Indian Institute Of Technology, Kanpur

STAMATICS

A Project Report on

Student Alcohol Consumption Analysis using Multiple Linear Regression Model

Guided by

Sandipan Mitra & Rohan Kumar

Submitted by

Prithwijit Ghosh - 211349

Amit Meena - 211259

Abstract

In our modern days, in one side student's grades are gradually decreasing from year to year and on the other hand some bad habits among the students such as -**Alcoholism,favour in relationship** etc.. are also present among many of the students. Hence by the data-set in our project,***Student Alcohol Consumption***, we first try to see whether alcoholism,relationship play a big role in student's grade in long run. If not then among many other variables such as - **family support,family educational background,study hours,grades in previous exams**etc.. which are most important for predicting the student's grade more accurately.

For this, we first do some sort of Exploratory Data Analysis on the response variable and the remaining regressor variables. we start with the usual error assumption and fit a multiple linear regression model where the response variable is the student's final grade and the regressors are the student's alcohol consumption habit,relationship etc..If this regression is insignificant then we will choose all the variables as regressors and fit the multiple linear regression model and finally we will choose only the best subset of regressors that can best explain the response variable.

Acknowledgment

As it is rightly said that the real learning comes from a practical work.

The success and final outcome of this assignment required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along with the completion of our project work. Whatever we have done in this project is only due to the wonderful guidance and assistance of our project mentors **Sandipan Mitra & Rohan Kumar**, Society of Stamatics, IIT Kanpur for giving us this great opportunity to do the project on ***'Student Alcohol Consumption and Their Grades'*** and providing us all support and guidance which made us complete the project work on time. Without his valuable guidance and motivation, it was nearly impossible to work on this project as a team and understand the practical aspect of the topic ***"Regression Analysis***.
Last but not the least we are grateful to all the faculty members and the seniors who constantly remained in touch with us and supported us at many stages.

Yours Sincerely,
Prithwijit Ghosh - 211349
Amit Meena - 211259

Contents

List of Tables

List of Figures

1 Introduction

Drinking has negative effects on young students, their families, and their respective schools or colleges. According to an extensive research from the **NIAAA** in 2015, drinking has been prevalent among 86.4% of students ages 18 and above. The same report noted that 1,825 college students 18 to 24 years old lost their lives due to alcohol-related road accidents. Roughly 97,000 students in the same age range have been involved in sexual assaults and rape due to excessive drinking. So in between the age period 15-22 years students get destroyed by the attraction to consumption of alcohol , they are losing their real ability and efficiency and creating a bad environment for others. Obviously this bad habit also affects their academic career.
But not only alcohol consumption but also several other effects can damage the progress in the student's life silently. For an example, in student life, if a student involves in relationship, his/her academic life may be ruined. This happens mainly because in a romantic relationship many student forget their original route("Education") and go with spurious or a wrong route mainly because of immaturity and lovesick behavior in childhood.
Also their family condition e.g. parental education, richness affects affect immensely to their educational life. This thing happens mainly because there is a general tendency that the student from the lower middle class or the middle class are more accurate to their study and education than those from heavily richer class. Students from lower wealth have a stronger insight into their goals mainly because of their responsibility and maturity due to their financial background. But the students with the richer background have nothing to achieve in their life in the sense money,fame etc..So they mainly involves in some relationship, alcohol and so on.. So these factors together decide whether a student get a proper **prosperity in their life or not** in long margin or more conveniently **Statistically**.
So we were interested in analyzing the academic performance of those students who started consumption of alcohol.

2 Data-set Description

2.1 data-set Description

Our data contain full information about the relationship of students grades with their other activities e.g. alcohol consumption habit, in the student life whether they fall into relationship etc. So, obviously here our target variable is the final grade of the student i.e.
G3 → final grade,where these grades are related with the course subject, Math or Portuguese.

Now Our regressor variables are both types i.e. categorical and continuous. In categorical columns student's school,sex,address etc. are included and simultaneously in the continuous columns student's grade in the first period(G1),grade in the second period,health,workday alcohol consumption habit,weekend alcohol consumption habit etc. are included.

In our data-set there exists a tidy number of binary variables i.e. discrete variables taking only two possible values corresponding to happening or non-happening of a particular event. In this case they are namely student's extra educational support,family educational support,internet access,whether he/she fall in relationship etc.

So the complete description of our regressors and response variable is given below –

Index	Data	Type	Description
1.	School	Binary	Student's School ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2.	Sex	Binary	student's sex ('F' - female or 'M' - male)
3.	Age	Numeric	student's age (from 15 to 22)
4.	Address	Binary	student's home address type ('U' - urban or 'R' - rural)
5.	Fam Size	Binary	family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3)
6.	Pstatus	Binary	parent's cohabitation status ('T' - living together or 'A' - apart)
7.	Medu	Numeric	mother's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8.	Fedu	Numeric	father's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9.	Mjob	Numeric	mother's job('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at-home' or 'other')
10.	Fjob	Nominal	father's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at-home' or 'other')
11.	Reason	Nominal	reason to choose this school (: close to 'home', school 'reputation', 'course' preference or 'other')
12.	Guardian	Nominal	student's guardian ('mother', 'father' or 'other')
13.	Traveltime	Numeric	home to school travel time (1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour)
14.	Studytime	Numeric	weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15.	Failures	Numeric	number of past class failures (n if $1 <= n < 3$, else 4)
16.	School sup	Binary	extra educational support (yes or no)
17.	Fam sup	Binary	family educational support (yes or no)
18.	Paid	Binary	extra paid classes within the course subject (Math or Portuguese) (yes or no)
19.	Activities	Binary	extra-curricular activities (yes or no)
20.	Nursery	Binary	attended nursery school (yes or no)
21.	Higher	Binary	wants to take higher education (yes or no)
22.	Romantic	Numeric	with a romantic relationship (yes or no)
23.	Femrel	Numeric	quality of family relationships (from 1 - very bad to 5 - excellent)
24.	Freetime	Numeric	free time after school (from 1 - very low to 5 - very high)
25.	Go Out	Numeric	going out with friends (from 1 - very low to 5 - very high)
26.	Dalc	Numeric	workday alcohol consumption (from 1 - very low to 5 - very high)
27.	Walc	Numeric	weekend alcohol consumption (from 1 - very low to 5 - very high)
28.	Health	Numeric	current health status (from 1 - very bad to 5 - very good)
29.	Absences	Numeric	number of school absences (from 0 to 93)
30.	Internet	Binary	Internet access at home (yes or no)
31.	G1	numeric	first period grade (from 0 to 20)
31.	G2	numeric	second period grade (from 0 to 20)
Target	**G3**	**numeric**	**final grade (from 0 to 20) → (This is our target variable)**

Table 1: Data-set Description with explanation of all the variables

2.2 Glimpse of our Data-set

We have done a sufficient discussion based on our data-set. Now we move forward and go for a further study. at first we look into the data-set at a glance and then we pictorially observe the columns of our data-set.

 # Color data frame (class colorDF) 33 x 395:
 # (Showing rows 1 - 20 out of 395)

	D$G3	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime
1	6	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2
2	6	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1
3	10	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1
4	15	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1
5	10	GP	F	16	U	GT3	T	3	3	other	other	home	father	1
6	15	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1
7	11	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1
8	6	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2
9	19	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1
10	15	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1
11	9	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1
12	12	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3
13	14	GP	M	15	U	LE3	T	4	4	health	services	course	father	1
14	11	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2
15	16	GP	M	15	U	GT3	A	2	2	other	other	home	other	1
16	14	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1
17	14	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1
18	10	GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3
19	5	GP	M	17	U	GT3	T	3	2	services	services	course	mother	1
20	10	GP	M	16	U	LE3	T	4	3	health	other	home	father	1

Figure 1: Glimpse of our Data-set

Now let us see what are the variables in our data-set.... This time we would not draw a table and write about all the variables. In this time we would see pictorially our data-set for much better understandings.

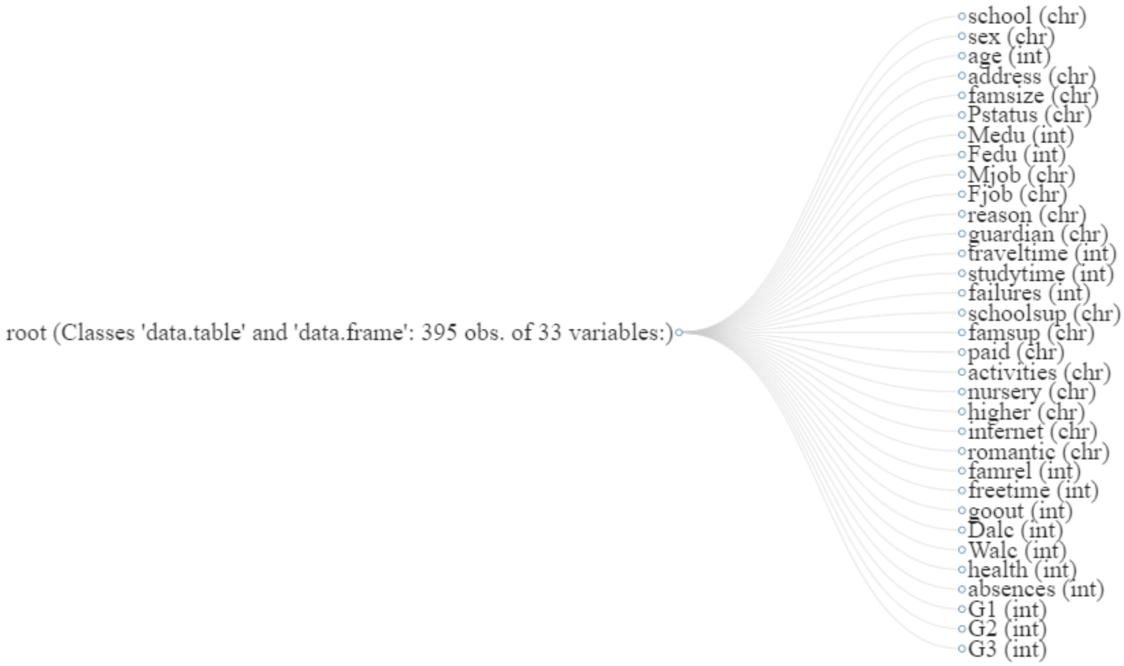


Figure 2: Glimpse of our Data-set

3 Goal of Our Study

Here Our response variable is G3 i.e. **final grade,where these grades are related with the course subject, Math or Portuguese** and we try to fit an MLR model based on the remaining variables. So the final grade of the student may be predicted based on the other given information on that particular student. In our model we follow the next steps respectively—

1. Basic Exploratory Data Analysis(**EDA**) on the variables in our data-set.
2. Encoding the categorical columns into corresponding **dummy variables**.
3. checking for **outliers** and if it exists then replace them by their suitable estimates.
4. Checking for **missing values** and replace them by their corresponding estimates.
5. Fit our Multiple Linear Regression(**MLR**) with all the variables.
6. **Residual analysis** and taking necessary actions based on that.
7. Scrutinize the model for **multicollinearity issue**.
8. Finally, select the variables which are necessary and drop all other by some variable selection methods and compare them by their **adjusted** R^2 value.

4 Methodology

We already discussed about the data-set (**i.e. Student alcohol consumption**) in somewhat subjectively. Now we will discuss about our plannings of the necessary steps from the very beginning to the bottom end. At first we will encode our categorical variables to dummy variables as necessary. Then we will check whether our data contains missing values or not. Outliers will be detected (if any) for each of the regressors in the model. This outliers will be considered as missing values and replaced by the corresponding sample estimates.Then we will fit an MLR model taking **G3** as the response variable. After that, various residual plots may be plotted and necessary actions may have been taken depending on that scenario. After fitting the model, multicollinearity issue will be detected very seriously.Then from the fitted model **Model Adequacy**(i.e. mainly R^2 and adjusted R^2) will be checked and finally using some variable selection techniques we may arrive at a handful of necessary regressors containing deliberate information from the model.

5 Exploratory Data Analysis(EDA)

For every data-set we have to perform the some amount of exploratory data analysis. At first we have to visualize the pattern of the data for necessary columns and also their interrelations via some sort of critical analysis and then some great deal of visualizing tools.

5.1 Data Standardization

If the data is the usual raw data as taken from the field, then the further statistical analysis become too much cumbersome. Ruling out this difficulty and making the usual compatibility we have standardize the data. By standardization we simply mean that for a particular variable at first the mean has to be subtracted and then the resultant has to be divided by the standard deviation of that variable for each observation.So, mathematically,

$$Standardized\ Variable = \frac{Raw\ Variable - mean}{Standard\ Deviation}$$

And our data-set after standardizing -

 # Color data frame (class colorDF) 41 x 395:
 # (Showing rows 1 - 20 out of 395)

	Y	D.age	D.Medu	D.Fedu	D.traveltime	D.studytime	D.famrel	D.freetime	D.goout	D.Dalc	D.Walc	D.health
1	-0.964	1.06	1.14	1.37	1.03	0.16	-0.15	-0.38	0.800	-0.54	-1.00	-0.40
2	-0.964	0.25	-1.60	-1.43	-0.65	0.16	1.37	-0.38	-0.098	-0.54	-1.00	-0.40
3	-0.091	-1.35	-1.60	-1.43	-0.65	0.16	-0.15	-0.38	-0.996	1.15	0.55	-0.40
4	1.001	-1.35	1.14	-0.50	-0.65	1.70	-1.67	-1.54	-0.996	-0.54	-1.00	1.04
5	-0.091	-0.55	0.23	0.44	-0.65	0.16	-0.15	-0.38	-0.996	-0.54	-0.23	1.04
6	1.001	-0.55	1.14	0.44	-0.65	0.16	1.37	0.77	-0.996	-0.54	-0.23	1.04
7	0.128	-0.55	-0.68	-0.50	-0.65	0.16	-0.15	0.77	0.800	-0.54	-1.00	-0.40
8	-0.964	0.25	1.14	1.37	1.03	0.16	-0.15	-0.38	0.800	-0.54	-1.00	-1.84
9	1.874	-1.35	0.23	-0.50	-0.65	0.16	-0.15	-1.54	-0.996	-0.54	-1.00	-1.84
10	1.001	-1.35	0.23	1.37	-0.65	0.16	1.37	1.93	-1.894	-0.54	-1.00	1.04
11	-0.309	-1.35	1.14	1.37	-0.65	0.16	-1.67	-0.38	-0.098	-0.54	-0.23	-1.12
12	0.346	-1.35	-0.68	-1.43	2.71	1.70	1.37	-1.54	-0.996	-0.54	-1.00	0.32
13	0.782	-1.35	1.14	1.37	-0.65	-1.38	-0.15	-0.38	-0.098	-0.54	0.55	1.04
14	0.128	-1.35	1.14	0.44	1.03	0.16	1.37	0.77	-0.098	-0.54	-0.23	-0.40
15	1.219	-1.35	-0.68	-0.50	-0.65	1.70	-0.15	1.93	-0.996	-0.54	-1.00	-0.40
16	0.782	-0.55	1.14	1.37	-0.65	-1.38	-0.15	0.77	0.800	-0.54	-0.23	-1.12
17	0.782	-0.55	1.14	1.37	-0.65	1.70	-1.67	-1.54	-0.098	-0.54	-0.23	-1.12
18	-0.091	-0.55	0.23	0.44	2.71	0.16	1.37	-0.38	-0.996	-0.54	-1.00	0.32
19	-1.182	0.25	0.23	-0.50	-0.65	-1.38	1.37	1.93	1.699	1.15	1.33	1.04
20	-0.091	-0.55	1.14	0.44	-0.65	-1.38	-1.67	-0.38	-0.098	-0.54	0.55	1.04

Figure 3: Data-set with standardized variables

5.2 EDA on response variable

Here our response or target variable is **G3** and it is a continuous variable.

So, we can plot the histogram because in any MLR model the response variable should have a normal distribution for a smooth propagation of the classical linear model theory.

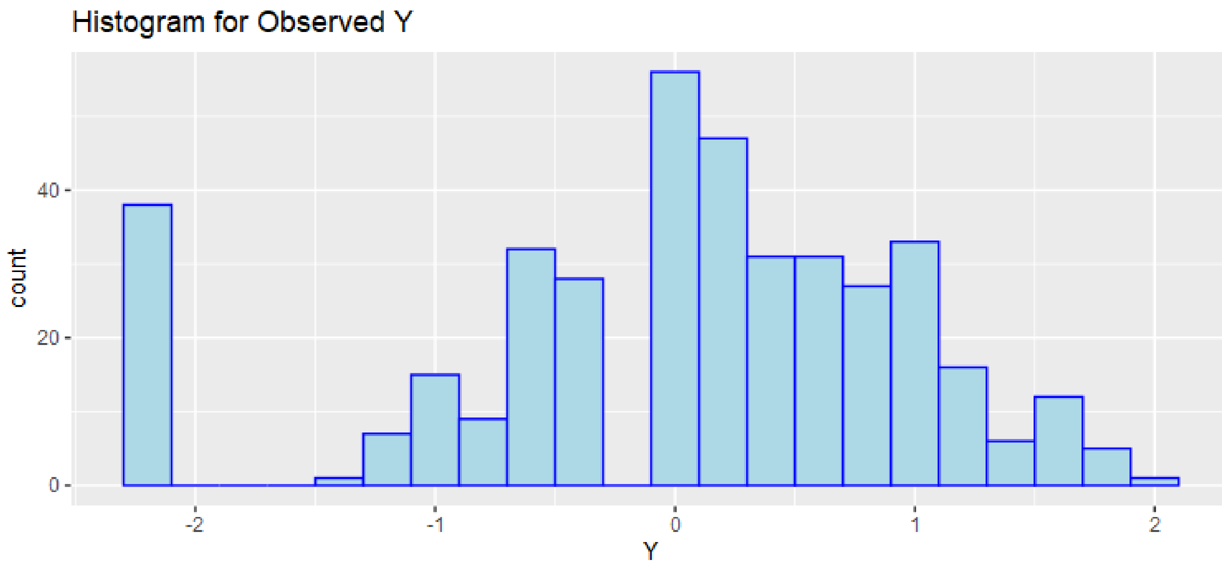


Figure 4: Histogram for the observed response variable G3

Here observing the histogram, it seems reasonable to assume that the distribution of the standardized response variable is nearly normal but only at the begging it takes a high jump.

5.3 EDA on remaining regressor variable

Let we will focus on the remaining regressor variable by simple bar-plot,histogram etc..since the regressor variables are treated as constant in MLR model so we only devote a small amount of EDA on them and we will move into our main model fitting purpose.

Box-plot

We first try to observe the box-plots for different regressor variables all taking together. Because box-plot gives 5-statistics summary not only in a neat and clean pictorial way but also shows the outlier(if any) by the dots. So looking all the regressors at one glance is best fitted by the box-plot. The box-plot for our data set is the following –

Now the above box-plot clearly shows that the variables in our data-set are largely affected by the outliers and we will consider this part later.

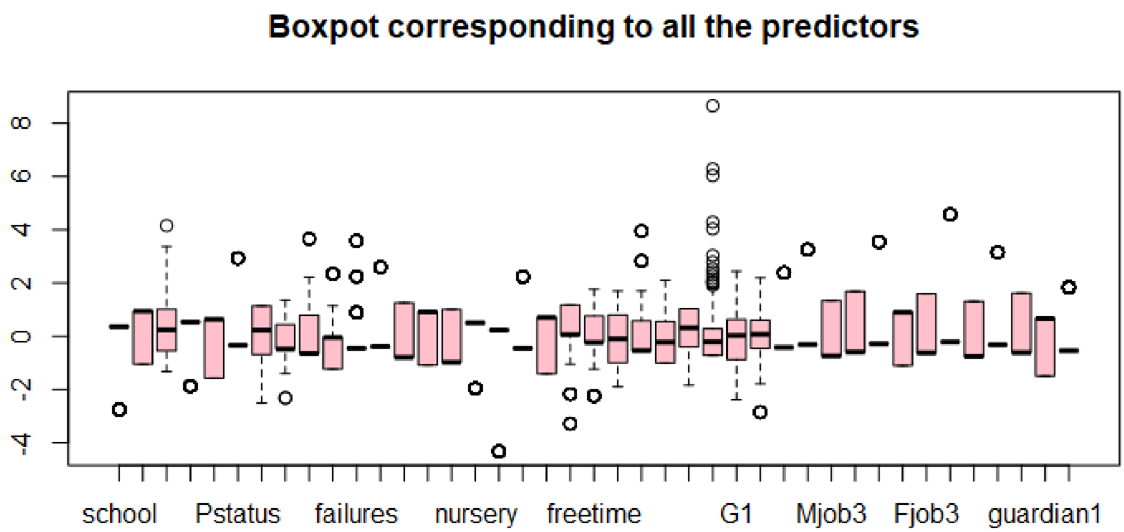


Figure 5: Box-plot corresponding to the remaining regressor variables from the data-set

Histogram and Bar-plot

Box-plot says overall all the things except the fact the distribution of the of that variable can not be explained by a simple box-plot.So, we have to take care of this fact and hence for the discrete case we use the bar-plot and for the continuous case we use the histogram. for our data-set, those visualizations are described below -

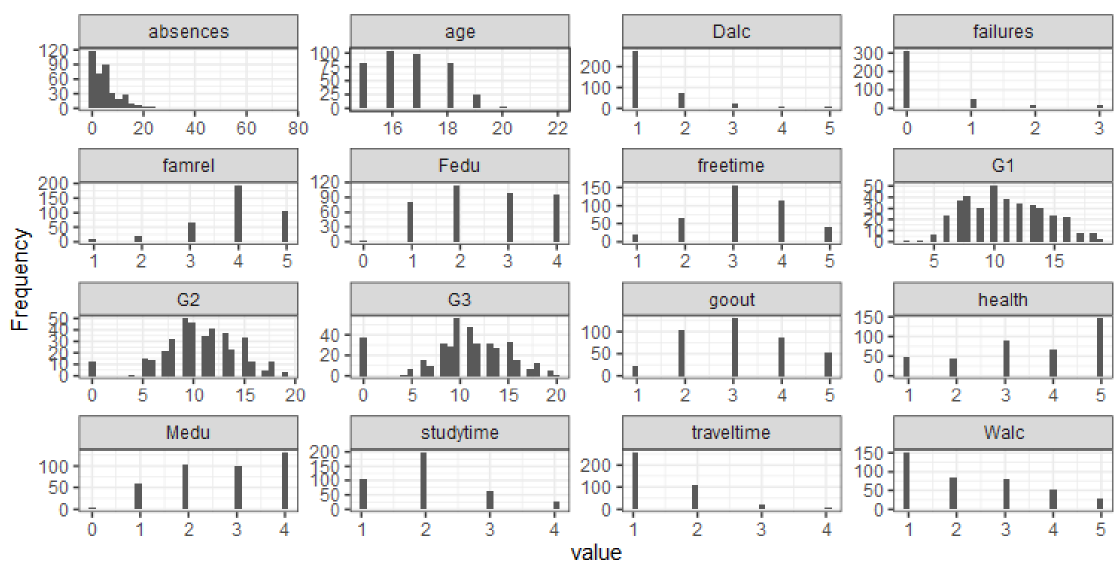


Figure 6: Box-plot corresponding to all the variables from the data-set

Now from the histograms we can easily say that -

absences → positively skewed distribution.
G1 → symmetric distribution.
G2→ symmetric distribution.
G3→ symmetric distribution.
All others → Difficult to interpret due small range of the variables.

Let us now look into the binary and categorical variables by their corresponding horizontal bar-plot.

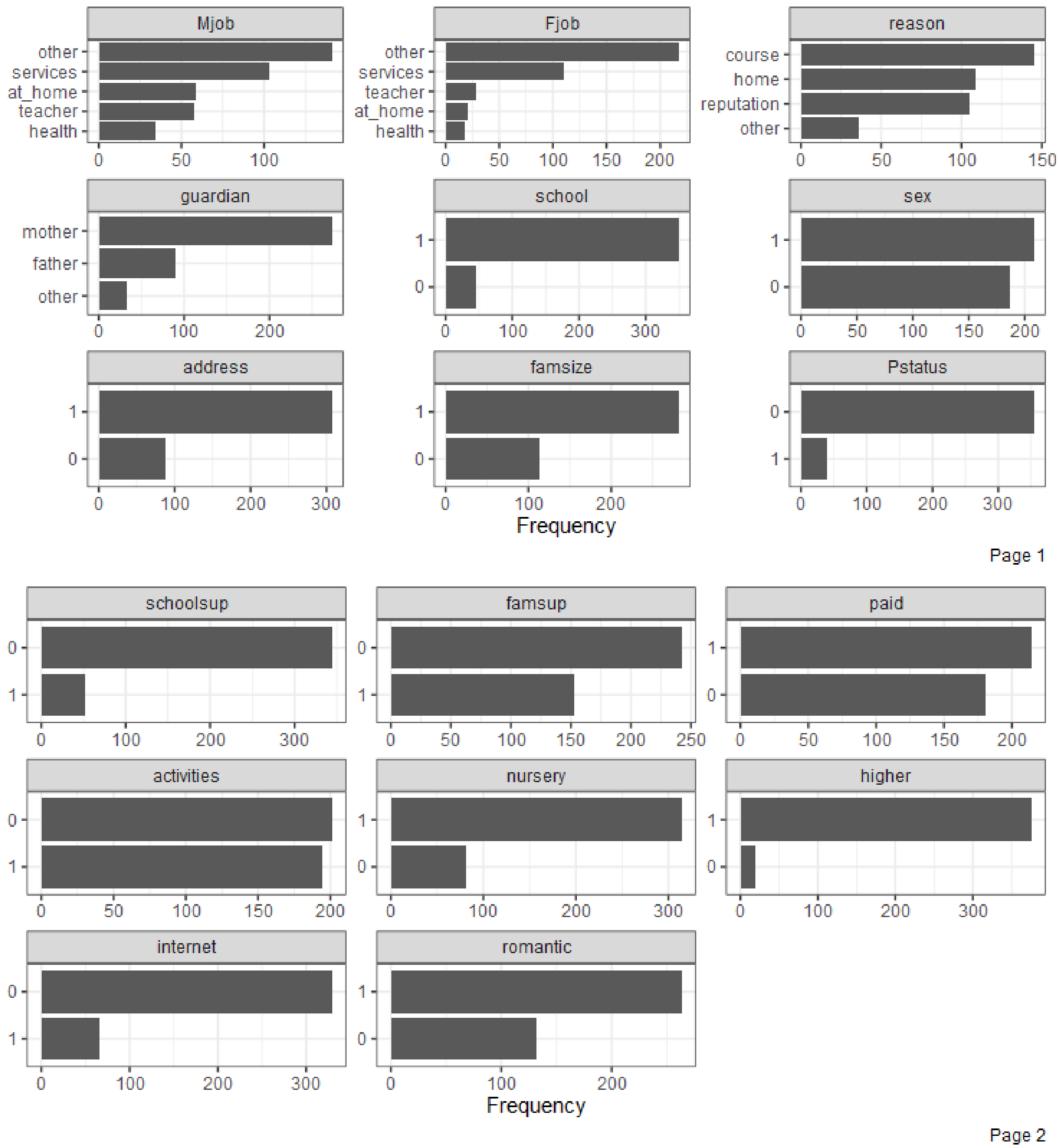


Figure 7: bar-plot corresponding to the binary variables

Let us now discuss about the categorical and the binary variables separately.

- **categorical Variables**
  Mjob & reason → Each Category fills with a sufficient number of observations.
  Fjob & guardian → Only the first two category fills with a large number of observation and all the remaing categories contains only a small.

- **Binary Category**
  School,Pstatus,Schoolsup,higher,nursery,internet → occurring of the particular event happens too much than the non-occurring.
  famsize,address,famsup,roamntic → occurring of the particular event happens better than the non-occurring.
  sex,paid,activities → occurring of the particular event and non-occurring of that event are nearly equal.

Correlation Heat-map

Now for observing the interrelation among the variables, here we mainly use the correlation matrix. But the correlation matrix is very difficult for visualization.

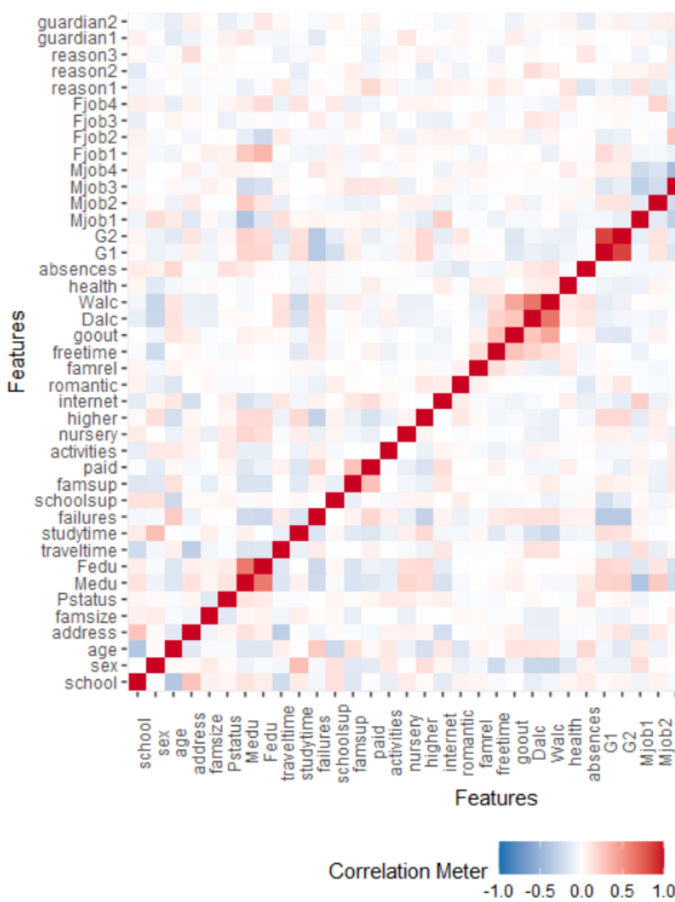


Figure 8: Correlation among all the variables

Hence here we goes into the correlation heat-map where the gradient in the color interpret the higher or lower correlation among the variables. So below is the correlation heat-map for our data-set.

from the above colored correlation heat-map,reddish color indicates that the correlation is positive and the bluish color indicates that the the correlation is negative and the the zero correlation is displayed by the whitish color.

6 Data Encoding

Data encoding is a principle step for every analysis of categorical data. Here, every categorical columns have to be encoded i.e. turned them into numerical values appropriately.
During encoding -

1. We first have to keep in mind that the design matrix must not be singular.
2. There is no restriction on the model.
3. Ordinality have to be explained properly.

Now we have four categorical columns and each category is a nominal category. Mow if we want to encode this category as the ordinal number e.g. 1,2,3,4... then a restriction on the model parameter is superimposed and we have to do restricted optimization for selecting the model parametr in our model, which is unnecessarily complicated. Hence here we use the usual "one hot encoding" technique to encode our categorical columns into binary variables.

we have four categorical variables and we have to convert them into the binary one's. These categorical columns are - **Mjob,Fjob,reason& guardian** Total number of categories in the Mjob and the Fjob columns are 5 each,reason has 4 categories and finally guardian has only 3 categories. So as by the basic rule of encoding we can encode it into 4,4,3,2 numerical variables that takes on the values 0 or 1 respectively.The categories for these four columns are – **Mjob→"at_home" , "health" , "other" , "services", "teacher"**
Fjob→"at_home" , "health" , "other" , "services", "teacher"
reason→"course" , "other" , "home", "reputation"
guardian→"father","mother","other"

We are ready to encode this categorical columns in to numerical ones. Now the following table gives the summarization of what actually we want to say–

Table 2: **Encoded Categorical variables Column: Mjob and Fjob**

	Encoded Categories : Mjob					Encoded Categories : Fjob			
Raw Categories	Mjob1	Mjob2	Mjob3	Mjob4	Raw Categories	Fjob1	Fjob2	Fjob3	Fjob4
at_home	1	0	0	0	at_home	1	0	0	0
health	0	1	0	0	health	0	1	0	0
other	0	0	1	0	other	0	0	1	0
services	0	0	0	1	services	0	0	0	1
teacher	0	0	0	0	teacher	0	0	0	0

Table 3: **Encoded Categorical Variables Column: Reason and Guardian**

Raw Categories	Encoded Categories : reason		
	reason1	reason2	reason3
course	1	0	0
other	0	1	0
home	0	0	1
reputation	0	0	0

Raw Categories	Encoded Categories : guardian	
	guardian1	guardian2
father	1	0
mother	0	1
other	0	0

In the above four table we have finally converted the categorical columns into the binary dummy variables. Now we proceed with our further analysis.

7 Outlier Detection

Outlier in the data plays a very important role when we came into the estimation purpose for the model parameters.So we have to remove or estimate the outliers based on the given scenario.

For our model let us draw the box plots to see whether heavy amount of outliers present in our data or not for only the continuous columns. Because if we remove the outliers from the binary variables then it is not only meaningless but also there exists a situation where we can all the values corresponding to the lower class is removed and for further prosperity we have to drop the columns. If the number of outliers are quite small we can simply delete the corresponding rows, otherwise we proceed into further analysis.

Let us first see that the box-plots of the corresponding columns–

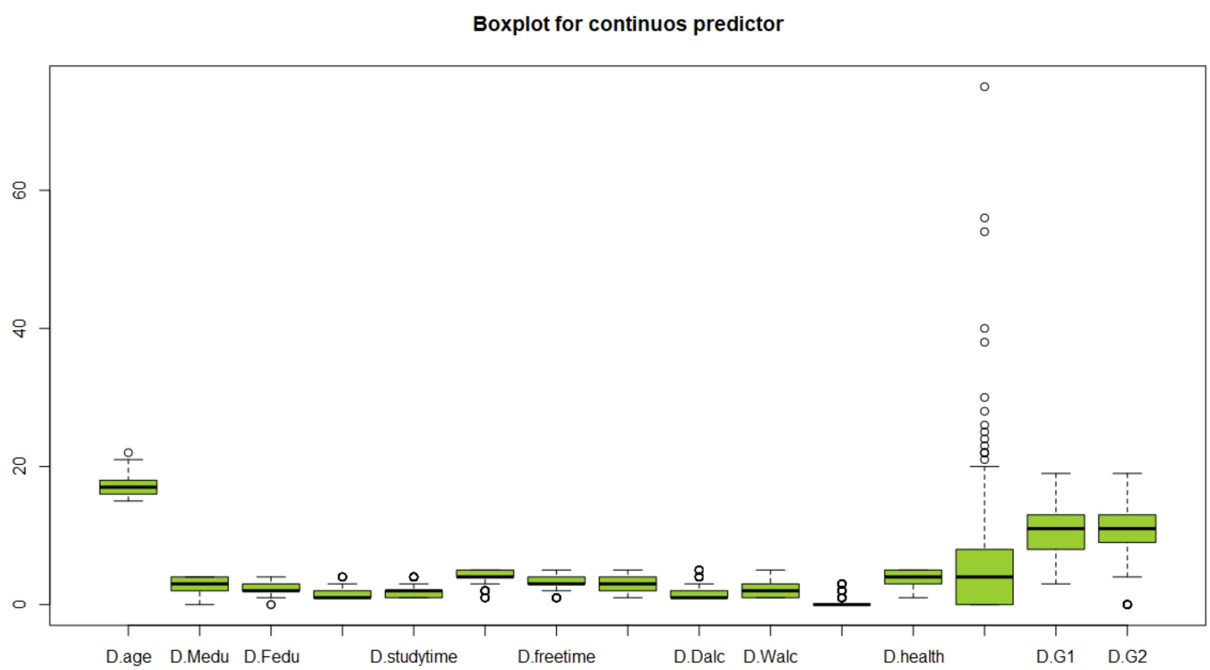


Figure 9: Box-plot for detecting outliers in the present data

Since too much outliers are present in the data, we have to estimate them. Here we apply Interquartile Range method (IQR) for outlier detection and removal of outliers. Observing the boxplot, it is almost sure that the variables are skewed. So, IQR method suits well for such kind of data.

Let, X denotes the variable
Then the IQR =(q_3(X)-q_1(X)) Where, q_3(X)= 3rd quartile ; q_1(X)=1st quartile
Now, let us define,
lower(X)= q_1(X) − IQR (X)*1.5
Upper(X)= q_3(X)+IQR (X)*1.5
If any values of X, say i-th value of X, X[i], say Now
That is our general procedure

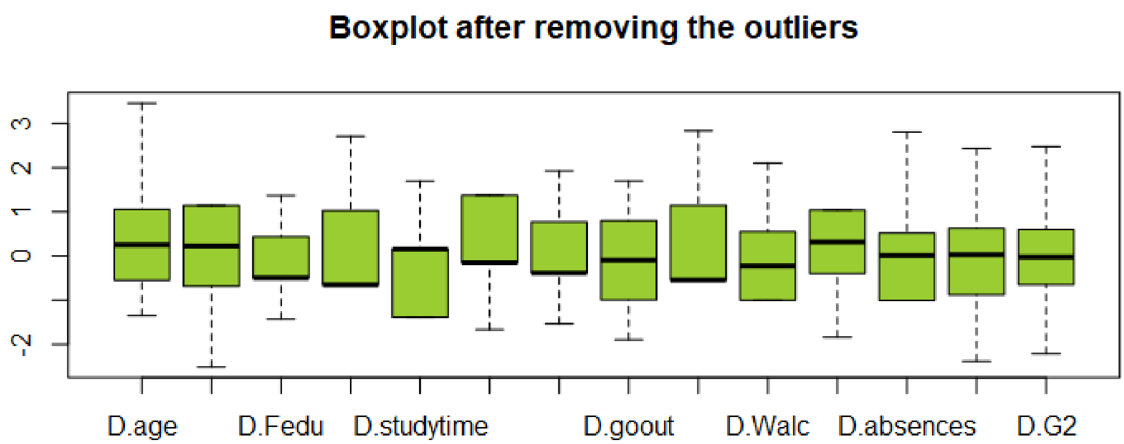


Figure 10: Box-plot after removing the all the outliers

In this way we can estimate all the outliers present in the data and hence we can say that, now our data is quite structured than the previous one.

8 Linear regression on alcohol consumption

Now we are begin with our preliminary consideration that whether student's final grade(G3) is affected by alcohol consumption, relationship etc.. So we start with our basic linear model assumption i.e. -

$$y = X\alpha \ + \ \delta$$

where

$$\delta \sim some\ distribution\ with\ E(\delta) = 0\ \&\ var(\delta) = \sigma^2 \tag{1}$$

Here y is our response variable and X is the design matrix and α is the parameter in the model what we have estimate.
Dalc and Walc provides the alcohol consumption data also the relationship represents by romantic, so the columns of X are -
First Column $\rightarrow$ Columns of all 1's
Second Column $\rightarrow$ Dalc
Third Column $\rightarrow$ Walc
Fourth Column $\rightarrow$ Romantic
So let us fit fit our MLR model with three regressors Dalc,Wlac and romantic.

```
> 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
```

Figure 11: Model summary corresponding to the three regressor variable

From the summary of the above model we can see that the R^2 value is 0.02472 and the *adjusted* R^2 value is 0.01723, which are very small to conclude any thing about the model. So a natural question arises that whether the regression is valid or not i.e. we have to test the hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \qquad against \qquad H_1 : not\ H_0$$

For testing this hypothesis our test statistic is –

$$F_0 = \frac{(Sum\ of\ square\ due\ to\ regression)/(degrees\ of\ freedom = 3)}{(Sum\ of\ square\ due\ to\ error)/(degrees\ of\ freedom = 391)}$$

Here we assume that additionally,

$$\epsilon_i \sim independently\ Normal\ distribution(0, \sigma^2) \quad \forall i = 1(1)395$$

and

From the above figure we see that the value of the F-statistic is 3.303 and the corresponding p-value is 0.02037,which is greater than 0.01 and hence we can finally fail to reject H_0 at 1% level of significance. So the values of all the parameter associated with the regressors are exactly equal to zero at 1% level of significance. Hence we finally conclude that the regression is not valid on the present scenario or equivalently Dalc, Walc and romantic does not affect significantly on the student's final grade.

9 Full model Building

Since from the previous discussion we see that the alcohol consumption has statistically no effect or insignificant for predicting anything about the final grade of the student. Hence we fit the full linear regression model i.e. the model with all the columns(except the column "failure", since it is a null column) as regressors and G3 as the response variable. So more mathematically,

$$Y = X\beta + \epsilon$$

where

$$\epsilon \sim \ some\ distribution\ with\ E(\epsilon) = 0\ \&\ var(\epsilon) = \sigma^2 \tag{2}$$

Here $\beta = (\beta_0, \beta_1, ..., \beta_{40})^\intercal$ and X is the design matrix with 1's in te first column and the remaining columns as the other regressor variables.
At first we try to find the least square estimate of the parameter vector β, and by the basic theory of the least square estimation procedure the parameter estimate is -

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

So let us put the values of the estimated $\beta's$ and also see the model summary simultaneously.

```
> 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
```

Figure 12: The full moedl and the corresponding summary

In the above picture the first column indicates the parameter estimates,second columns indicates the standard error of the estimate,third column indicates the corresponding t-value and the final column indicates the p-value corresponding to th t-value.
As before we first check whether the regression is valid or not i.e.

$$H_0 : \beta_1 = \beta_2 = ...\beta_{40} = 0 \qquad against \qquad H_1 : not\ H_0$$

As above explained —
For testing this hypothesis our test statistic is –

$$F_0 = \frac{(Sum\ of\ square\ due\ to\ regression)/(degrees\ of\ freedom = 40)}{(Sum\ of\ square\ due\ to\ error)/(degrees\ of\ freedom = 354)}$$

Here additionally we assume that,

$$\epsilon_i \sim independent\ Normal\ distribution(0, \sigma^2) \quad \forall i = 1(1)395$$

From the above figure we see that the value of the F-statistic is 26.98 and the corresponding p-value is 2.2e-16,which is less than 0.01 and hence we can finally reject H_0 at 1% level of significance. We can conclude that the regression is valid for the given scenario and we can proceed with our final analysis.

9.1 R^2 value

For checking the the model adequacy, we can proceed as following.
Since in regression we actually predict some response variable by some other regressor variables. But actually the variation in the response variable is tried to explain by the regressor variables through the regression technique and this variation is measured by the corresponding sum of square. If the

model fits well then the sum of square due to regression has to be large and the sum of square error is small. So, a general measure of model adequacy is –

$$R^2 = \frac{Sum\ of\ square\ due\ to\ regression}{Sum\ of\ square\ due\ to\ error}$$

Whose value lies between 0 and 1, where near 0 value represents the worse model fitting and near 1 value represents the better model fitting.
For our case it is 0.753 which nor too bad but the model is also not to good hence we have to think better to fit the model better then the present.

9.2 Adjusted R^2 value

R^2 value represents the model adequacy well but it has a serious problem that if we increase the number of regressor then the R^2 value is also increases. It is a big drawback for any indicator which indicates the model adequacy because we can simply add too many regressors in our model and finally get a very good R^2 value quite easily. Hence a remedy for solving this issue is defining Adjusted R^2 which is same thing as R^2 except the fact that it is adjusted by the corresponding degrees of freedom of the sum of squares. So mathematically,

$$R^2 = \frac{(Sum\ of\ square\ due\ to\ regression)/(Degrees\ of\ freedom\ for\ regression)}{(Sum\ of\ square\ due\ to\ error)/(Degrees\ of\ freedom\ for\ error)}$$

It's value always less than the R^2 value but it has not any lower bound.
For our case it is 0.7251 which nor too bad but the model is also not to good hence we have to think better to fit the model better then the present.

9.3 Residual Analysis

Residuals are the observed errors i.e. difference between observed response variable and predicted response variable. Now the model part i.e. the linear part of the regressor variable is supposed to be independent of the ϵ (error component). Now error is unobserved, so it is best explained by the residuals of the model. Also, errors are linearly independent from the predicted response variable. Hence, if we plot the predicted response and the residuals then the points must be scattered within a horizontal strip around the origin. For our model –

Figure 13: The residual plot corresponding to the full model

It is difficult to interpret anything about this residual plot so, we go into some further details.

9.4 Homoscadasticity Checking

In the basic linear model assumption we assume that the distribution of error is homoscedastic but now we have to justify our assumption.Here we generally test the hypothesis that whether the distribution has a constant variance or not and if the p-value is less than 0.01 we reject our assumption of homoscedasticity at 1% level of significance.

```
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
```

Figure 14: The Breusch Pagan test for testing the homoscedasticity assumption

From the above hypothesis it is readily obvious that the given sample is not from the distribution with constant, since the p-value corresponding to each test is less than 0.01. So we can safely reject the null hypothesis i.e. the distribution has a constant variance.

9.5 Normality Assumption

For any further propagation it is necessary to assume the normality of the error as assumed previously. But for any further assumption we have to check it based on our given data.

Graphical Checking

Graphically we can easily check the normality assumption by the Q-Q plot where the ranked residuals are placed in the X-axis and the sample quantiles are placed in the Y-axis. If the middle 60-80% of the data must be in a straight line, we can easily conclude that the underlying distribution is normal. For our model the Q-Q plot is the following –

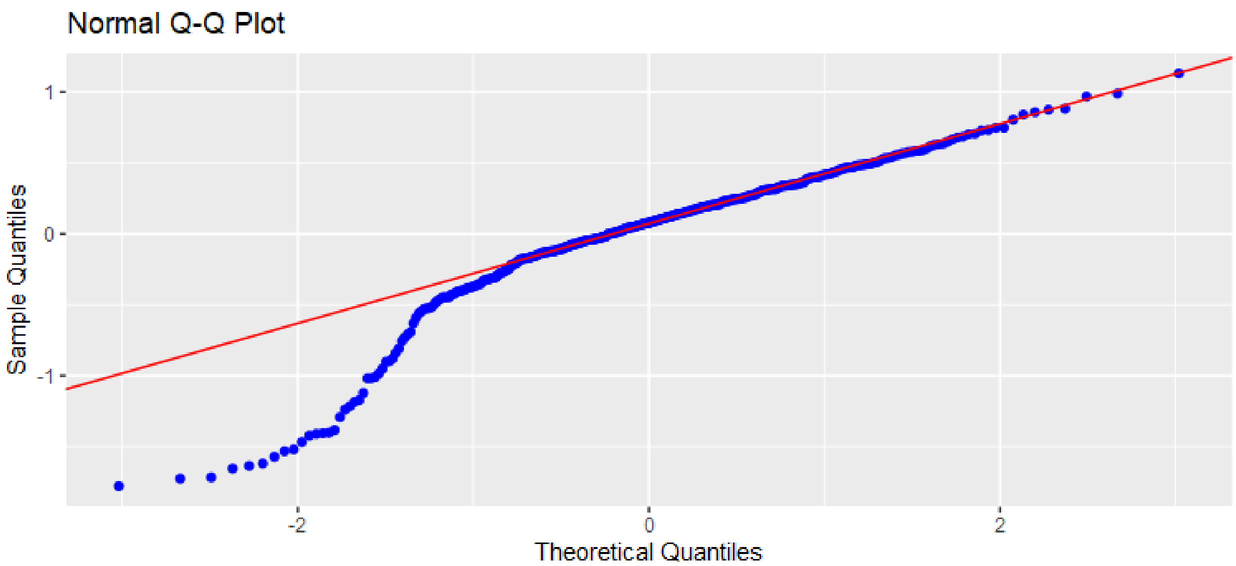


Figure 15: The Q-Q plot for normality checking

Since nearly the beginning, approximately 25% observations falls bellow the normality line we can't safely assume the normality assumption for our model.

Theoretical Checking

Form the graphical method we can't conclude anything about normality, here we go for the theoretical checking. For theoretical purpose we actually test the hypothesis whether the given sample is from normal distribution or not. So for our test –

```
> 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
-----------------------------------------------
```

Figure 16: Theoretical value for testing of normality assumption

From the above hypothesis it is readily obvious that the given sample is not from the normal distribution, since the p-value corresponding to each test is less than 0.01. So we can safely reject the null hypothesis i.e. the distribution is normal.

10 Transformed Model

Since our original model violets all the assumption that we assume in our theoretical consideration,so it is necessary to transform our model. There are two possible transformation one is Box-Cox and another is sin hyperbolic inverse. Since in the Box-Cox transformation we need to identify the variable or variables causing the heteroskadasticity, so we choose the sin hyperbolic inverse transformation.Now a particular sin hyperbolic inverse transformation is not reliable so we transform Y as $sinh^{-1}(Y)^x$, where x is unknown. We will change x and choose the model with maximum *adjusted* R^2 value and the best fitted Q-Q plot simultaneously.

Maximum AdjustedR^2 and Corresponding Q-Q plot

Let us see two plots corresponding to the maximum $AdjustedR^2$ and the corresponding Q-Q plot.

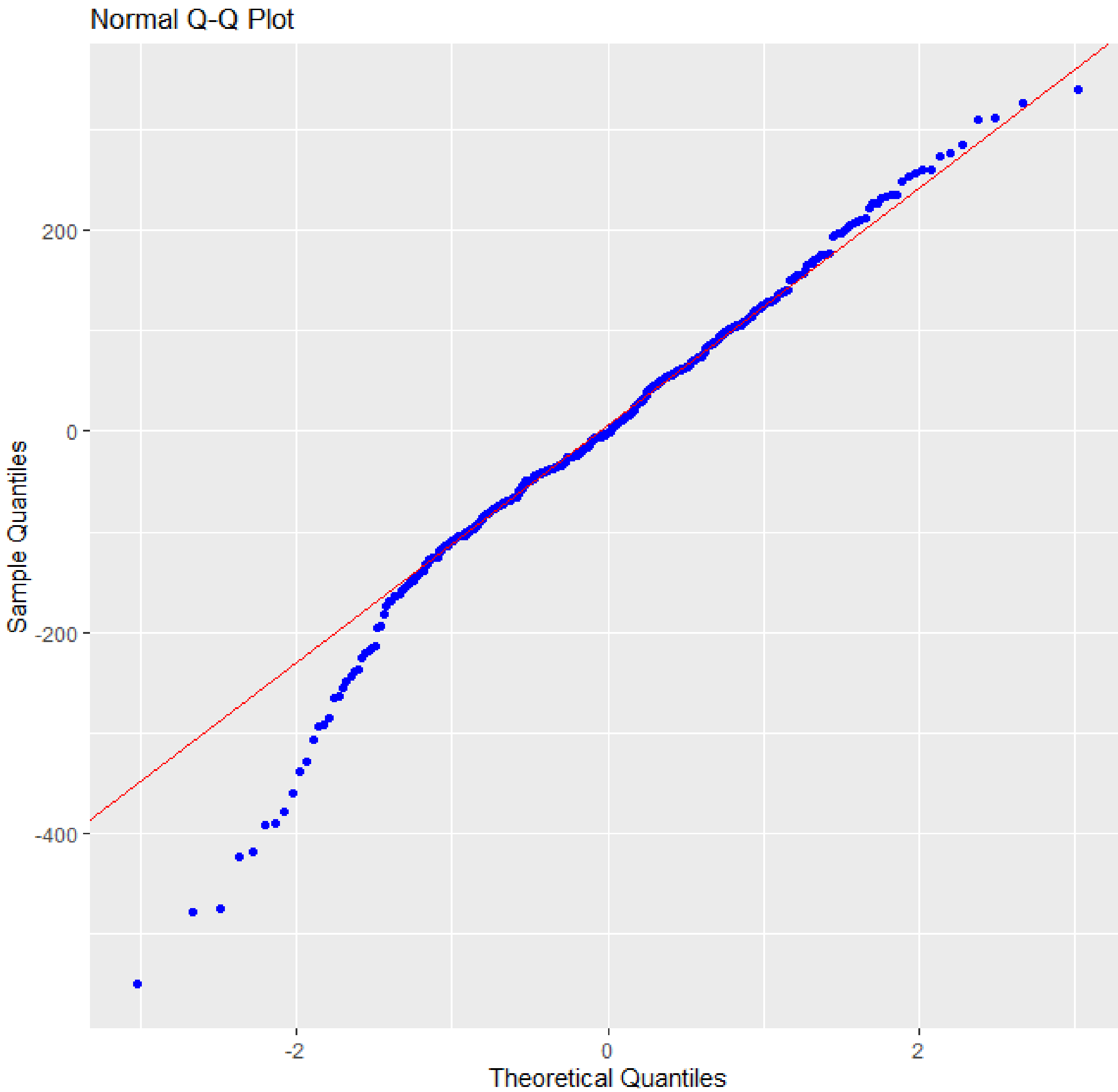


Figure 17: Normal Q-Q plot

This is the model with maximum adjusted R^2 but the Q-Q pot suggest that the normality assumption is till not valid. Now we plot the change of adjustedR^2 corresponding to the values of x.

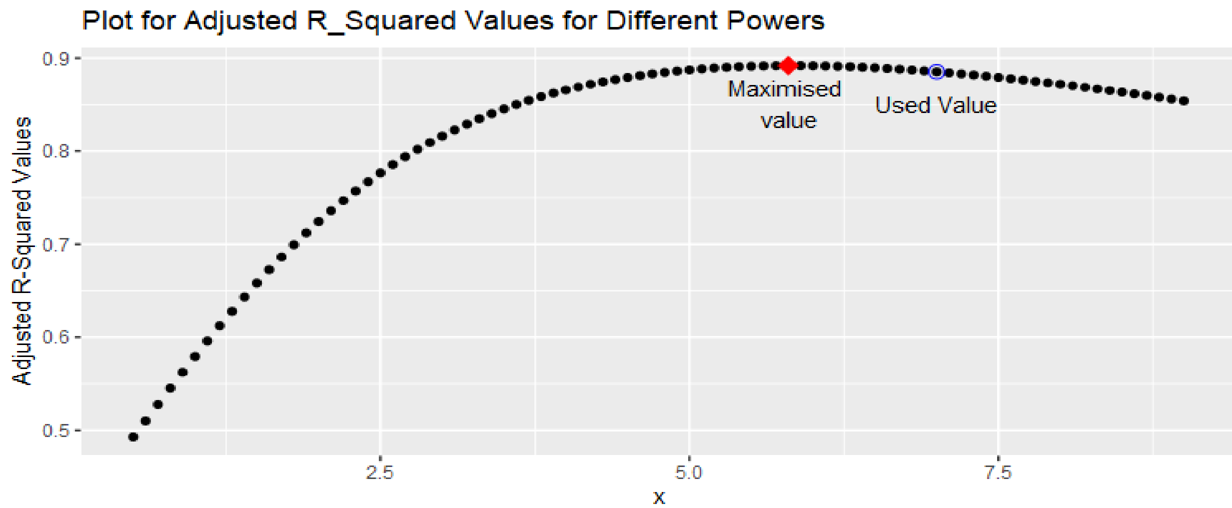


Figure 18: Normal Q-Q plot

In this plot the **red point** indicates the point with maximum adjusted R^2 and the **blue circle** indicates that the point with better Q-Q plot.Since the surface of maxima is flat so we can choose the value with **blue circle** and the value of x at that position is 7.0.
Let us now see the Q-Q plot corresponding to the transformed model –

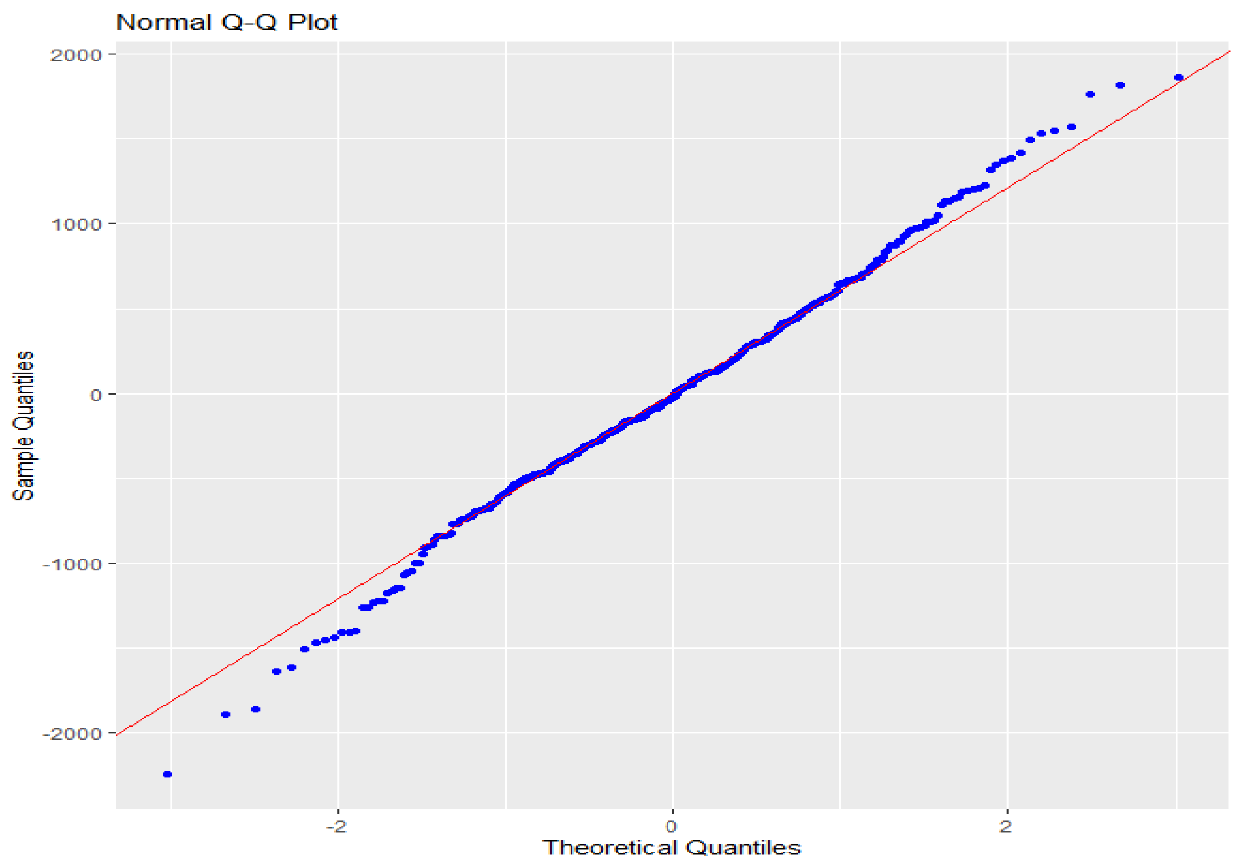


Figure 19: Normal Q-Q plot

Now the Q-Q plot is far more better then the previous one.Now we have to check the theoretical values for normality and homoskadasticity.

10.1 Homoskadasticity Checking for Transformed Model

As before for theoretical purpose we actually test the hypothesis whether the given sample is from normal distribution or not. So for our test –

```
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
```

Figure 20: The Breusch Pagan test for testing the homoscedasticity assumption

Since the p-value is greater than 0.05, so we fail to reject the null hypothesis and the our assumption of homoskadasticity is validated.

10.2 Normality Assumption

Since from the graphical method we can't conclude anything about normality, here we go for the theoretical checking. For theoretical purpose we actually test the hypothesis whether the given sample is from normal distribution or not. So for our test –

```
> 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
-----------------------------------------------
```

Figure 21: Theoretical value for testing of normality assumption

Now the above table of test statistic and their corresponding p-value suggests that the normality assumption is not violated, since the p-value corresponding to three tests out of four suggests that the sample is from normal distribution as the values are greater than 0.05. So we finally securely say that the error distribution is normal.

Now all of our basic assumptions are satisfied, so finally we try to see the model summary and residual analysis of our final transformed model.

10.3 Transformed Model Summary and Residual Analysis

Summary

For the final transformed model,let us summarize all the information. Here our main target is to see what is the adjustedR^2 for this model.

```
> 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
```

Figure 22: Transformed model summary

From the above summary it is readily seen that the test of regression is significant implies that the the regression is valid.Now the R^2 value is 0.897 and the adjustedR^2 value is 0.8854, which are too well than the previous and we can sufficiently be satisfied with those values.

Residual Plots

By different residual plot we actually try to see whether our model looks fine or it again includes some of the difficulty. Fo this we first plot the R studentized residuals or the deleted studentized residuals against the predicted response variable. And the plot is tyhe following –

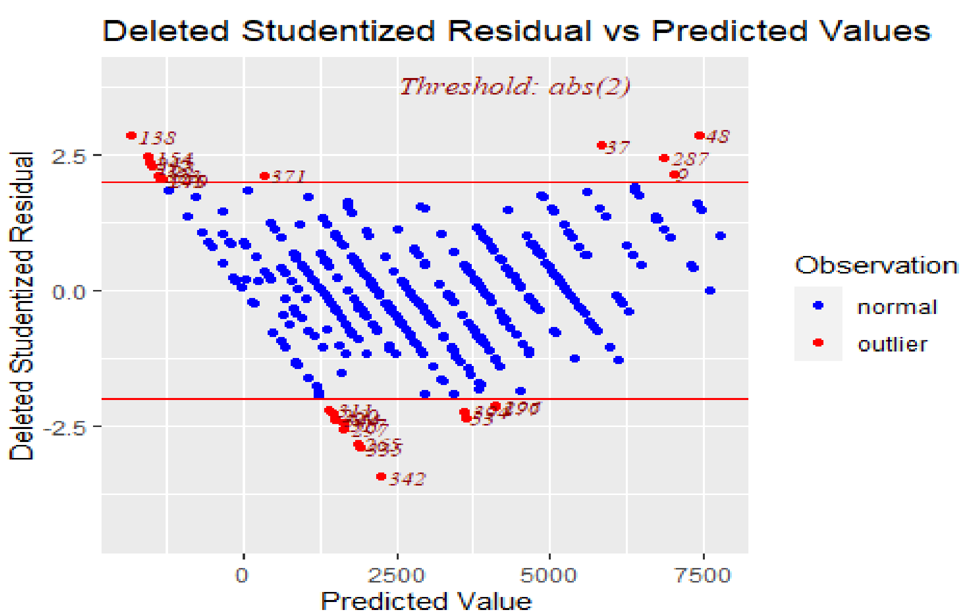


Figure 23: Threshold plot for detecting the residuals

The observations lying outside the threshold line are the outliers and since they are small in compare to the data-set so we will ignore the outliers for our model.

Leverage points

In statistics and in particular in regression analysis, leverage is a measure of how far away the independent variable values of an observation are from those of the other observations. High-leverage points, if any, are outliers with respect to the independent variables. That is, high-leverage points have no neighboring points in R^p space, where p is the number of independent variables in a regression model. This makes the fitted model likely to pass close to a high leverage observation.[1] Hence high-leverage points have the potential to cause large changes in the parameter estimates when they are deleted i.e., to be influential points. Although an influential point will typically have high leverage, a high leverage point is not necessarily an influential point. The leverage is typically defined as the diagonal elements of the hat matrix. Let us now see the residual and the leverage plot for our data-set –

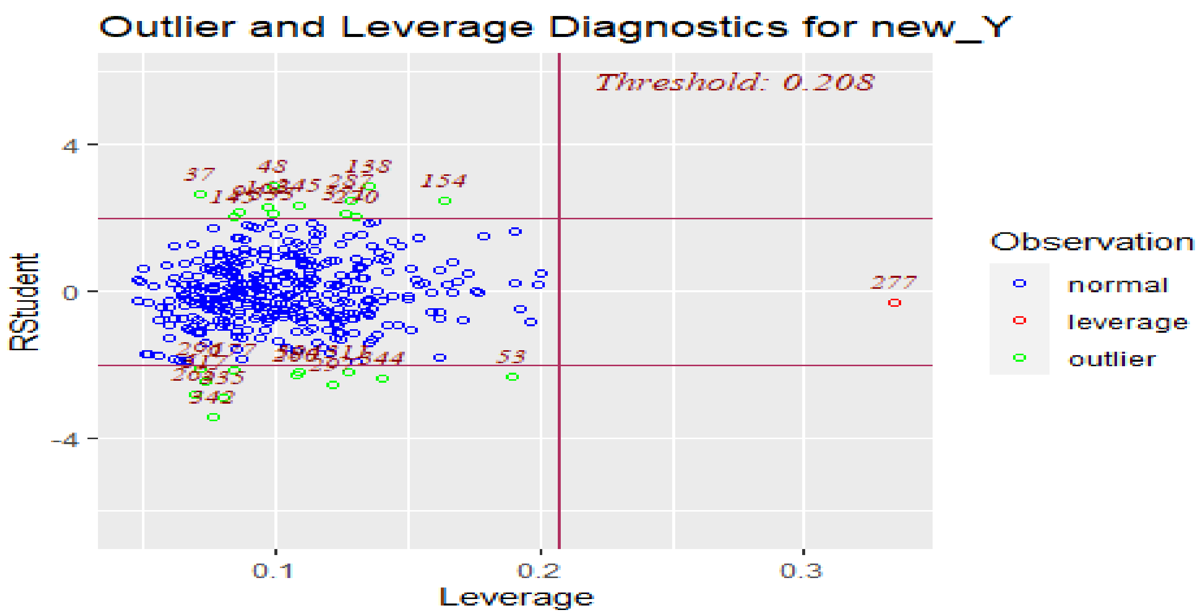


Figure 24: Threshold plot for detecting the outliers and the leverage points

The only red point which is far apart from the the usual residuals is the only leverage point for our data-set and we have to remove this point for better fitting out model.

Removing Leverage point

Let us now remove the leverage point and fit our model again to see whether there is an significant change in our model parameter estimation or other theoretical calculation purpose.

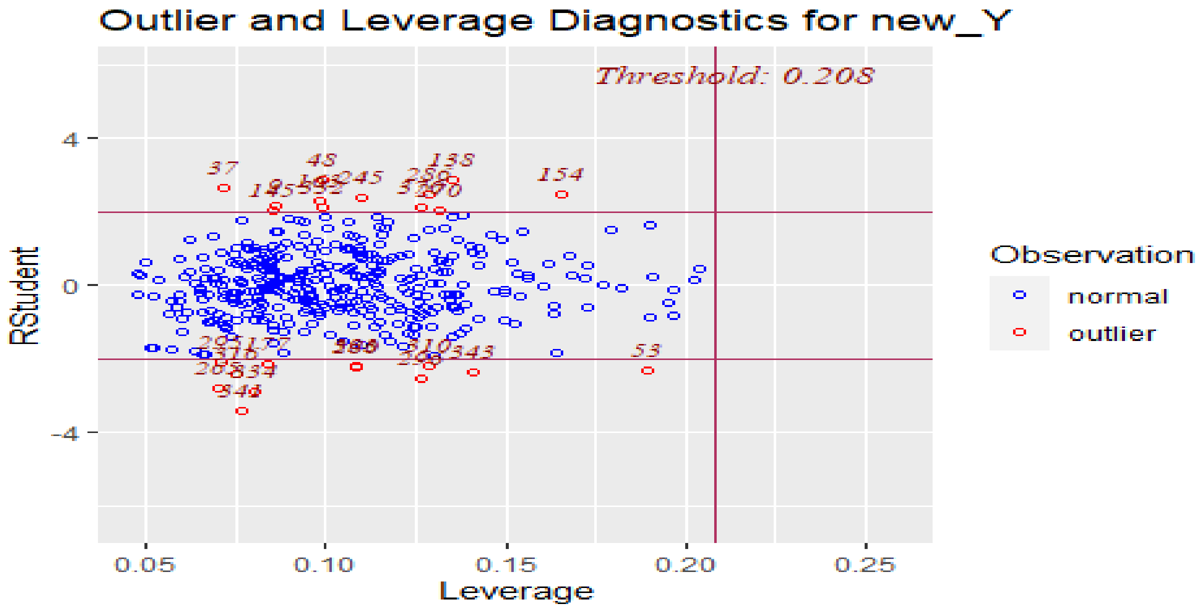


Figure 25: Leverage and residual plot fter removing the leverage point

Model Summary

After removing the leverage point let us see the fitted model summary and the corresponding measurement based on that 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
```

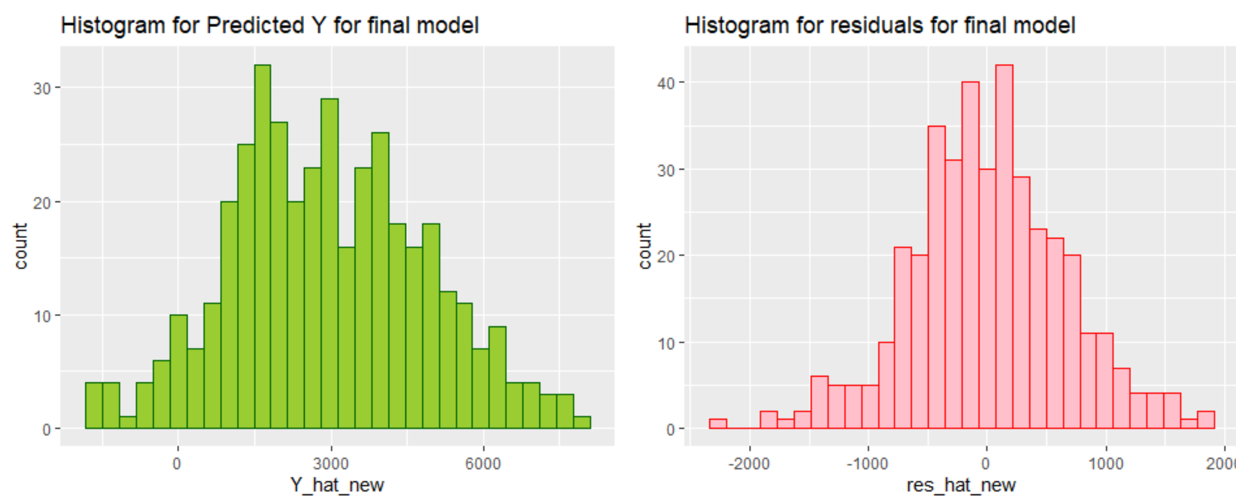
```
famrel        128.172     41.562   3.084   0.0022 **
freetime      -12.854     40.193  -0.320   0.7493
goout          24.015     38.380   0.626   0.5319
Dalc          -59.106     55.836  -1.059   0.2905
Walc           -5.339     42.112  -0.127   0.8992
health        -12.203     27.369  -0.446   0.6560
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
```

Figure 26: Model summary after removing the leverage point

Since the R^2 and the adjustedR^2 value and also the model parameters are approximately same as that of the previous model, hence we can say that there is not a significant effect of th leverage point in our model.

Now let us plot the residuals and the predicted response variable by the histogram to see the ultimate behavior of our final transformed model.



(a) Histogram of the fitted response after transforming the model

(b) Histogram of the residuals after transforming the model

Figure 27: Histogram : Predicted response and the residuals

11 Multicollinearity

Multicollinearity is a silent killer of any regression model. If the multicollinearity is present in the data then all the coefficient magnitudes becomes very large, but the predicted value of the response variables may be highly satisfactory. But if we delete one row or one column then the model change drastically, the negative coefficient becomes positive, the largest coefficient may be smaller in the newer model and finally the prediction may be reversed the previous one.

Let us consider an example: based on our present data we fit a model and estimate response variable, i.e. we give a mathematical logic for predicting some near future observations. The model gives well data fit, all the model accuracy say that this model is a good one. But after one or two week we again collect some data and add to the data-set and the model changes drastically. If may possible that it's prediction based on the previous data also changes dramatically and we finally conclude that the model is completely worthless.

Let us consider our MLR model –

Now the least square estimators of the model parameters i.e. β's are explained as

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

Now the matrix $(X^{\intercal}X)$ must be non-singular for the inverse to exist.Since the elements of $(X^{\intercal}X)$ are continuous real values it has a very high probability that it is non-singular.Now the original problem occurs when the matrix is non-singular but it is actually near singular i.e the determinant of $(X^{\intercal}X)$ is ≈ 0. Then this $(X^{\intercal}X)$ matrix is said to be the ill-conditioned matrix.

11.1 Simple Correlation Check

Let us first look into the simple correlation matrix for the whole data-set. Now the correlation matrix is not only hard to interpret but also very difficult to visualize. So what we want to say that instead of correlation matrix we try to visualize it by the correlation heat map. Here light colors represent the high positive correlation and the dark color represent the high negative correlation, in between the color changes with the from light to dark gradient as the value of the correlation changes from 1.0 to -1.0.

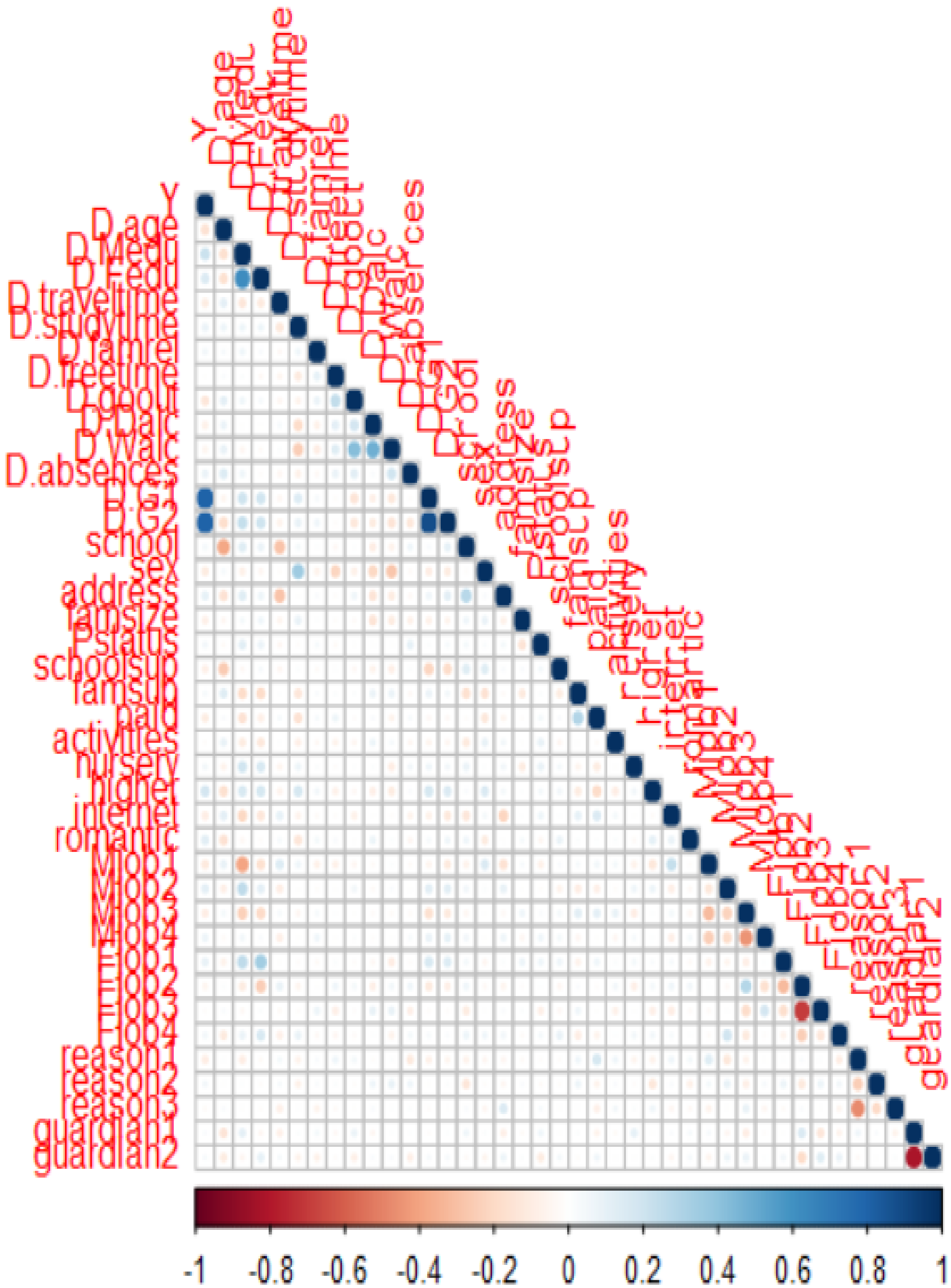


Figure 28: The circle correlation heat-map of our data-set

The correlation heat-map is only a graphical consideration so we can't consider it as a real criterion for checking out the multicollinearity. We have to choose some of the theoretical measurement for checking the multicollinearity in the model.

11.2 Variance Inflation Factor

Since the correlation matrix is not very informative for our data, we choose our second alternative procedure for finding multicollinearity in the data,if presents. Now multicollinearity exists only if the continuous columns are nearly linearly dependent. so we first standardize the continuous columns and try to apply the VIF method to completely get rid of the multicollinearity problem. We first check the VIF values for our model for each of the regressor variables and as a rule of thumb if

- If VIF value is greater then 5, we can say that multicollinearity is present in the model.
- Else multicollinearity is not present in our model.

Now th VIF values for our model –

```
> library(DescTools)
> sort(VIF(MM))
    famrel    Pstatus    famsize activities   romantic    nursery     health
  1.138377   1.143370   1.153333   1.156468   1.167666   1.169243   1.183598
 schoolsup   internet   absences     higher   freetime     famsup       paid
  1.255243   1.256796   1.269135   1.290706   1.304345   1.307425   1.319573
traveltime    address  studytime    reason2      goout        sex     school
  1.323228   1.378876   1.389910   1.444882   1.482833   1.484660   1.509813
   reason3      Mjob2        age    reason1       Dalc       Fedu      Fjob4
  1.660698   1.716523   1.797482   1.815808   2.026085   2.097141   2.109837
      Walc      Fjob1      Mjob4       Medu      Mjob1      Mjob3  guardian2
  2.405020   2.689030   2.695511   2.939196   3.221013   3.788406   3.955019
 guardian1         G2         G1      Fjob3      Fjob2
  4.022799   4.401860   4.592133   5.340952   6.130967
```

Figure 29: The VIF values for all of our regressors

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

```
> sort(VIF(MM))
    famrel    Pstatus    famsize activities    nursery   romantic     health
  1.130963   1.141372   1.151718   1.155598   1.156560   1.166193   1.183489
     Fjob4   internet  schoolsup   absences     higher   freetime traveltime
  1.248615   1.252201   1.252428   1.266218   1.290705   1.296121   1.302502
    famsup       paid      Fjob3    address  studytime      Fjob1    reason2
  1.305550   1.317914   1.330835   1.378876   1.389192   1.415343   1.444866
       sex      goout     school    reason3      Mjob2        age    reason1
  1.478752   1.482785   1.504526   1.657971   1.710385   1.788573   1.815754
      Dalc       Fedu       Walc      Mjob4       Medu      Mjob1      Mjob3
  2.005758   2.096419   2.365465   2.695487   2.934064   3.212194   3.777425
 guardian2  guardian1         G2         G1
  3.940240   4.021043   4.369571   4.534113
```

Figure 30: The VIF values for all of our regressors f after deleting the correlated variable

Since none of the variable has VIF value greater than 5, hence we can conclude that the number of regressor included in our model are free from multicollinearity and this is our final model before the variable selection procedure has been done.

Now thefinal model summary is described as –

```
> 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   
```

```
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     health
  1.130963   1.141372   1.151718   1.155598   1.156560   1.166193   1.183489
     Fjob4   internet  schoolsup   absences     higher   freetime traveltime
  1.248615   1.252201   1.252428   1.266218   1.290705   1.296121   1.302502
    famsup       paid      Fjob3    address  studytime      Fjob1    reason2
  1.305550   1.317914   1.330835   1.378876   1.389192   1.415343   1.444866
       sex      goout     school    reason3      Mjob2        age    reason1
  1.478752   1.482785   1.504526   1.657971   1.710385   1.788573   1.815754
      Dalc       Fedu       Walc      Mjob4       Medu      Mjob1      Mjob3
  2.005758   2.096419   2.365465   2.695487   2.934064   3.212194   3.777425
 guardian2  guardian1         G2         G1
  3.940240   4.021043   4.369571   4.534113
```

Figure 31: Summary of the final model before variable selection

12 variable Selection

Now in our model there are too many regressor variables so it not compatible for many situation to collect those information about these variables correctly. So it is necessary to choose a perfect set of regressors that are as good as the full set of regressors. So there are mainly three methods to choose these variables and we would choose the variobles based on the AIC(Akaike Information Criterion) whose main principal is smaller the value better value result.

12.1 Forward Selection

The basic criterion for the variable selection by the forward selection method is described by the following algorithm –

- We start with the intercept model and compute the AIC for the model.
- We then compute AIC for all the possibilities of adding one more variable in our intercept only model. We select the variable with the smallest AIC if it has a lower AIC than intercept only model.

- We then again compute AIC for adding one more variable in the model. We sort the values by ascending order and variables are added depending on the value of AIC.
- We continue performing these steps until any further addition results in increase of AIC of the model.

For our model the forward selection curve and the summary of the selected model is the following –

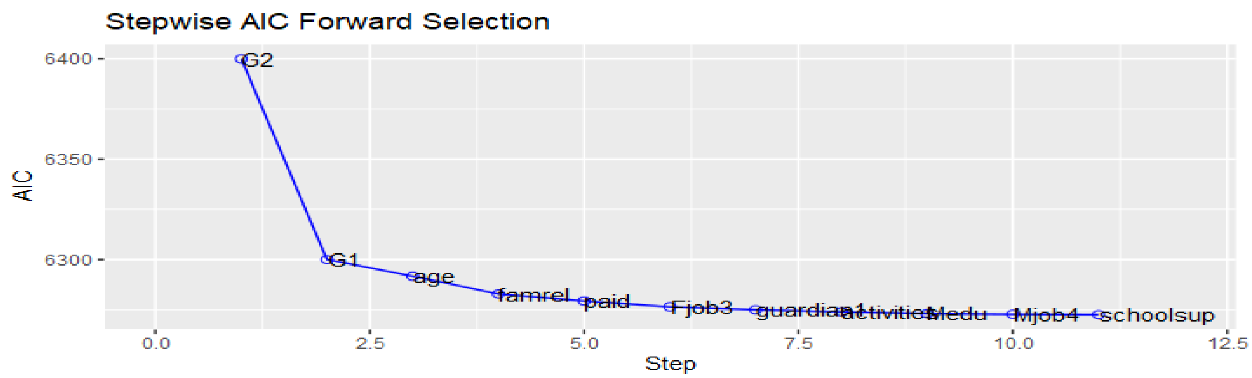


Figure 32: Variables selection sequence in forward selection

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

Figure 33: summary of the model selected by the forward selection

This model has only 11 variables and the $adjusted R^2$ value is 0.8892 which is nearly the same as the previous. So this model is as good as the previous one but has a big advantage that here the number of variables is only 11 nearly one-forth than the previous.

12.2 Backward Elimination

The basic criterion for the variable deletion by the backward elimination method is described by the following algorithm –

- We start with the full model(model with all the regressors and the intercept) and compute the AIC for the model.
- We then compute AIC for all the possibilities of deleting one more variable from our full model. We delete the variable with the smallest AIC if it has a lower AIC than full model.
- We then again compute AIC for deleting one more variable from the model. We sort the values by ascending order and the existing variable is subtracted depending on the value of AIC.
- We continue performing these steps until any further deletion results in increase of AIC of the model.

For our model the backward elimination curve and the summary of the selected model is the following –

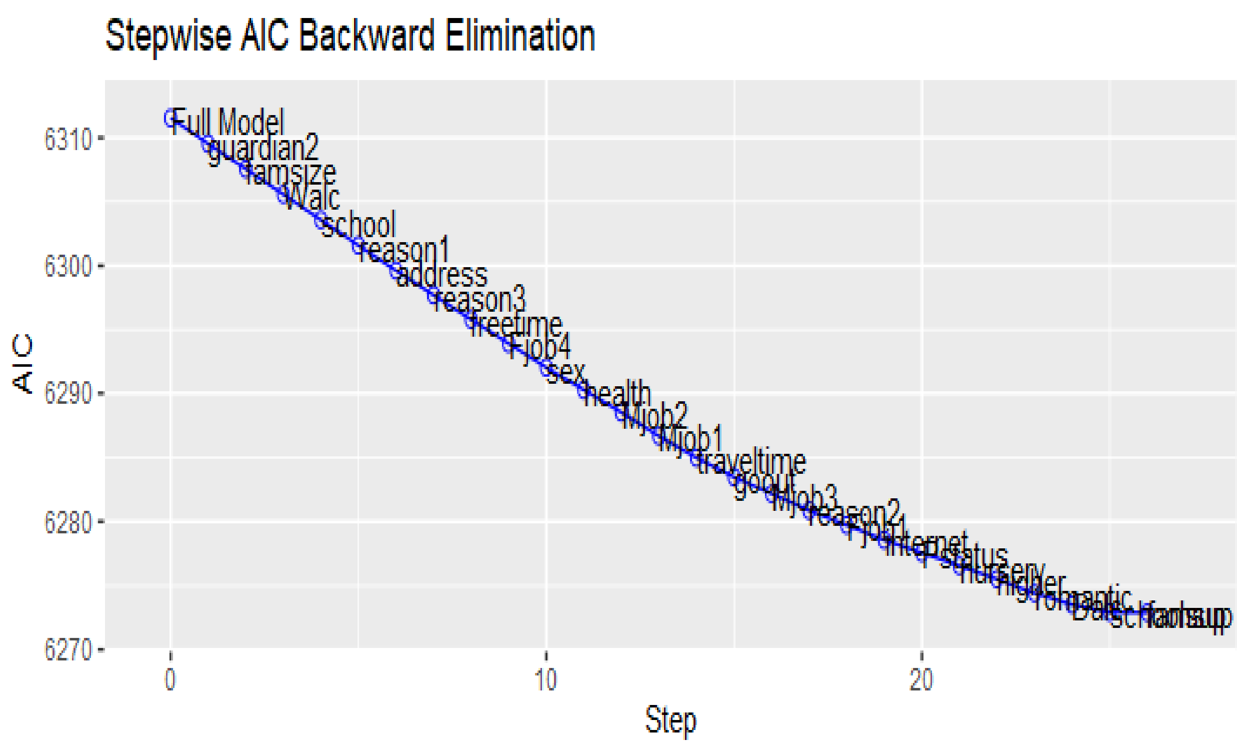


Figure 34: Variables deletion sequence in backward elimination

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

Figure 35: summary of the model selected by the backward elimination

This model has only 13 variables and the $adjusted R^2$ value is 0.8897 which is nearly the same as the previous. So this model is as good as the previous one but has a big advantage that here the number of variables is only 13 nearly one-forth than the original model and approximately same as the forward selection model.

12.3 Stepwise Selection

A combination of forward selection and backward elimination procedure is the stepwise regression for selecting a basket of good variables. It is a not only a modification of forward selection procedure but also the backward elimination one and has the following steps.

- We start with the intercept model and compute the AIC for the model.
- We then compute AIC for all the possibilities of adding one more variable in our intercept only model. We select the variable with the smallest AIC if it has a lower AIC than intercept only model.
- We then again compute AIC for adding one more variable in the model along with the AIC for removing the already added variable. We sort the values by ascending order and variables are either added or the existing variable is subtracted depending on the value of AIC.
- We continue performing these steps until any further action - addition or subtraction, results in increase of AIC of the model.

For our model the stewise selection curve and the summary of the selected model is the following –

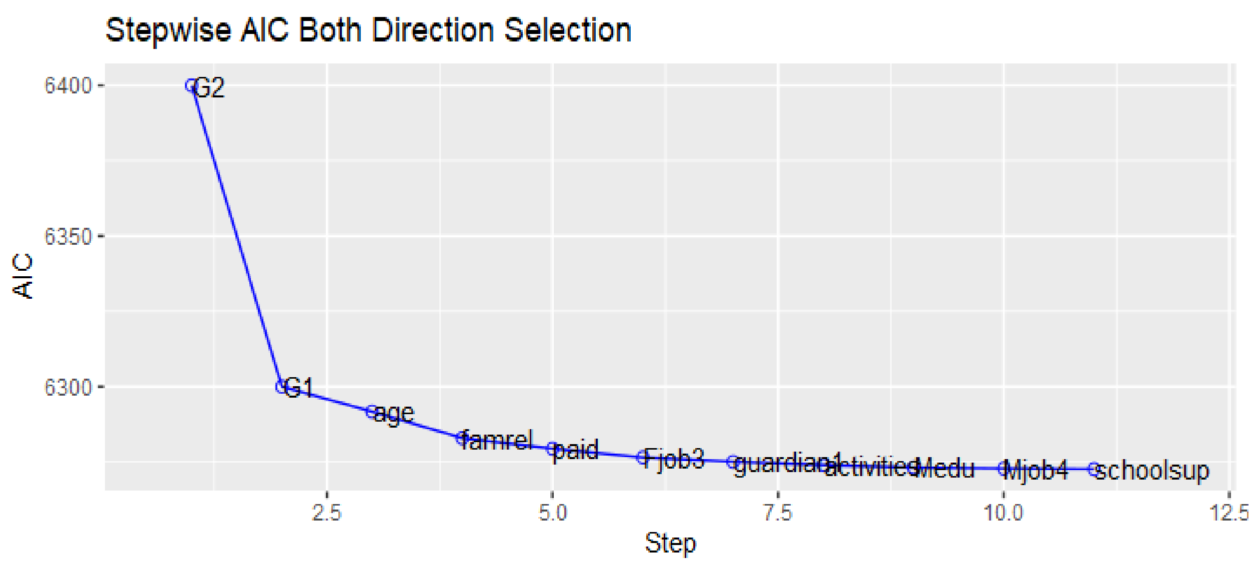


Figure 36: Variables selection or deletion sequence in stepwise selection

```
> 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
```

Figure 37: summary of the model selected by the stepwise selection

This model has only 13 variables and the $adjustedR^2$ value is 0.8897 which is exactly the same as the previous. So this model is as good as the original one but has a big advantage that here the number of variables is only 13 nearly one-forth than the original model and approximately same as the forward selection model.

13 Conclusion

Here all the three model suggests nearly the same amount of regressor variables 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 he backward elimination i.e. a combination of the above two method, we can select the final model as the model selected by the stepwise selection procedure. 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 stepwise selection procedure.

14 reference

1. **Introduction to Linear Regression Analysis** - Douglas C Montgomery, Elizabeth A Peck , G. Geoffrey Vining

2. **Applied Regression Analysis, Third Edition** - Norman R. Draper, Harry Smith

3. **Regression Analysis Lecture Notes by Prof. Sharmishtha Mitra(Course MTH - 416A)**

4. https://en.wikipedia.org/wiki/Leverage_(statistics)

5. http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram

6. https://cran.r-project.org/web/packages/olsrr/vignettes/intro.html

7. https://www.rdocumentation.org/packages/ggplot2/versions/3.3.6