         166.767    218.984   0.762  0.44684
## Fjob2          28.306    173.548   0.163  0.87053
## Fjob3         -92.762    179.679  -0.516  0.60599
## Fjob4          78.326    242.611   0.323  0.74700
## reason1       -14.836     97.350  -0.152  0.87896
## reason2        88.747    145.520   0.610  0.54235
## reason3       -33.450    100.440  -0.333  0.73931
## guardian1     118.273    150.488   0.786  0.43244
## guardian2      -4.041    164.446  -0.025  0.98041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 692.4 on 354 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8854
## F-statistic: 77.08 on 40 and 354 DF,  p-value: < 2.2e-16
```

Now the regression is significant and but the R-squared value is 0.897 and the adjusted R-squared value is 0.8854, which are quite better than the previous. So let us proceed further–

Residual Analysis 1. Residual vs different regressors

```
library(olsrr)
ols_plot_comp_plus_resid(MM)
```

2. Homoscedasticity checking

```
ols_test_breusch_pagan(MM)
```

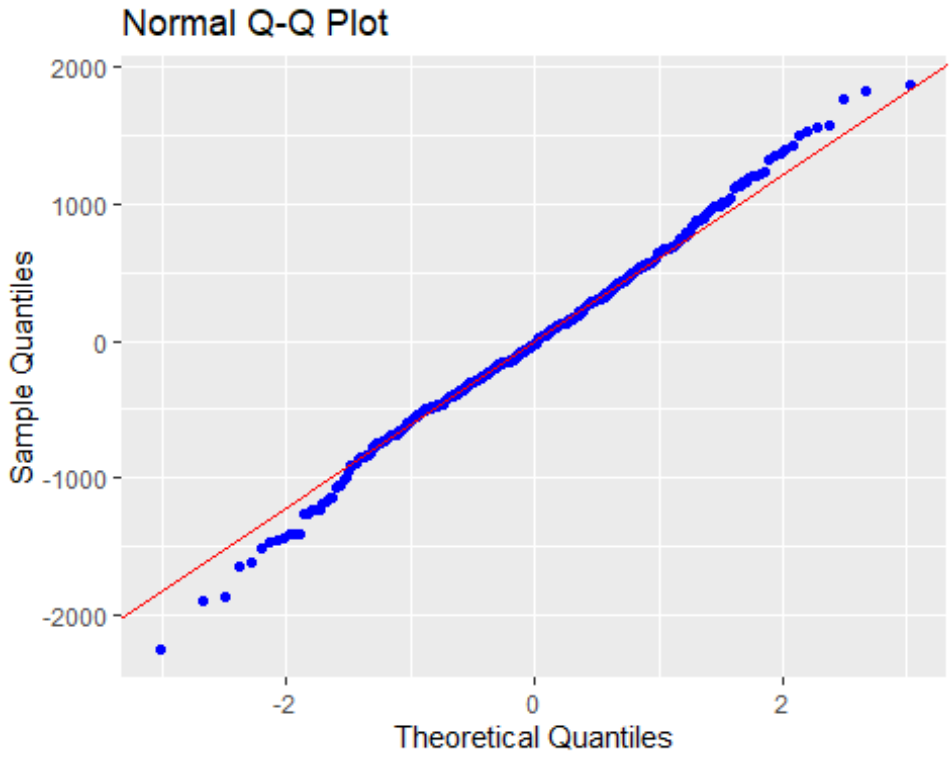
```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
```

```
##                Data
##  --------------------------------
##  Response : new_Y
##  Variables: fitted values of new_Y
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    1.357253
##  Prob > Chi2   =    0.2440142
```

Now th p-value is greater than 0.05, So we can say that the error distribution is homoscedastic under 5% level of significance.

3.a.Normality Checking : Q-Q Plot

```
ols_plot_resid_qq(MM)
```



The above Q-Q plot suggests that the error distribution is approximately normal. 3.b.Normality Checking : Theoretical Checking via hypothesis

```
ols_test_normality(MM)
```

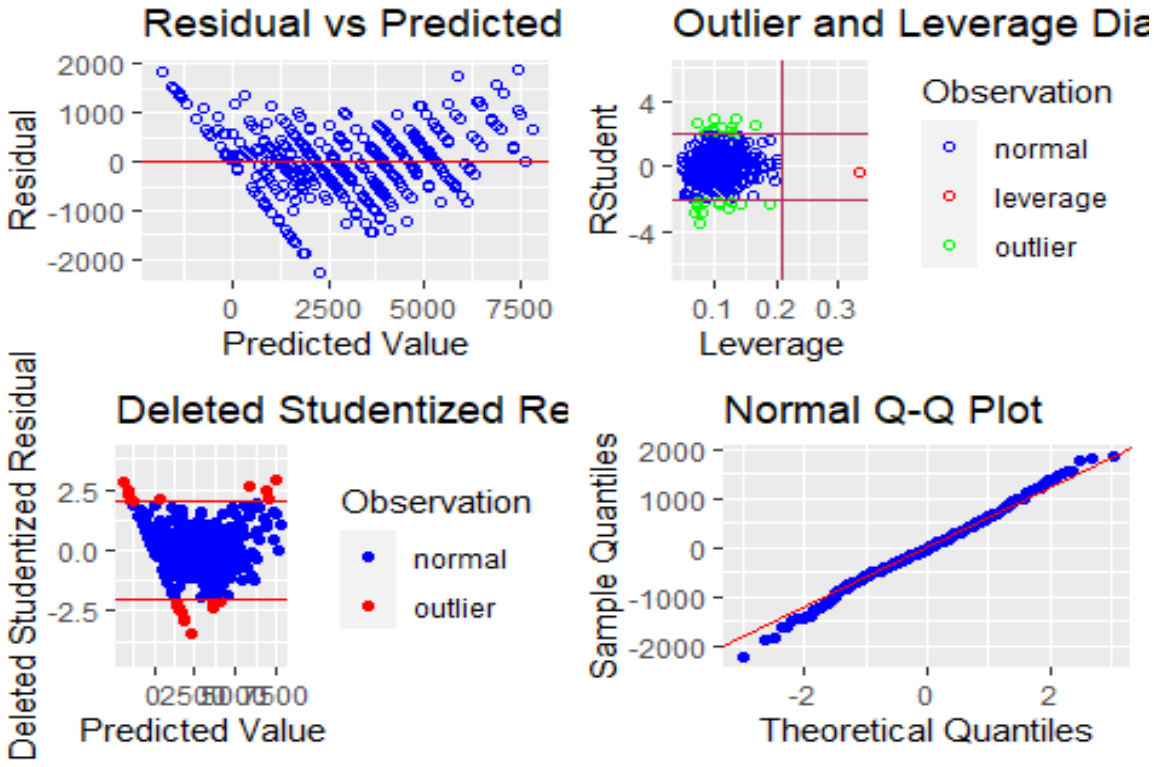
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9949         0.2200
## Kolmogorov-Smirnov        0.036          0.6853
```

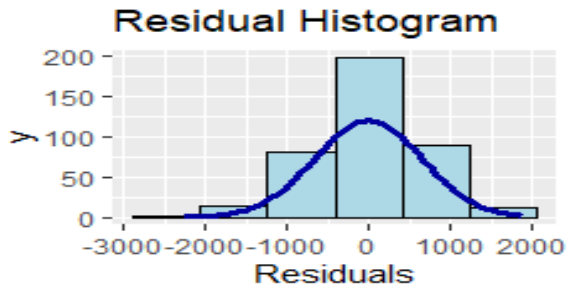
```
## Cramer-von Mises          32.9477          0.0000
## Anderson-Darling           0.5871          0.1253
## ----------------------------------------------
```

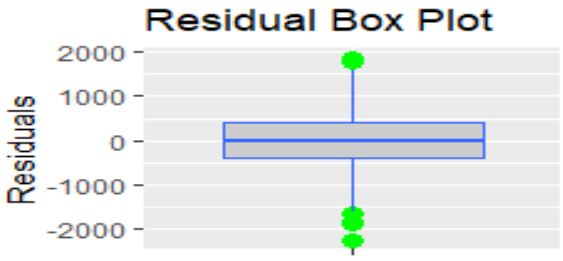
Since 3 out of 4 test validates the assumption i.e. the error distribution is normal. So we can safely assume that the error distribution is normal.

4. model diagnostics

```
ols_plot_diagnostics(MM)
```

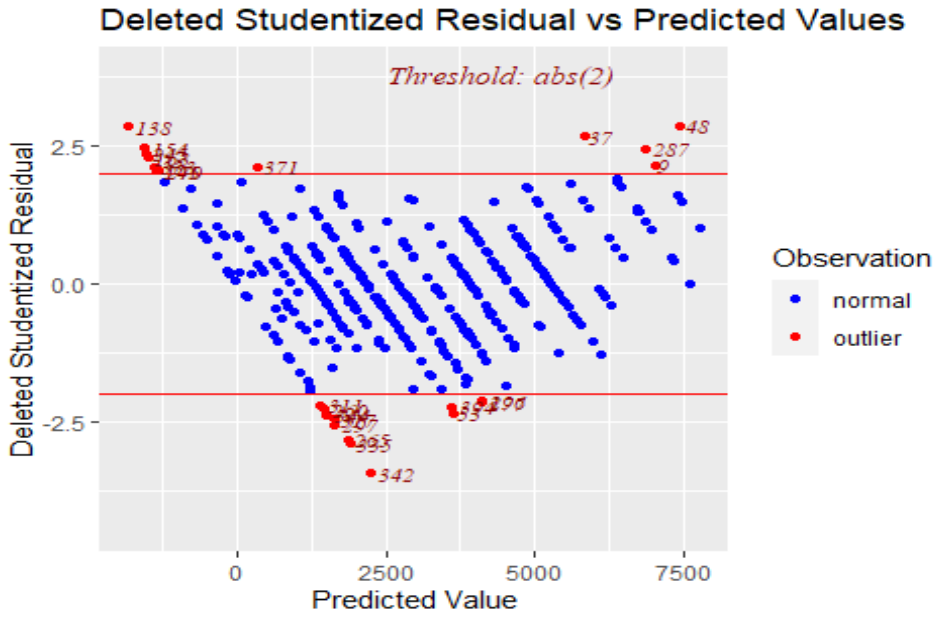






5. Studentized residual vs predicted response variaible plot

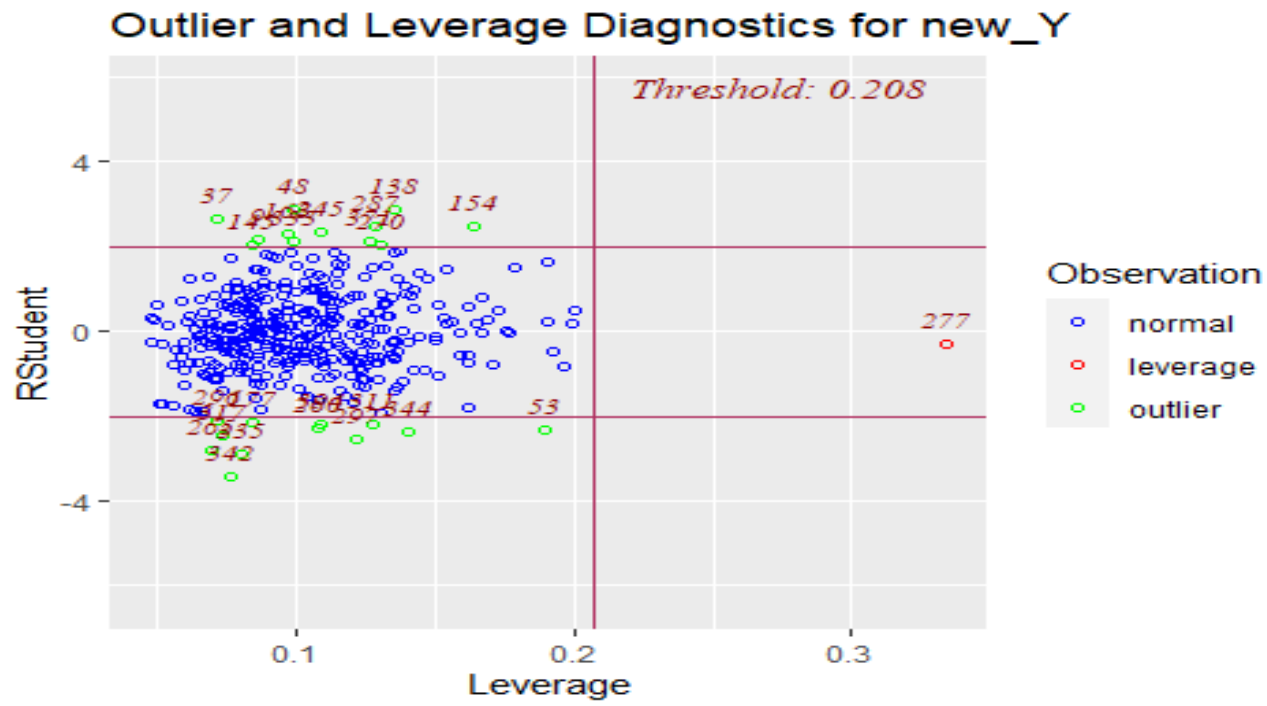
```
ols_plot_resid_stud_fit(MM)
```



The above plot suggets that the outlier are qite small in our model.

Leverage point checking

```
ols_plot_resid_lev(MM)
```



There exists a leverage point at the index 277. Let us now remove that.

```
De = data.frame(new_Y,De)

dim(De)

## [1] 395  41

colnames(De)

##  [1] "new_Y"      "school"     "sex"        "age"        "address"
##  [6] "famsize"    "Pstatus"    "Medu"       "Fedu"       "traveltime"
## [11] "studytime"  "schoolsup"  "famsup"     "paid"       "activities"
## [16] "nursery"    "higher"     "internet"   "romantic"   "famrel"
## [21] "freetime"   "goout"      "Dalc"       "Walc"       "health"
## [26] "absences"   "G1"         "G2"         "Mjob1"      "Mjob2"
## [31] "Mjob3"      "Mjob4"      "Fjob1"      "Fjob2"      "Fjob3"
## [36] "Fjob4"      "reason1"    "reason2"    "reason3"    "guardian1"
## [41] "guardian2"

De = De[-277,]
new_Y = De$new_Y


De = De[,-1]
dim(De)

## [1] 394  40
```

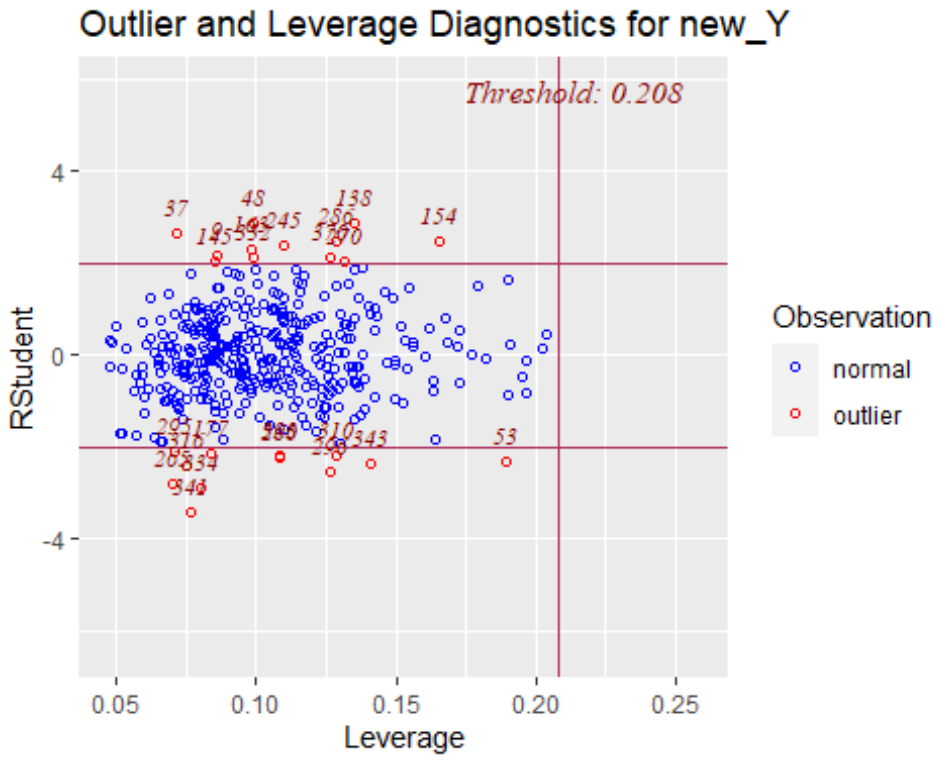
Let us again fit the model –

```
MM = lm(new_Y~.,data = De)    #again model checking
summary(MM)

##
## Call:
## lm(formula = new_Y ~ ., data = De)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2242.10  -400.65   -18.91   418.16  1860.07
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2145.392    886.735  -2.419   0.0160 *
## school        -18.344    133.648  -0.137   0.8909
## sex           -36.082     85.227  -0.423   0.6723
## age           -81.169     36.747  -2.209   0.0278 *
## address        15.159     98.879   0.153   0.8782
## famsize         5.552     82.721   0.067   0.9465
## Pstatus       120.297    123.661   0.973   0.3313
## Medu          101.397     54.703   1.854   0.0646 .
```

```
## Fedu          -69.575     46.495  -1.496   0.1354
## traveltime     24.636     57.649   0.427   0.6694
## studytime     -58.650     49.066  -1.195   0.2328
## schoolsup    -142.648    116.574  -1.224   0.2219
## famsup       -116.070     82.045  -1.415   0.1580
## paid          177.262     80.512   2.202   0.0283 *
## activities    101.260     75.138   1.348   0.1786
## nursery       -85.638     93.889  -0.912   0.3623
## higher       -144.682    185.222  -0.781   0.4352
## internet      -95.484    104.856  -0.911   0.3631
## romantic       93.866     80.116   1.172   0.2421
## famrel        128.172     41.562   3.084   0.0022 **
## freetime      -12.854     40.193  -0.320   0.7493
## goout          24.015     38.380   0.626   0.5319
## Dalc          -59.106     55.836  -1.059   0.2905
## Walc           -5.339     42.112  -0.127   0.8992
## health        -12.203     27.369  -0.446   0.6560
## absences       -5.996      5.465  -1.097   0.2734
## G1            220.530     22.552   9.779   <2e-16 ***
## G2            329.381     19.487  16.903   <2e-16 ***
## Mjob1          90.128    175.679   0.513   0.6083
## Mjob2          74.070    162.969   0.455   0.6497
## Mjob3         128.041    142.043   0.901   0.3680
## Mjob4         191.786    130.505   1.470   0.1426
## Fjob1         164.844    219.343   0.752   0.4528
## Fjob2          26.404    173.868   0.152   0.8794
## Fjob3         -91.669    179.939  -0.509   0.6108
## Fjob4          76.615    242.977   0.315   0.7527
## reason1       -13.272     97.594  -0.136   0.8919
## reason2        88.284    145.712   0.606   0.5450
## reason3       -30.784    100.907  -0.305   0.7605
## guardian1     123.414    151.520   0.815   0.4159
## guardian2       1.210    165.458   0.007   0.9942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 693.3 on 353 degrees of freedom
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8853
## F-statistic: 76.81 on 40 and 353 DF,  p-value: < 2.2e-16
```

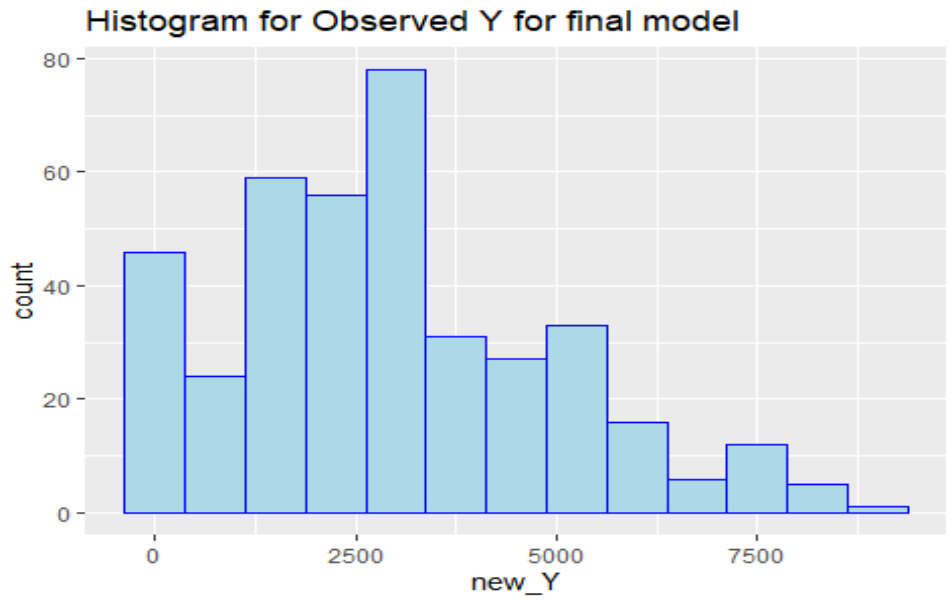
Finally check for the leverage

```
ols_plot_resid_lev(MM)        #again leverage plot
```



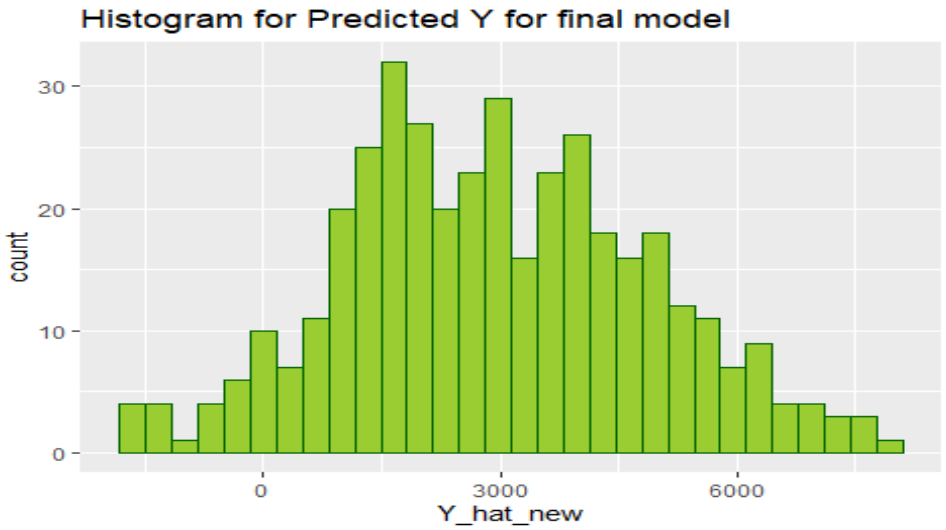
So there is not the leverage point. So the final histogram of 1. response variable

```
Y_hat_new = predict.lm(MM)
res_hat_new = MM$residuals
Predicted_final = data.frame(new_Y,Y_hat_new,res_hat_new)
ggplot(data = Predicted_final,mapping = aes(new_Y)) + geom_histogram(color =
"blue",fill = "lightblue",binwidth =750) + ggtitle("Histogram for Observed Y
for final model")
```



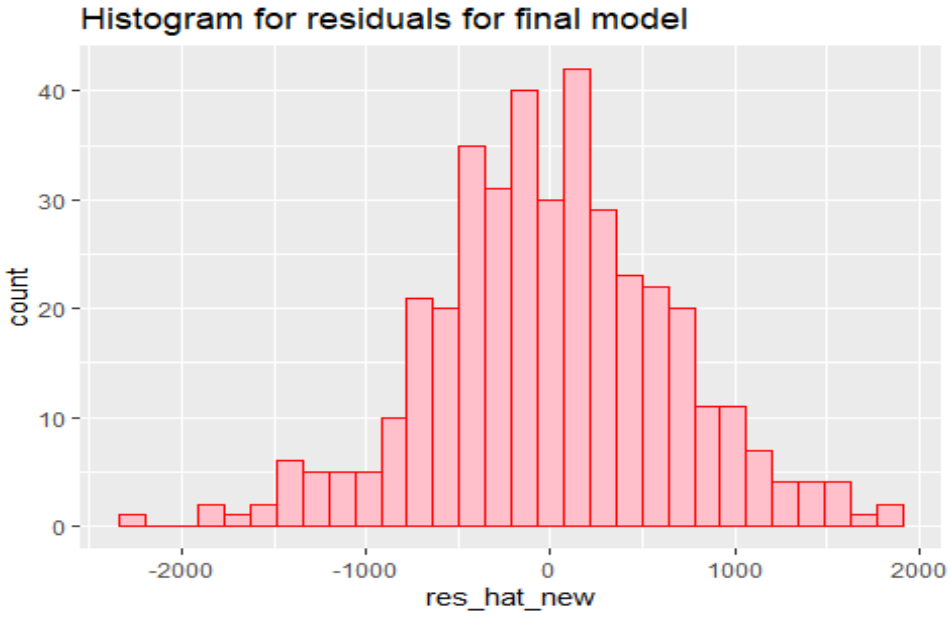
2.Predicted response variable

```
ggplot(data = Predicted_final,mapping = aes(Y_hat_new)) + geom_histogram(colo
r = "darkgreen",fill = "yellowgreen") + ggtitle("Histogram for Predicted Y fo
r final model")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.residual

```
ggplot(data = Predicted_final,mapping = aes(res_hat_new)) + geom_histogram(co
lor = "red",fill = "pink") + ggtitle("Histogram for residuals for final model
")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Checking For Multicollinearity

Let us check via the VIF method

```
library(DescTools)
sort(VIF(MM))
```

```
##     famrel    Pstatus    famsize activities   romantic    nursery     heal
th
##   1.138377   1.143370   1.153333   1.156468   1.167666   1.169243   1.1835
98
##  schoolsup   internet   absences     higher   freetime     famsup       pa
id
##   1.255243   1.256796   1.269135   1.290706   1.304345   1.307425   1.3195
73
## traveltime    address  studytime    reason2      goout        sex     scho
ol
##   1.323228   1.378876   1.389910   1.444882   1.482833   1.484660   1.5098
13
##    reason3      Mjob2        age    reason1       Dalc       Fedu      Fjo
b4
##   1.660698   1.716523   1.797482   1.815808   2.026085   2.097141   2.1098
37
##       Walc      Fjob1      Mjob4       Medu      Mjob1      Mjob3  guardia
n2
##   2.405020   2.689030   2.695511   2.939196   3.221013   3.788406   3.9550
19
##  guardian1         G2         G1      Fjob3      Fjob2
##   4.022799   4.401860   4.592133   5.340952   6.130967
```

Since Fjob3 and Fjob2 has VIF value greater than 5 So we can drop the highest one and check tha VIF agian.

```
De = within(De,rm("Fjob2"))
MM = lm(new_Y~.,data = De)
summary(MM)
```

```
##
## Call:
## lm(formula = new_Y ~ ., data = De)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2239.50  -402.62   -20.54   419.93  1858.33
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2116.827    865.360  -2.446  0.01492 *
## school        -17.143    133.230  -0.129  0.89769
## sex           -36.899     84.940  -0.434  0.66426
## age           -81.562     36.605  -2.228  0.02650 *
```

```
## address        15.158     98.743   0.154  0.87809
## famsize         6.022     82.549   0.073  0.94188
## Pstatus       119.512    123.382   0.969  0.33339
## Medu          101.745     54.580   1.864  0.06313 .
## Fedu          -69.706     46.423  -1.502  0.13411
## traveltime     25.732     57.117   0.451  0.65262
## studytime     -58.481     48.986  -1.194  0.23334
## schoolsup    -143.486    116.283  -1.234  0.21804
## famsup       -115.598     81.873  -1.412  0.15885
## paid          176.829     80.350   2.201  0.02840 *
## activities    101.573     75.006   1.354  0.17654
## nursery       -87.123     93.249  -0.934  0.35078
## higher       -144.653    184.966  -0.782  0.43471
## internet      -94.522    104.520  -0.904  0.36643
## romantic       94.298     79.954   1.179  0.23903
## famrel        128.681     41.369   3.111  0.00202 **
## freetime      -13.339     40.011  -0.333  0.73904
## goout          24.048     38.327   0.627  0.53076
## Dalc          -59.956     55.478  -1.081  0.28057
## Walc           -4.518     41.707  -0.108  0.91379
## health        -12.243     27.330  -0.448  0.65445
## absences       -5.956      5.452  -1.092  0.27537
## G1            220.145     22.378   9.837  < 2e-16 ***
## G2            329.635     19.389  17.001  < 2e-16 ***
## Mjob1          88.732    175.196   0.506  0.61284
## Mjob2          75.550    162.453   0.465  0.64218
## Mjob3         129.202    141.641   0.912  0.36229
## Mjob4         191.727    130.324   1.471  0.14214
## Fjob1         141.919    158.912   0.893  0.37243
## Fjob3        -115.347     89.697  -1.286  0.19930
## Fjob4          53.040    186.662   0.284  0.77646
## reason1       -13.353     97.458  -0.137  0.89110
## reason2        88.209    145.510   0.606  0.54477
## reason3       -31.405    100.685  -0.312  0.75529
## guardian1     122.933    151.278   0.813  0.41698
## guardian2      -0.326    164.920  -0.002  0.99842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 692.3 on 354 degrees of freedom
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8856
## F-statistic:    79 on 39 and 354 DF,  p-value: < 2.2e-16

sort(VIF(MM))

##     famrel    Pstatus    famsize activities    nursery   romantic     heal
th
##   1.130963   1.141372   1.151718   1.155598   1.156560   1.166193   1.1834
89
##      Fjob4   internet  schoolsup   absences     higher   freetime travelti
```

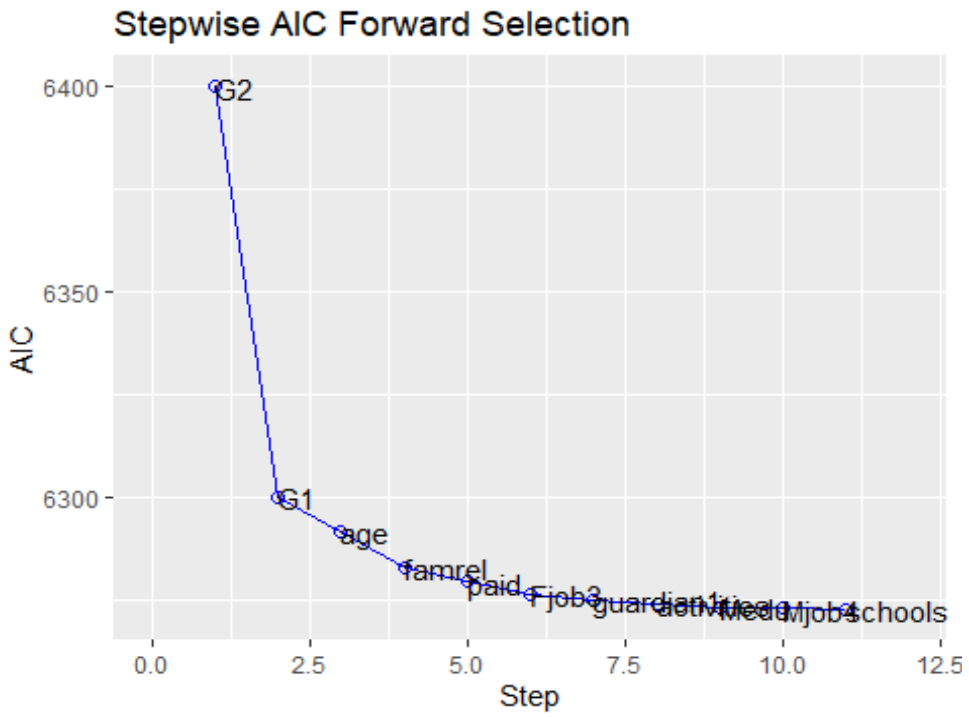
```
me
##   1.248615   1.252201   1.252428   1.266218   1.290705   1.296121   1.3025
02
##     famsup       paid      Fjob3    address  studytime      Fjob1     reaso
n2
##   1.305550   1.317914   1.330835   1.378876   1.389192   1.415343   1.4448
66
##        sex      goout     school    reason3      Mjob2        age     reaso
n1
##   1.478752   1.482785   1.504526   1.657971   1.710385   1.788573   1.8157
54
##       Dalc       Fedu       Walc      Mjob4       Medu      Mjob1       Mjo
b3
##   2.005758   2.096419   2.365465   2.695487   2.934064   3.212194   3.7774
25
##  guardian2  guardian1         G2         G1
##   3.940240   4.021043   4.369571   4.534113
```

Now there is no multicollinearity in the model and the adjusted R_squred value is also approximately equal to the preious one. So our model building job done and we continue with our final job i.e. variable selection.

We select the variable by the minimum AIC criterion.

1.Forward Selection

```
for_aic = ols_step_forward_aic(MM)
plot(for_aic)
```



So the sumary –

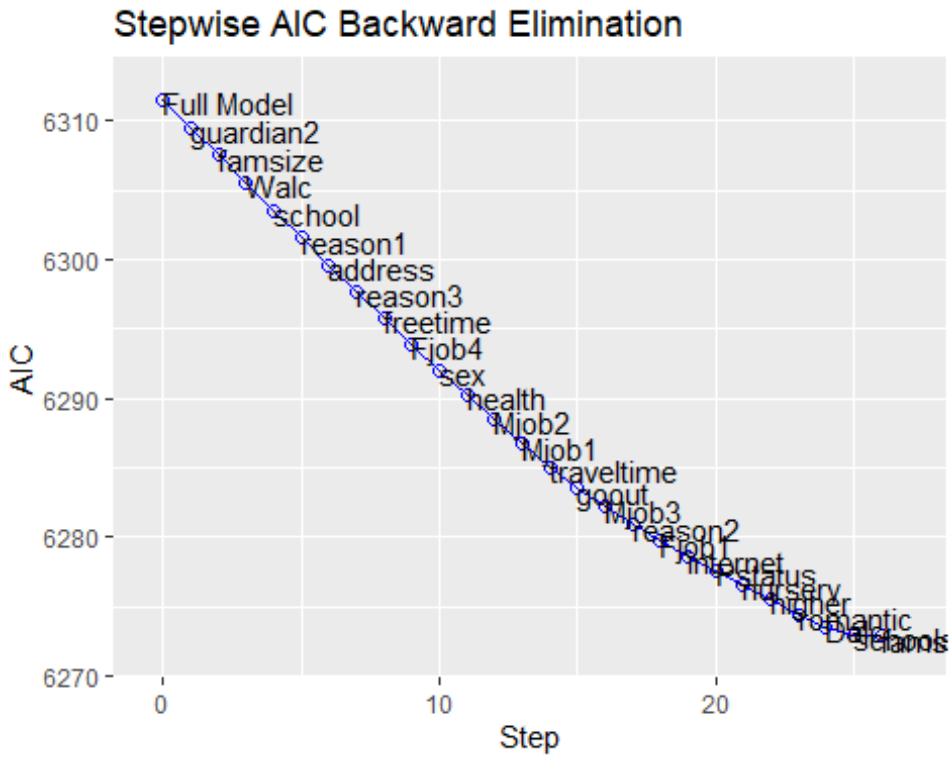
```
summary(for_aic$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2255.52  -400.44   -15.28   429.29  2056.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2255.90     555.68  -4.060 5.96e-05 ***
## G2            337.08      18.06  18.659  < 2e-16 ***
## G1            208.04      20.60  10.099  < 2e-16 ***
## age           -95.61      29.02  -3.295 0.001077 **
## famrel        136.07      38.56   3.529 0.000467 ***
## paid          197.48      70.53   2.800 0.005367 **
## Fjob3        -167.38      79.40  -2.108 0.035683 *
## guardian1     127.10      76.14   1.669 0.095885 .
## activities    129.29      69.81   1.852 0.064785 .
## Medu           51.69      33.17   1.558 0.119962
## Mjob4         129.55      80.17   1.616 0.106933
## schoolsup    -159.87     109.86  -1.455 0.146440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 681.1 on 382 degrees of freedom
## Multiple R-squared:  0.8924, Adjusted R-squared:  0.8893
## F-statistic: 287.9 on 11 and 382 DF,  p-value: < 2.2e-16
```

This model has only 11 variables and the adjusted R_squared value is 0.8892 which is nearly the same as the previous.

2.Backward Elimination

```
ba_aic = ols_step_backward_aic(MM)
plot(ba_aic)
```



So the model summary is –

```
summary(ba_aic$model)
```

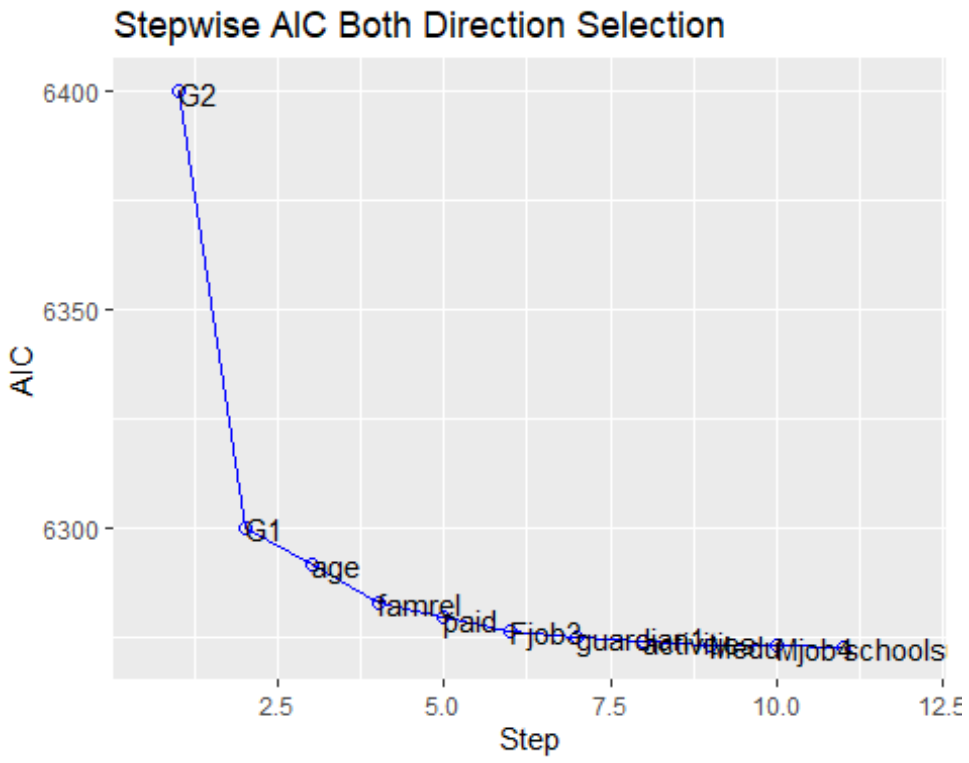
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2272.49  -424.56   -13.94   435.89  2023.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2369.965    543.438  -4.361 1.67e-05 ***
## age           -80.367     28.676  -2.803 0.005329 **
## Medu           99.036     41.962   2.360 0.018774 *
## Fedu          -62.921     41.422  -1.519 0.129587
## studytime     -59.089     42.426  -1.393 0.164506
## paid          176.714     71.390   2.475 0.013746 *
## activities    117.869     69.990   1.684 0.092986 .
## famrel        132.739     38.551   3.443 0.000639 ***
## absences       -7.334      4.912  -1.493 0.136209
## G1            218.867     20.363  10.748  < 2e-16 ***
## G2            332.471     18.011  18.459  < 2e-16 ***
## Mjob4         120.485     79.892   1.508 0.132364
```

```
## Fjob3        -158.264     79.199  -1.998 0.046396 *
## guardian1     111.872     76.923   1.454 0.146677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 679.7 on 380 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8897
## F-statistic:   245 on 13 and 380 DF,  p-value: < 2.2e-16
```

This model has only 13 variables and the adjusted R_squared value is 0.89 which is nearly the same as the previous.

3.Stepwise Selection

```
both_aic = ols_step_both_aic(MM)
plot(both_aic)
```



So the model Summary –

```
Dataframe = data.frame(De$G2,De$G1,De$age,De$famrel,De$paid,De$Fjob3,De$guard
ian1,De$activities,De$Medu,De$Mjob4,De$absences,De$Fedu,De$studytime)
summary(lm(new_Y~.,Dataframe))

##
## Call:
## lm(formula = new_Y ~ ., data = Dataframe)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2272.49  -424.56   -13.94   435.89  2023.45
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2369.965    543.438  -4.361 1.67e-05 ***
## De.G2           332.471     18.011  18.459  < 2e-16 ***
## De.G1           218.867     20.363  10.748  < 2e-16 ***
## De.age          -80.367     28.676  -2.803 0.005329 **
## De.famrel       132.739     38.551   3.443 0.000639 ***
## De.paid         176.714     71.390   2.475 0.013746 *
## De.Fjob3       -158.264     79.199  -1.998 0.046396 *
## De.guardian1    111.872     76.923   1.454 0.146677
## De.activities   117.869     69.990   1.684 0.092986 .
## De.Medu          99.036     41.962   2.360 0.018774 *
## De.Mjob4        120.485     79.892   1.508 0.132364
## De.absences      -7.334      4.912  -1.493 0.136209
## De.Fedu         -62.921     41.422  -1.519 0.129587
## De.studytime    -59.089     42.426  -1.393 0.164506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 679.7 on 380 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8897
## F-statistic:   245 on 13 and 380 DF,  p-value: < 2.2e-16
```

This model has only 13 variables and the adjusted R_squared value is 0.8897 which is nearly the same as the previous.

```
Here all the three model suggests nearly the same amount of regressor variabl
es and also approximately same adjusted R-squared value. So we can choose any
on of them safely. Since the stepwise selection is a selection procedure that
takes into account both the forward selection and the backward elemination i.
e. a combination of the above two method, we can select the final model as th
e model selected by the stepwise selction preocedure.
We can finally conclude that the student's grade can be modeled or predicted
with approximately 89% accuracy based on the variable selected by the stepwis
e selection procedure.
```