MBA652A STATISTICAL MODELLING FOR BUSINESS ANALYTICS

PROJECT REPORT
PREDICTING TEST SCORES OF STUDENTS

SUBMITTED TO:
Dr. Devlina Chatterjee
IME Department, IIT Kanpur

SUBMITTED BY:
Mr. Faraz Ahmad Khan – 19114006
Mr. Yash Kalpesh Panchal – 19114020
Ms. Anushri Singh – 19125010
Mr. Ashwin Bhide – 19125015

INDEX

OBJECTIVE:

The objective of this project is to determine the various factors affecting the test score and thus build a linear regression model depicting the effects of these factors. We will be formulating a model to depict the effect of these factors. The data set is collected from archives of the UCI Machine Learning Repository.

BACKGROUND:

It is important to know what are the casual-effect relationships between different entities in our day-to-day life. This will enable us not to understand what parameters affect our target entity and to what extent. It also helps us to know which are the variables that we can control as this will help us to get the desired output in any situation. We will be studying the student's performance in a test in this case.

METHODOLOGY:

We use a multiple linear regression model to predict the test scores by successively adding the most significant attribute affecting the test score based on various factors like correlation coefficient, adjusted R-square, RSE, etc.

DATA DESCRIPTION AND SOURCE:

Dataset: Student performance in secondary school (for 1 subject)
https://archive.ics.uci.edu/ml/datasets/Student+Performance

Source:
Paulo Cortez, University of Minho, GuimarÃ£es, Portugal

Attribute Information: Independent Variables

1. **School** - student's school (numeric: 1 - 'GP' - Gabriel Pereira or 2 - 'MS' - Mousinho da Silveira)
2. **Sex** - student's sex (numeric: 1- 'F' - female or 2- 'M' - male)
3. **Age** - student's age (numeric: from 15 to 22)
4. **Address** - student's home address type (numeric: 2- 'U' - urban or 1- 'R' - rural)
5. **Famsize** - family size (numeric: 1- 'LE3' - less or equal to 3 or 2- 'GT3' - greater than 3)
6. **Pstatus** - parent's cohabitation status (numeric: 2- 'T' - living together or 1- '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: 5-'teacher', 2-'health' care related, civil 4- 'services' (e.g. administrative or police), 1-'at_home' or 3-'other')
10. **Fjob** - father's job (nominal: 5-'teacher', 2-'health' care related, civil 4- 'services' (e.g. administrative or police), 1-'at_home' or 3-'other')
11. **Reason** - reason to choose this school (nominal: close to 2-'home', school 4-'reputation', 1-'course' preference or 3-'other')
12. **Guardian** - student's guardian (nominal: 2-'mother', 1-'father' or 3-'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 (numeric: 1-no or 2-yes)
17. **Famsup** - family educational support (numeric: 1-no or 2-yes)
18. **Paid** - extra paid classes within the course subject (Math or Portuguese) (numeric: yes or no)
19. **Activities** - extra-curricular activities (numeric: 1-no or 2-yes)
20. **Nursery** - attended nursery school (numeric: 1-no or 2-yes)
21. **Higher** - wants to take higher education (numeric: 1-no or 2-yes)
22. **Internet** - Internet access at home (numeric: 1-no or 2-yes)
23. **Romantic** - with a romantic relationship (numeric: 1-no or 2-yes)
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)

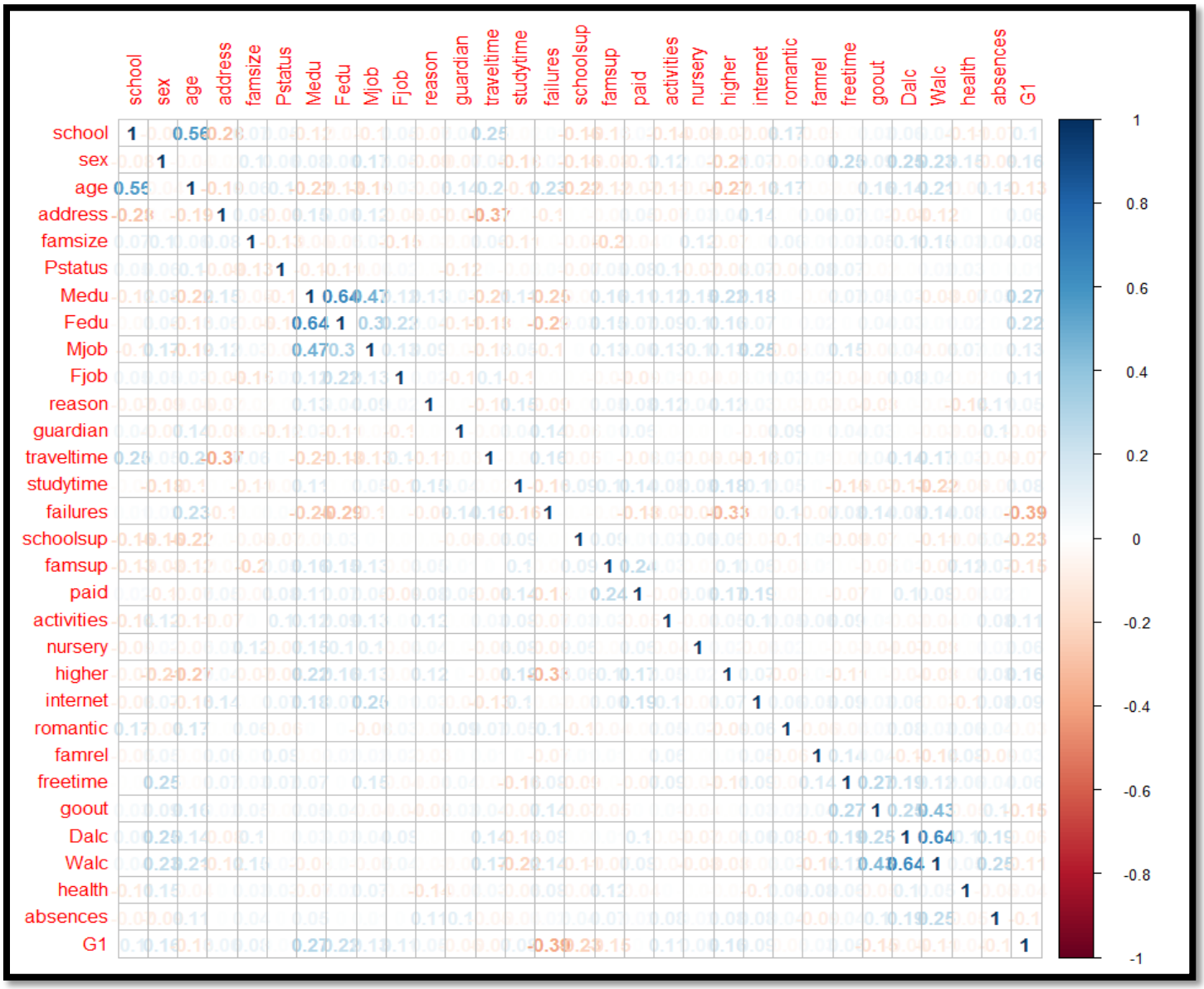
Attribute Information: Dependent Variables

G1(Marks) - Score (numeric: from 0 to 20)

Scatter Plots: Please refer to Annexure I for all scatter plots

CORRELATION MATRIX:

The matrix shows the correlation between independent variables amongst themselves and the correlation of independent variables with the dependent variable.



MULTICOLLINEARITY and VIF:

When two or more explanatory variables in a multiple regression model are highly linearly related, then the data set is said to have multicollinearity. It is evident from the correlation matrix that none of the variables are highly correlated with another, hence it can be deduced that the data is not having multicollinearity. Also, this is explained by the VIF (Variation Inflation Factor) of variables used to build for all variables.

As none of the variables are having VIF > 10, we can conclude that there is no multicollinearity between the variables.

Attribute	VIF
school	1.8548
sex	1.3889
age	1.9708
address	1.3559
famsize	1.1931
Pstatus	1.152
Medu	2.3254
Fedu	2.0097
Mjob	1.4713
Fjob	1.1839
reason	1.1401
guardian	1.1641
traveltime	1.3739
studytime	1.228
failures	1.4016
schoolsup	1.1718
famsup	1.234
paid	1.2604
activities	1.1618
nursery	1.1036
higher	1.356
internet	1.2247
romantic	1.1379
famrel	1.1147
freetime	1.3056
goout	1.4444
Dalc	1.9252
Walc	2.4124
health	1.1788
absences	1.2202

REGRESSION MODELS:

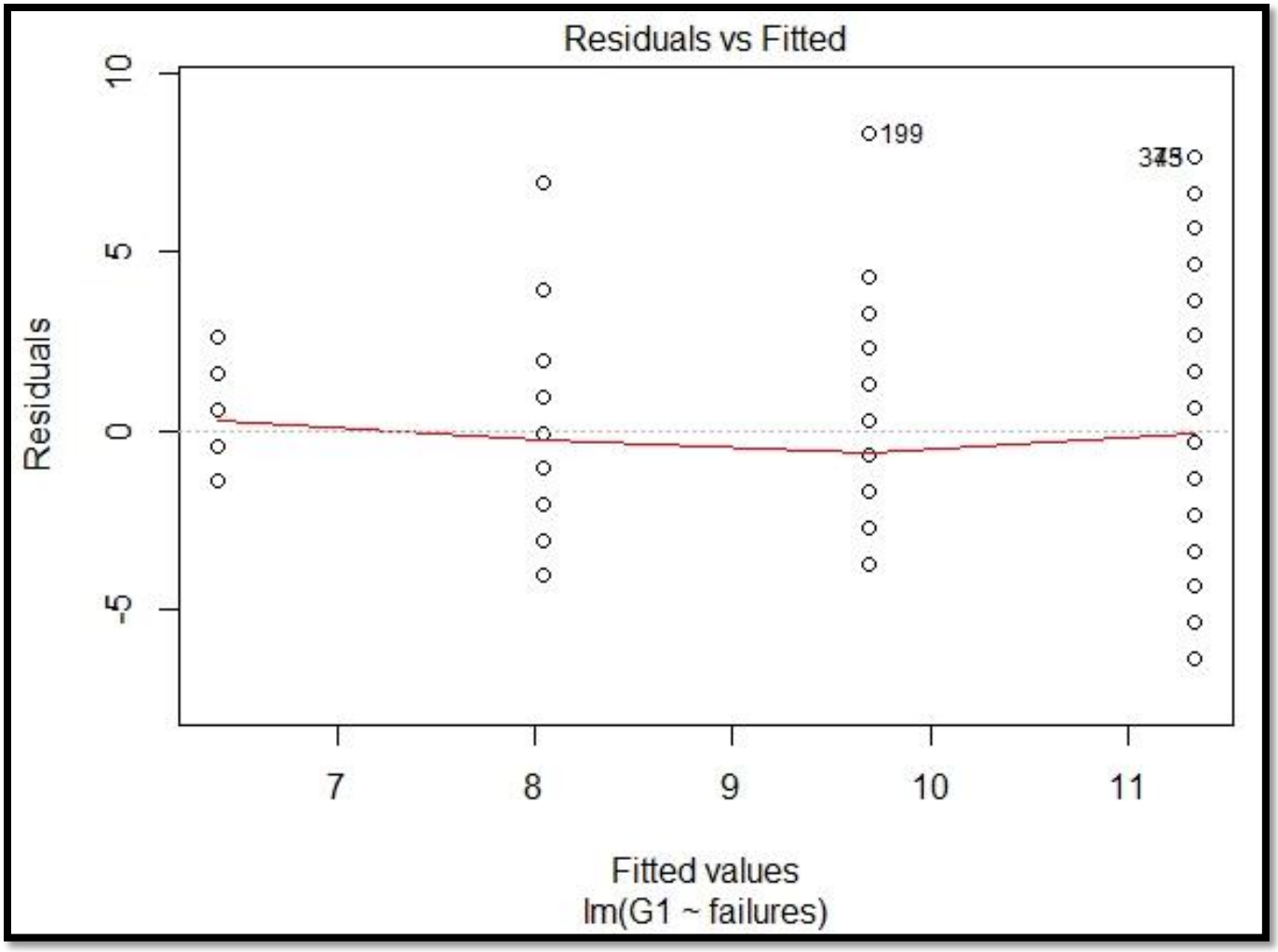
Here as we have 30 independent variables, we made single regression models with those variables only – having correlation more than 0.1, the following table shows the summary of these models

Attribute	Correlation with G1	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
school	0.1	3.327	0.009537	0.005909	2.629	0.1061	273
sex	0.16	3.301	0.02469	0.02112	6.912	0.009048	273
age	-0.13	3.315	0.01637	0.01276	4.542	0.03396	273
medu	0.27	3.219	0.0725	0.0691	21.34	5.94E-06	273
fedu	0.22	3.26	0.04865	0.04517	13.96	0.0002272	273
mjob	0.13	3.315	0.01641	0.01281	4.555	0.03372	273
fjob	0.11	3.322	0.01246	0.008845	3.445	0.06451	273
failures	-0.39	3.073	0.1547	0.1516	49.97	1.31E-11	273
schoolsup	-0.23	3.25	0.05472	0.05126	15.8	9.00E-05	273
famsup	-0.15	3.306	0.0217	0.01812	6.056	0.01448	273
activities	0.11	3.324	0.01117	0.007547	3.084	0.08021	273
higher	0.16	3.301	0.0247	0.02113	6.914	0.009037	273
goout	-0.15	3.307	0.02113	0.01754	5.893	0.01585	273
Walc	-0.11	3.323	0.01167	0.008049	3.223	0.0737	273
absences	-0.1	3.325	0.01028	0.006653	2.835	0.09337	273

From the results of single regression models, we see that the “failures” is the most significant variable.

The initial model can be written as,
Marks = 11.3341 – 1.6446*failures

Residual plot for single regression model Marks vs. failures is given below.



Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	3.073	0.1547	0.1516	49.97	1.31E-11	273

Multiple R^2 is very low for this model, only 15.47% of the data fits the curve (line) of our model, or we can say that only 15.47% of the data is explained by the curve (line), for Model_1, this is due to omitted variable bias, which can be tackled or we can increase this by adding more variables to our model. So, to increase R^2, we need to take into account other variables.

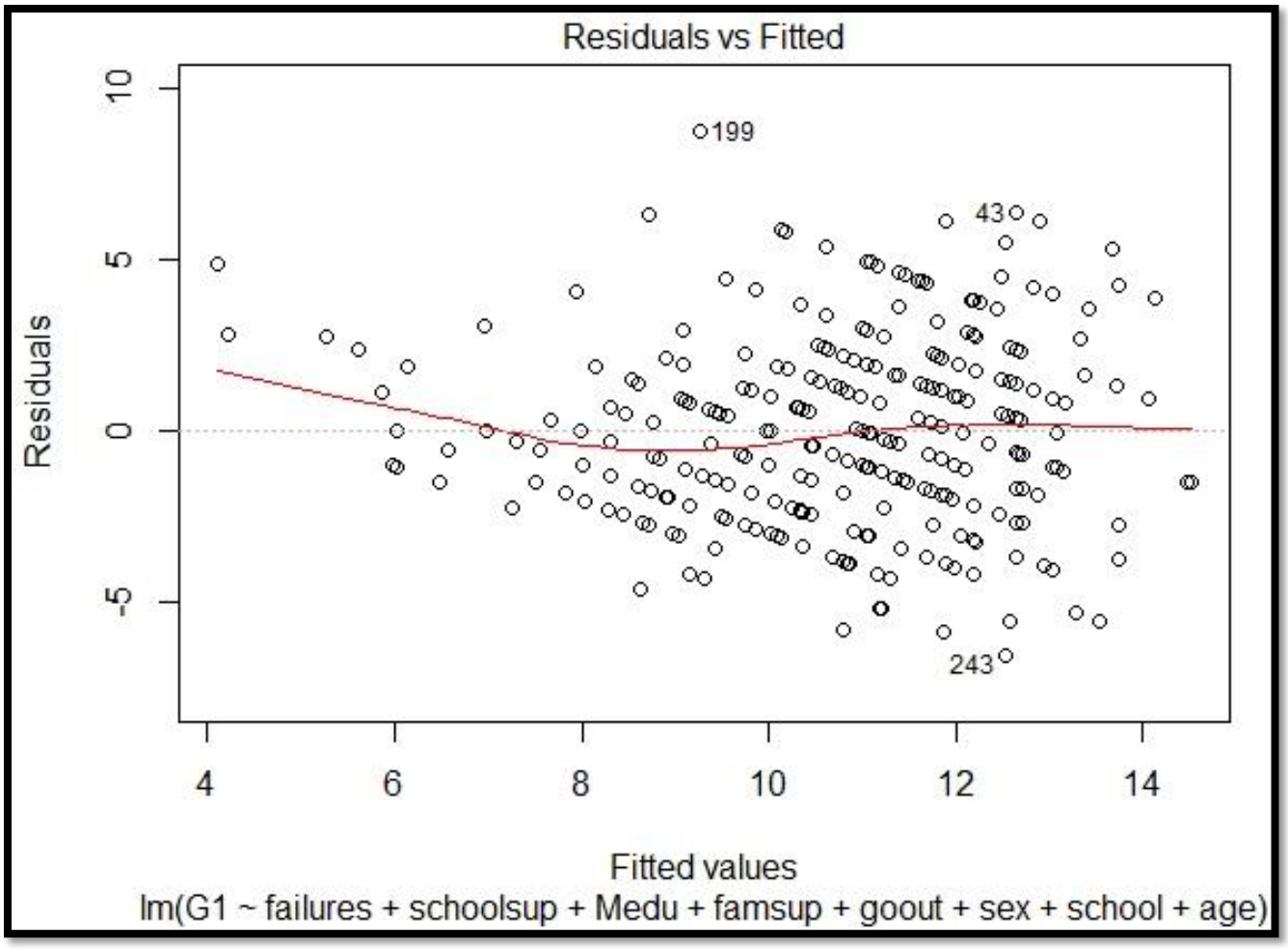
Significant Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null Hypothesis H_0 = Heteroskadasticity is not present in the data set.
Alternate hypothesis H1 = Heteroskadasticity is present in the data set.
This means that when we conduct the Breusch-Pagan (BP) test, the variance of dependent

variable relates to the set of independent variables. If p-value observed with this is less the confidence interval of 95% ($\alpha = 0.05$), then we reject the null hypothesis.

In order to do this, we will be using multiple regression models while adding one more variable at a time to the previously most significant variable added. We are ensuring that no two independent variables added to the multiple regression model are highly correlated. The variables added in multiple regression models are “schoolsup,” “Medu,” “famsup,” “goout," “sex,” “school,” and “age” in respective order. Please note that the initial seven multiple regression models are illustrated in Annexure-II. The final model obtained in the multiple regression model is shown below.

Marks = 19.0440 – 1.2720*failures – 1.8838*schoolsup + 0.5908*Medu –
1.1009*famsup – 0.3828*goout + 0.8123*sex + 1.7185*school – 0.4576*age



Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.797	0.3178	0.2973	15.49	<2.2e-16	266
schoolsup***						
Medu***						
famsup**						
goout*						
sex*						
school**						
age*						

Variables failures, schoolsup, and Medu are 99.9% significant, whereas famsup and school are 99% significant. Variable age, sex, and goout are 95% significant. Still, multiple R^2 is ~32%, which is not good enough, saying that only 32% of the total data is explained, but adding more variables to this model will result in the reduction of adjusted R^2 which is not good; hence we would not add any further variables. The above model is the optimal linear model that could be developed with this given data.

To check whether heteroskadasticity is present or not, the BP test is done. Following is the result of test. Data set used for test is of final model.

BP = 14.171, df = 8, **p-value = 0.07742**

CONCLUSION:

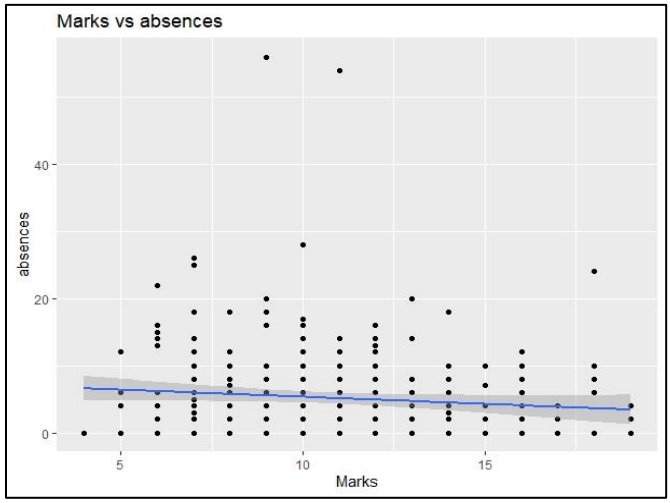
Marks = 19.0440 – 1.2720*failures – 1.8838*schoolsup + 0.5908*Medu –
1.1009*famsup – 0.3828*goout + 0.8123*sex + 1.7185*school – 0.4576*age

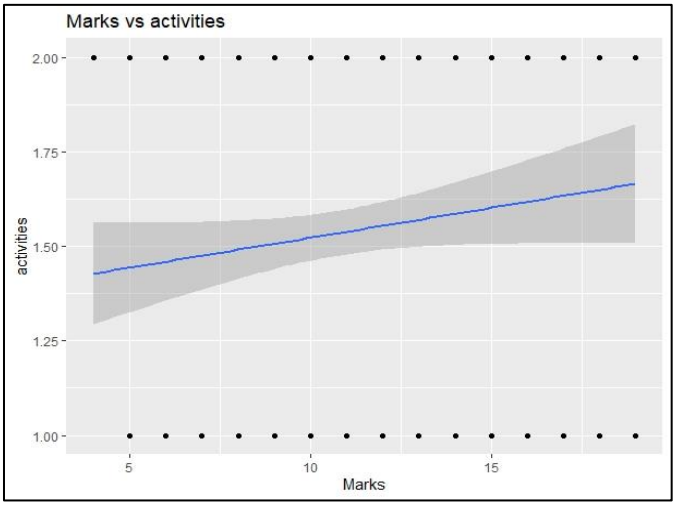
- Multiple R^2 observed may be less due to omitted variable bias. For example IQ of student (quantitative variable) and interest of student in that particular subject (qualitative variable).
- As most of the parameters in the data set are categorical in nature, the impact each parameter on the dependent variable cannot be measured in great depth. This might have contributed to lower multiple R^2 value.
- A non-linear regression model might be able to better correlate the independent and dependent variables in the data set.

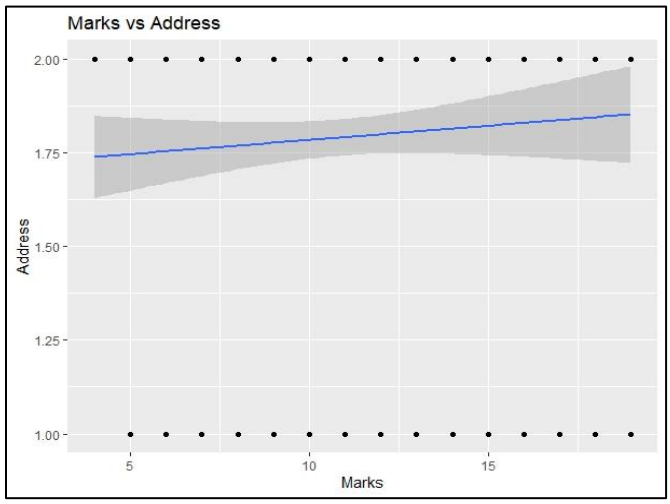
REFERENCE:

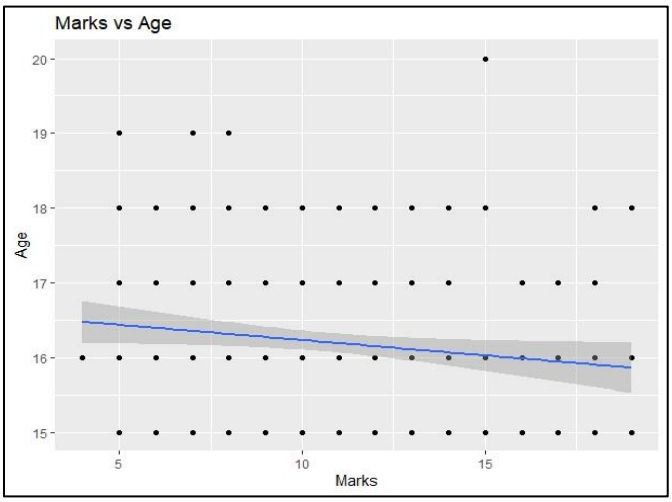
- Data set - https://archive.ics.uci.edu/ml/datasets/student%2Bperformance
- Breusch-Pagan test - https://rpubs.com/cyobero/187387
- Multicollinearity and VIF- http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/

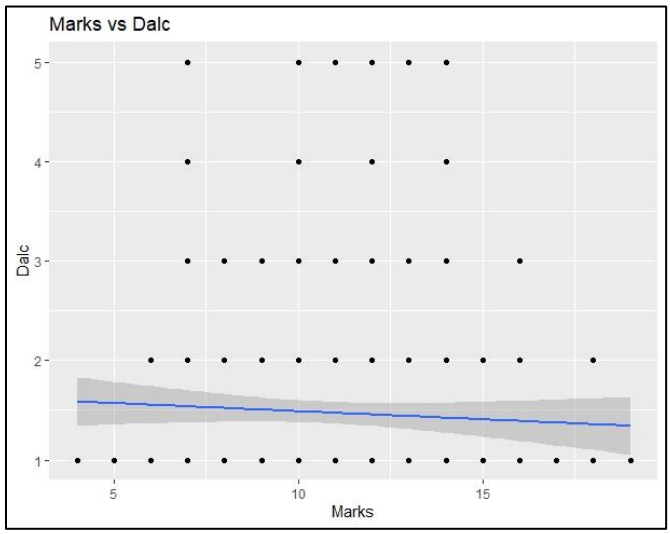
ANNEXURE I

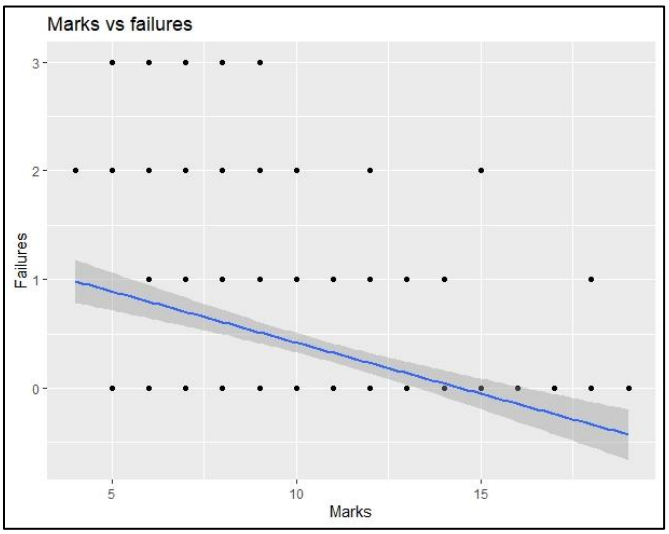

Marks vs absences
absences
Marks
0
20
40
5
10
15

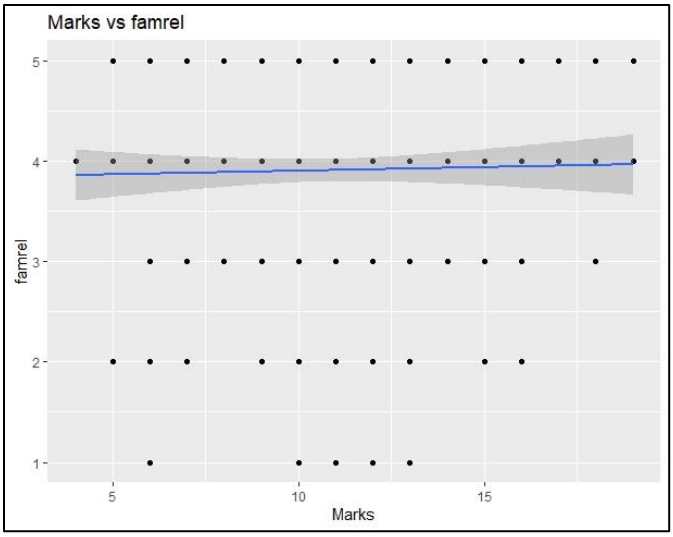

Marks vs activities
activities
Marks
1.00
1.25
1.50
1.75
2.00
5
10
15

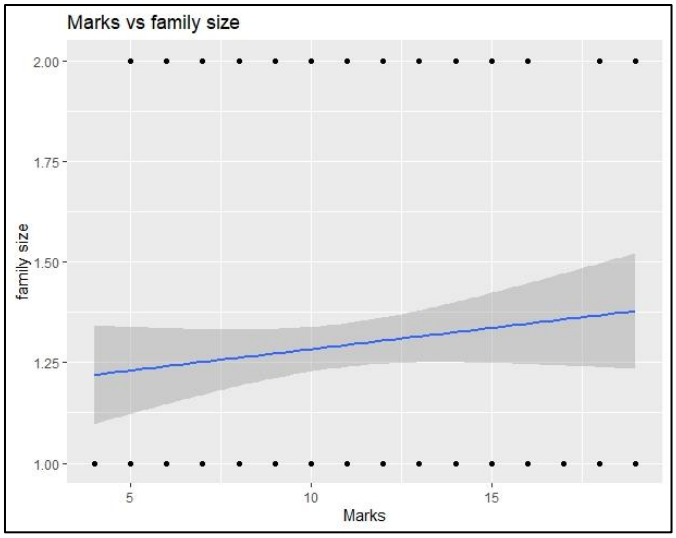

Marks vs Address
Address
Marks
1.00
1.25
1.50
1.75
2.00
5
10
15

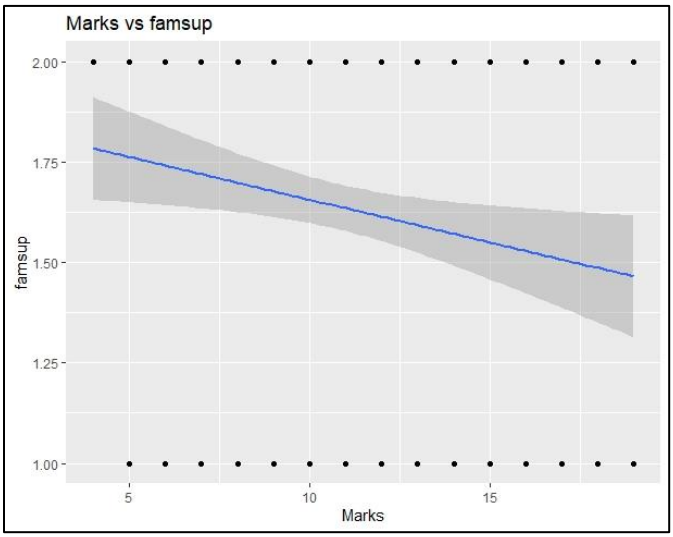

Marks vs Age
Age
Marks
15
16
17
18
19
20
5
10
15

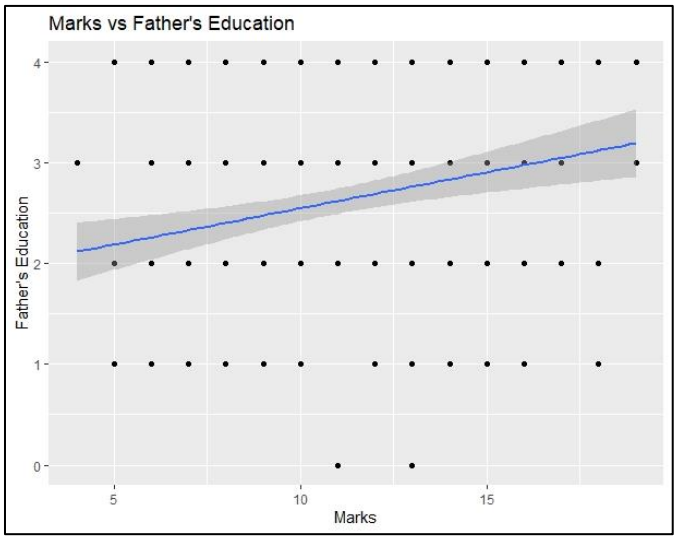

Marks vs Dalc
Dalc
Marks
1
2
3
4
5
5
10
15

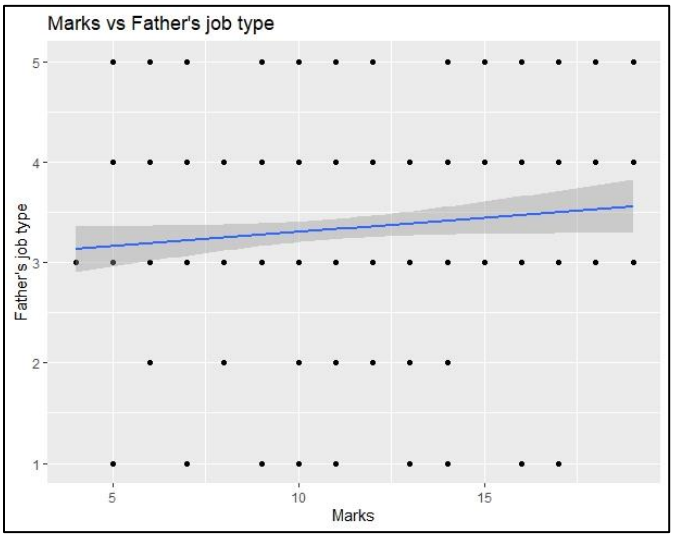

Marks vs failures
Failures
Marks
0
1
2
3
5
10
15

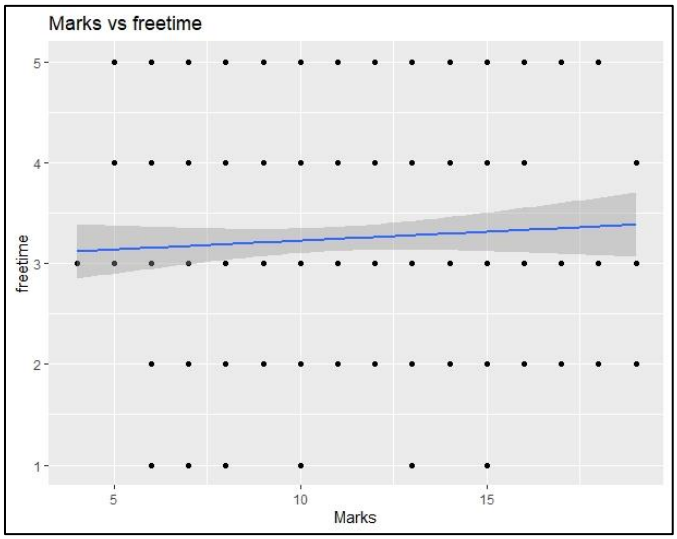

Marks vs famrel
famrel
Marks

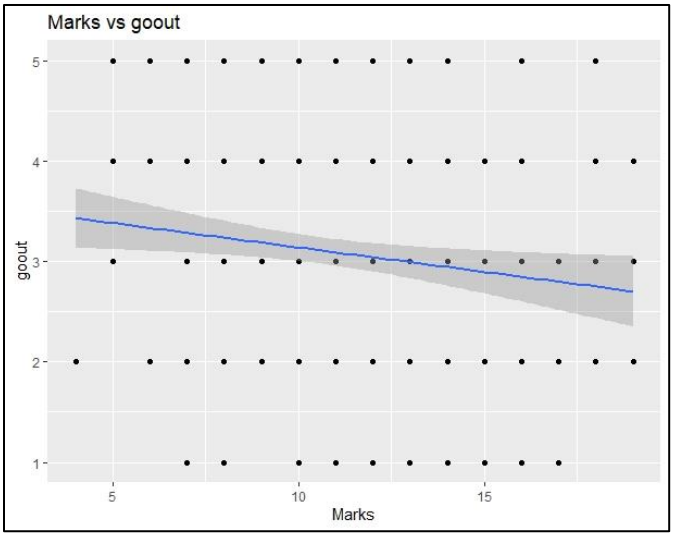

Marks vs family size
family size
Marks

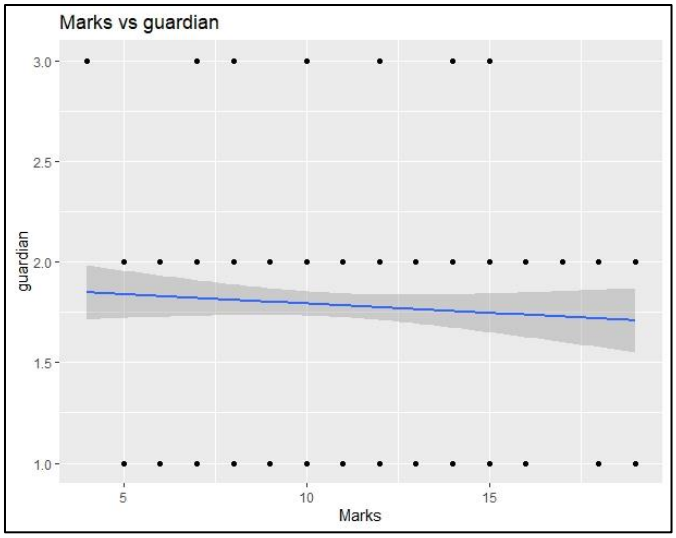

Marks vs famsup
famsup
Marks

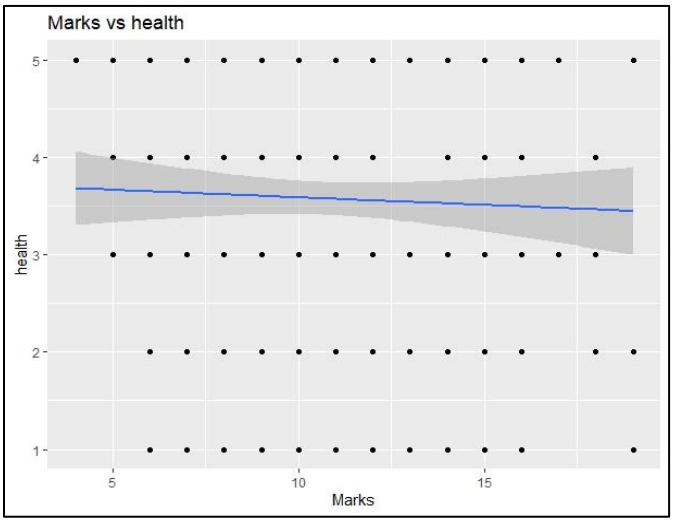

Marks vs Father's Education
Father's Education
Marks

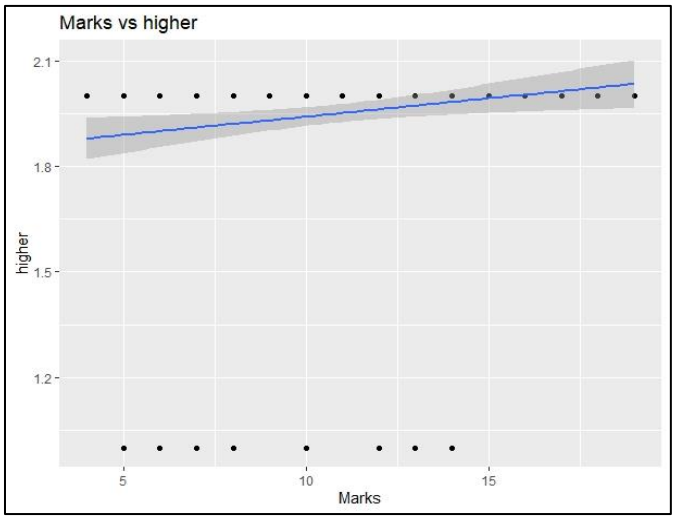

Marks vs Father's job type
Father's job type
Marks

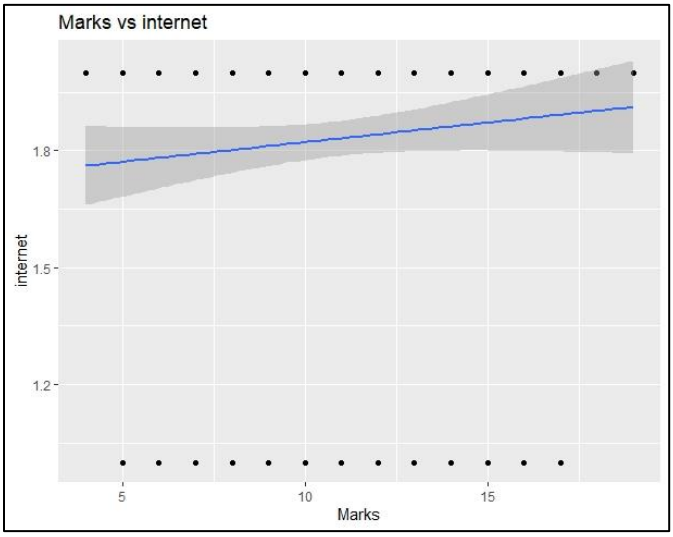

Marks vs freetime
freetime
Marks

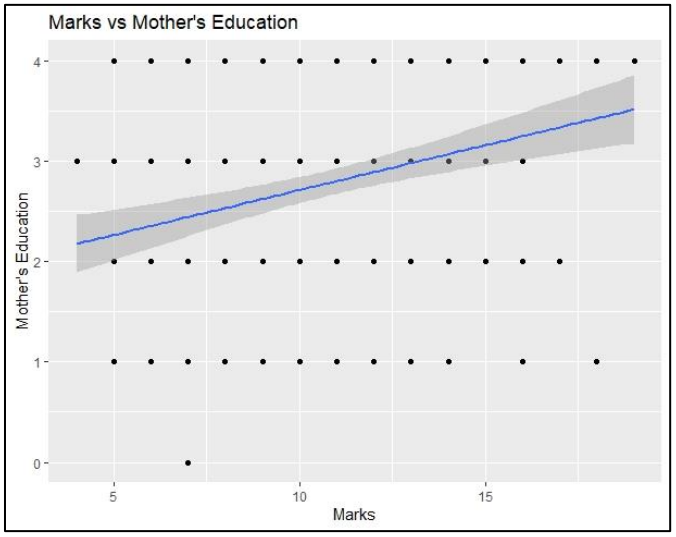

Marks vs goout
goout
Marks
5
4
3
2
1
5
10
15

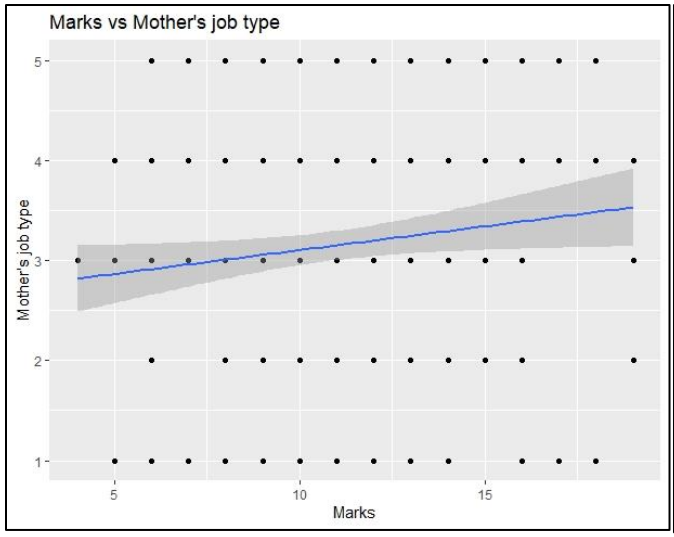

Marks vs guardian
guardian
Marks
3.0
2.5
2.0
1.5
1.0
5
10
15

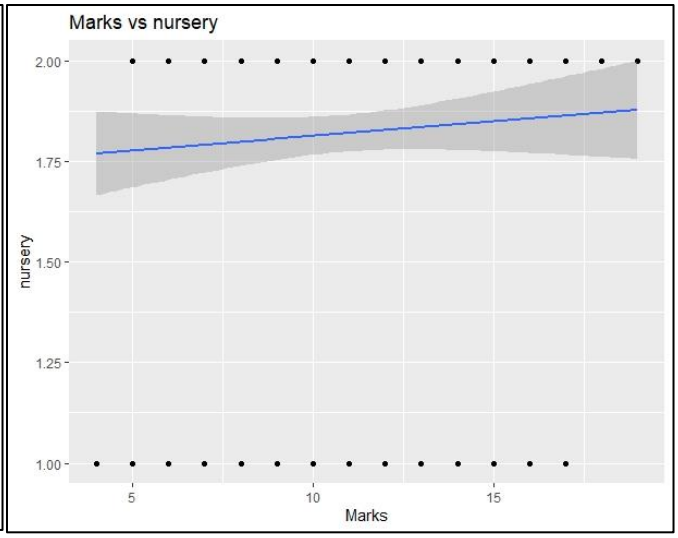

Marks vs health
health
Marks
5
4
3
2
1
5
10
15

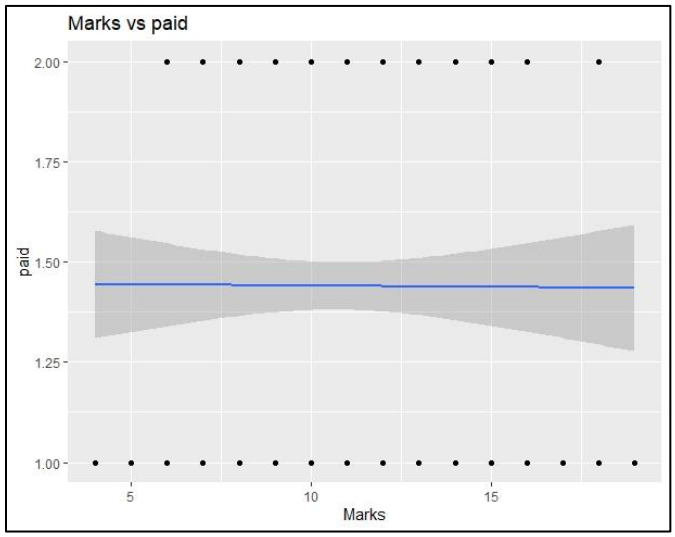

Marks vs higher
higher
Marks
2.1
1.8
1.5
1.2
5
10
15

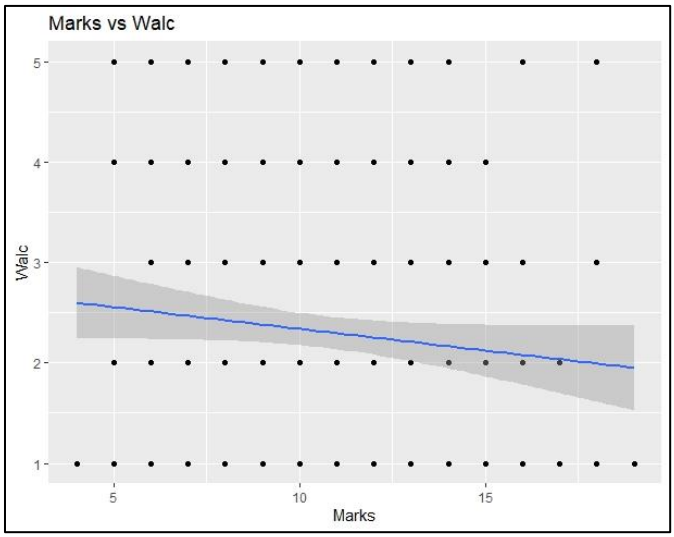

Marks vs internet
internet
Marks
1.8
1.5
1.2
5
10
15

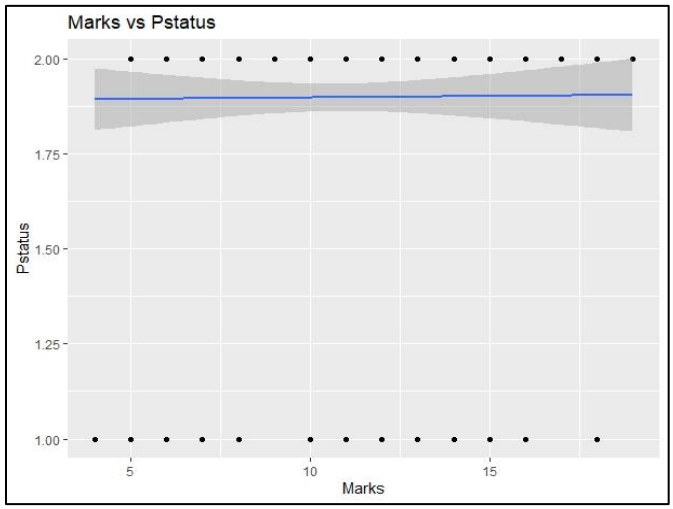

Marks vs Mother's Education
Mother's Education
Marks
4
3
2
1
0
5
10
15

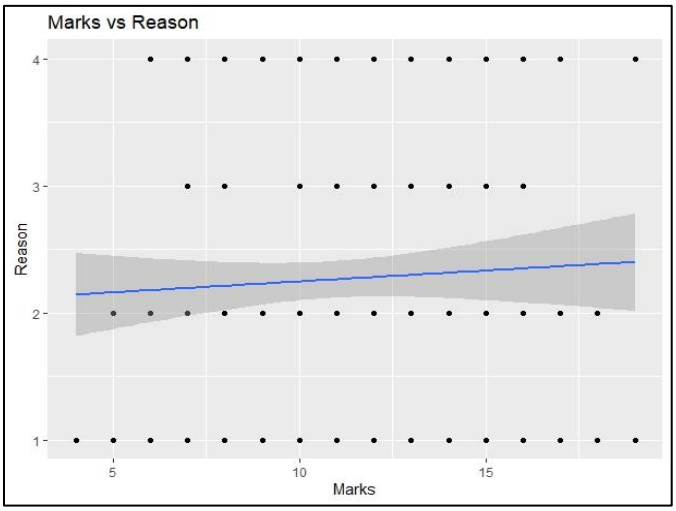

Marks vs Mother's job type
Mother's job type
Marks
5
4
3
2
1
5
10
15

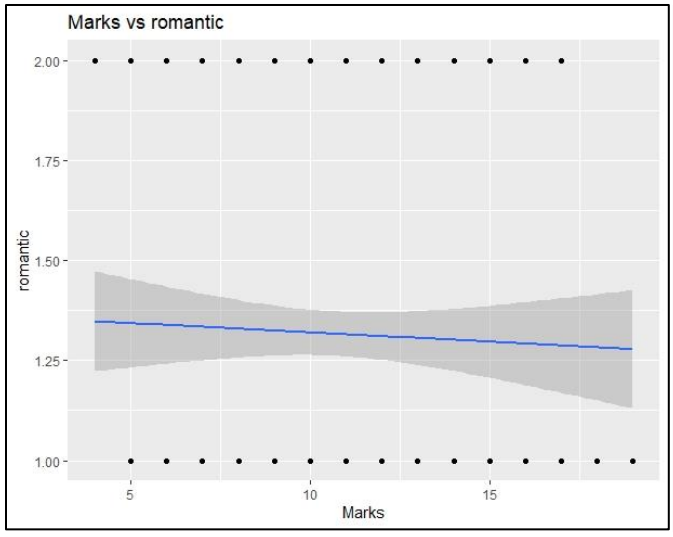

Marks vs nursery
nursery
Marks
2.00
1.75
1.50
1.25
1.00
5
10
15

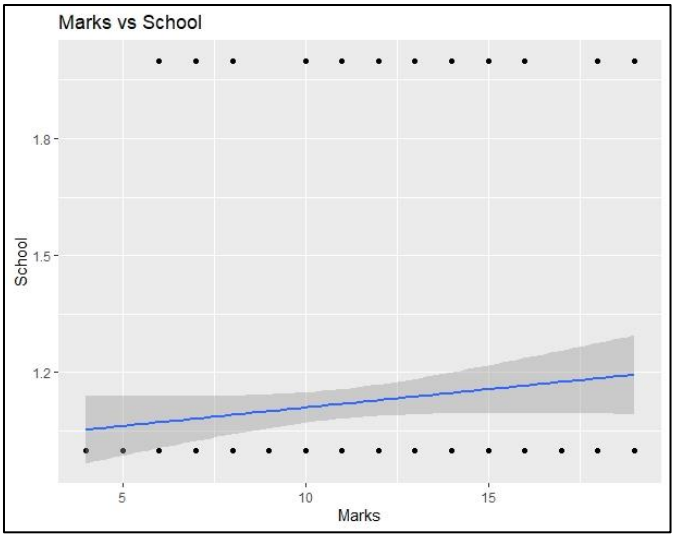

Marks vs paid
paid
Marks
2.00
1.75
1.50
1.25
1.00
5
10
15

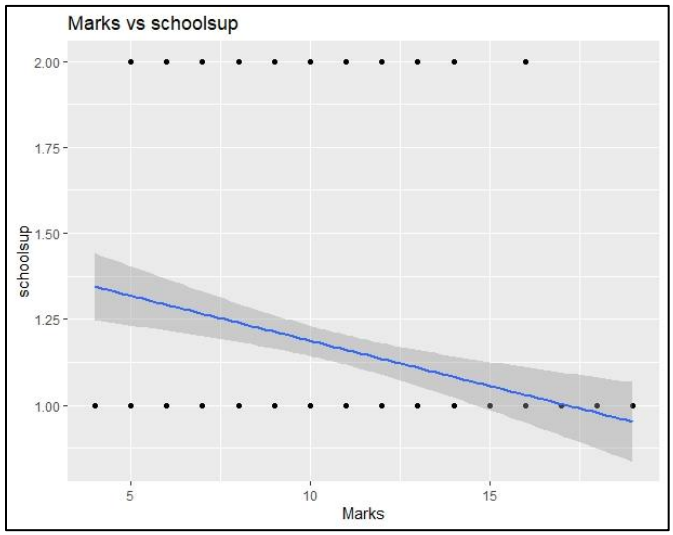

Marks vs Walc
Walc
Marks
5
4
3
2
1
5
10
15

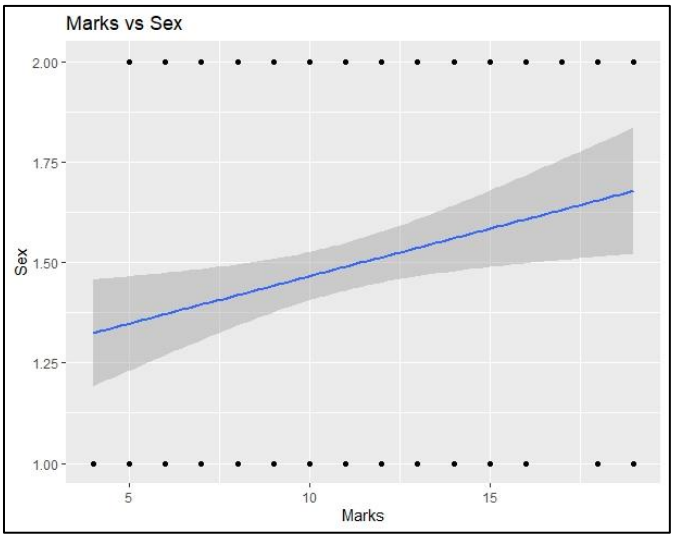

Marks vs Pstatus
Pstatus
Marks
2.00
1.75
1.50
1.25
1.00
5
10
15

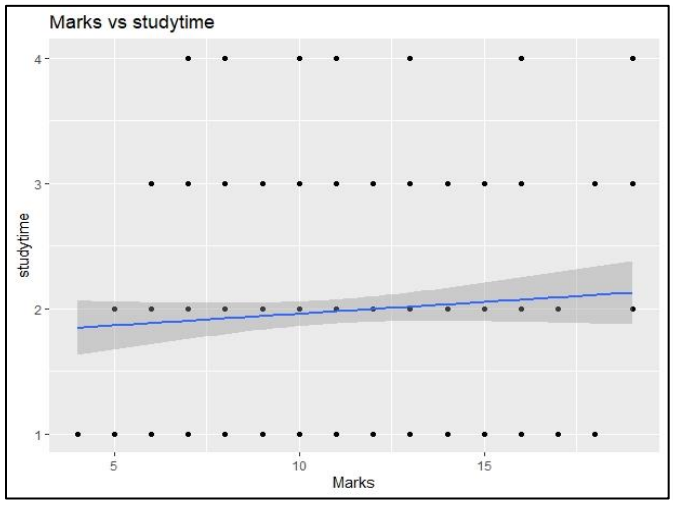

Marks vs Reason
Reason
Marks
4
3
2
1
5
10
15

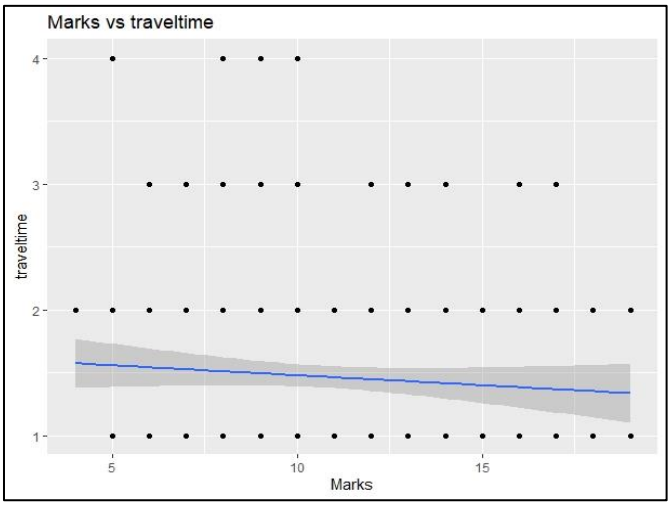

Marks vs romantic
romantic
2.00
1.75
1.50
1.25
1.00
5
10
15
Marks


Marks vs School
School
1.8
1.5
1.2
5
10
15
Marks


Marks vs schoolsup
schoolsup
2.00
1.75
1.50
1.25
1.00
5
10
15
Marks

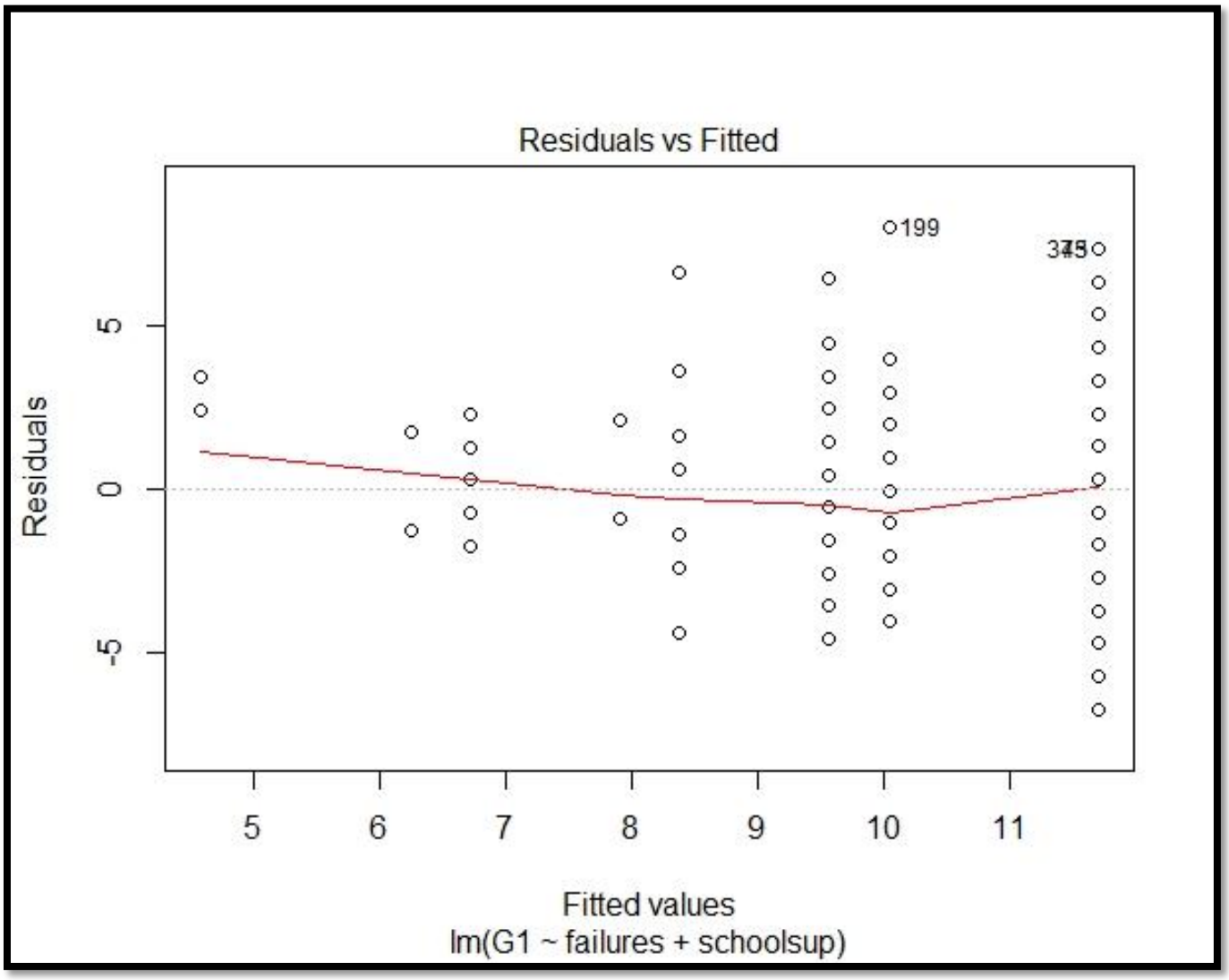

Marks vs Sex
Sex
2.00
1.75
1.50
1.25
1.00
5
10
15
Marks


Marks vs studytime
studytime
4
3
2
1
5
10
15
Marks


Marks vs traveltime
traveltime
4
3
2
1
5
10
15
Marks

ANNEXURE II

Model_1:
Marks = 13.8287 – 1.6575*failures – 2.1332*schoolsup



Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.973	0.2119	0.2061	36.56	8.67E-15	272
Schoolsup***						

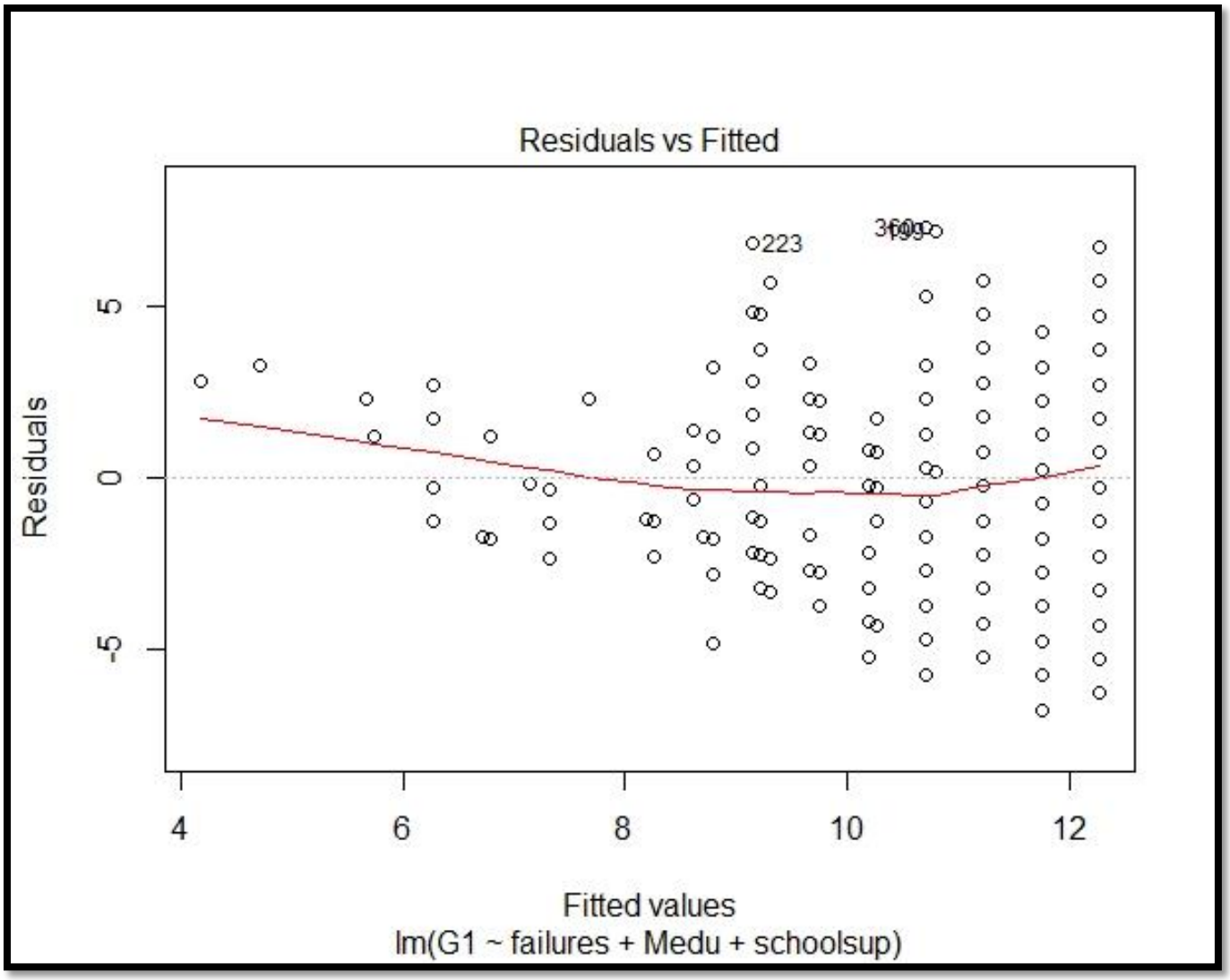
Including another variable schoolsup in a single regression model, we see that adjusted R^2 is increased from 0.1516 to 0.2061, which depicts that Model_1 is much better than a single regression model.

Failures and schoolsup are both 99.9% significant variables. But still, Model_1 is not enough, as R^2 is only 21.19%, which is low, explaining only 21.19% of the total data, still a better model has to be formed, for reducing the omitted variable bias we would add more variables to current Model_1.

Based on the correlation of independent variables and the dependent variable, we concluded that the 3rd most significant variable is Medu, so developing a model using failures, schoolsup, and Medu variables, we get the following model.

Model_2:

Marks = 12.2508 – 1.4758*failures – 2.0765*schoolsup + 0.5213*Medu



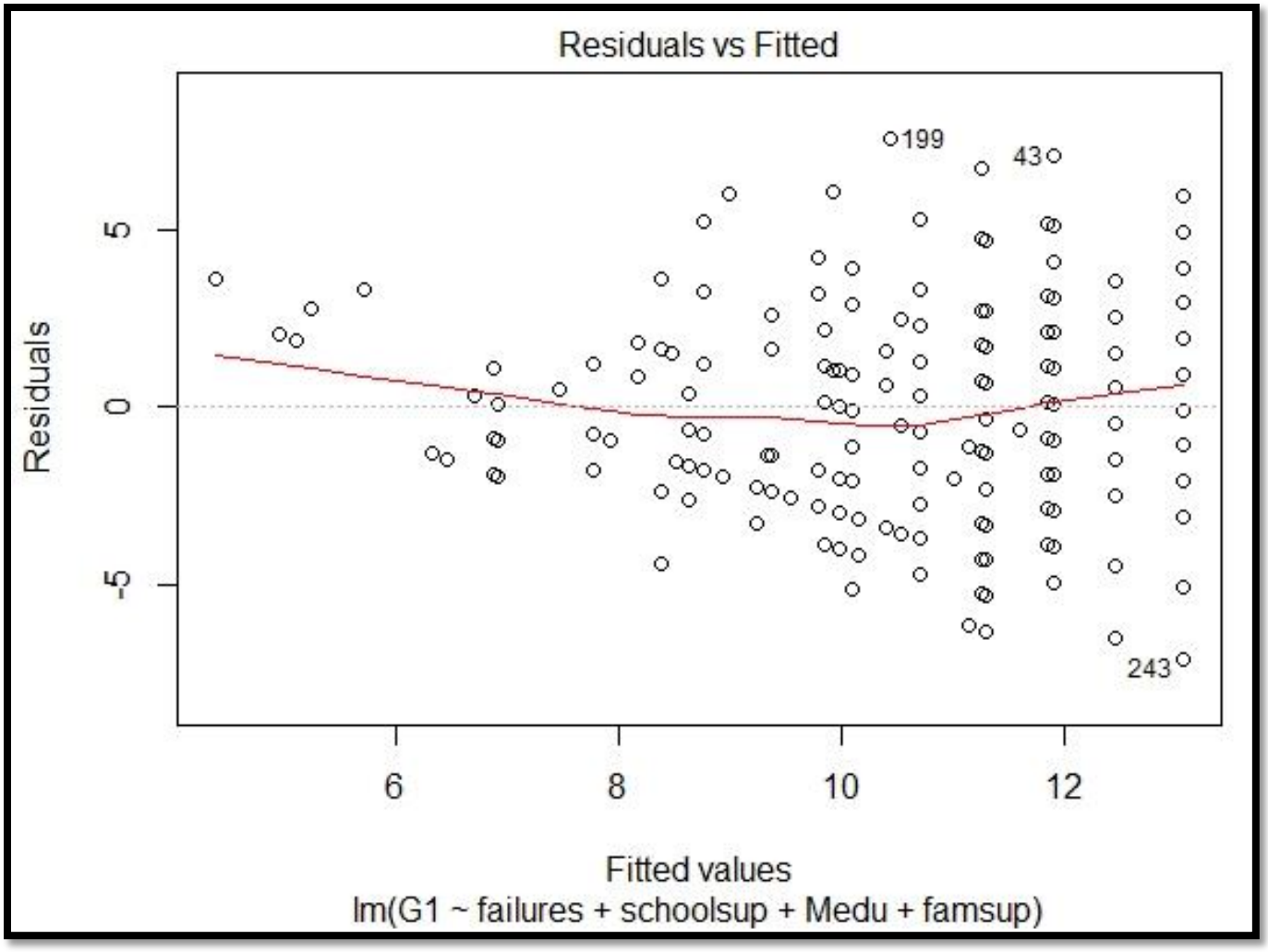
Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.925	0.2398	0.2314	28.5	4.76E-16	271
schoolsup***						
Medu**						

Here, failure and schoolsup are 99.9% significant variables, whereas Medu is 99% significant, but still, there is ~13% increase in adjusted R^2 value, and there ~14% increase in explanation of data on adding Medu to Model_2, hence adding Medu improved our model. But still, the value of multiple R^2 is 24%, which is less, hence still more variables need to be added, to avoid omitted variable bias. Adding famsup to Model_2, we will get the following model.

Model_3:
Marks = 13.7435 – 1.4620*failures – 1.9298*schoolsup + 0.6051*Medu – 1.1596*famsup

Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.878	0.2668	0.2559	24.56	< 2.2e-16	270
schoolsup***						
Medu***						
famsup**						

Further increase in adjusted R^2 is observed, which is ~11%, and ~12% increase in R^2 is seen, still not good enough as only 27% of the data is explained by this model, which is low. In order to increase these values, we need to add more and more variables, until the maximum possible value of adjusted R^2 is achieved.

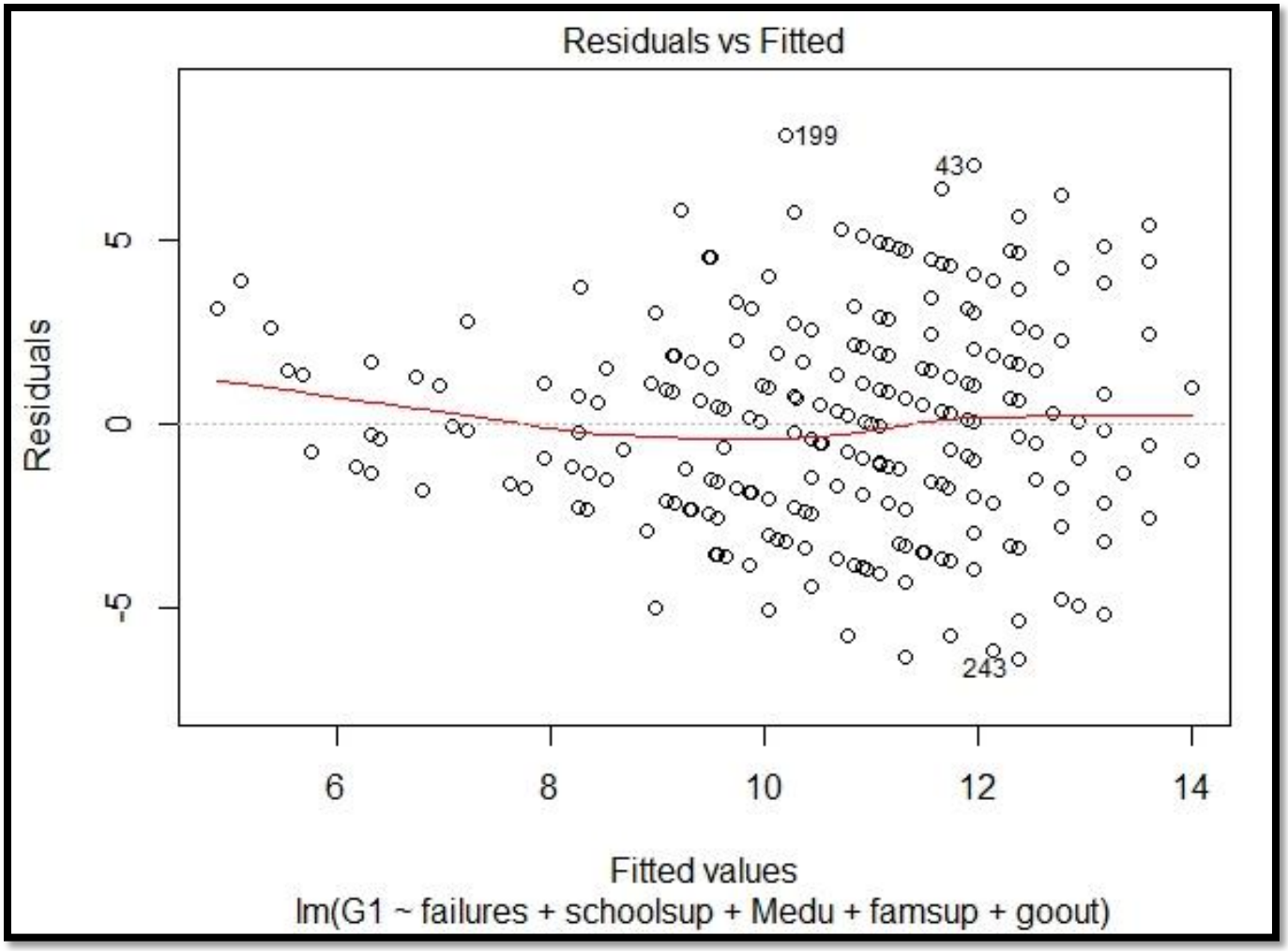


Based on the correlation of variables we add goout to Model_3 as it gave us the maximum increase in adjusted R^2, as elaborated in Model_4

Model_4:

Marks = 15.0353 – 1.3717*failures – 1.9981*schoolsup + 0.6455*Medu – 1.2167*famsup – 0.4069*gout

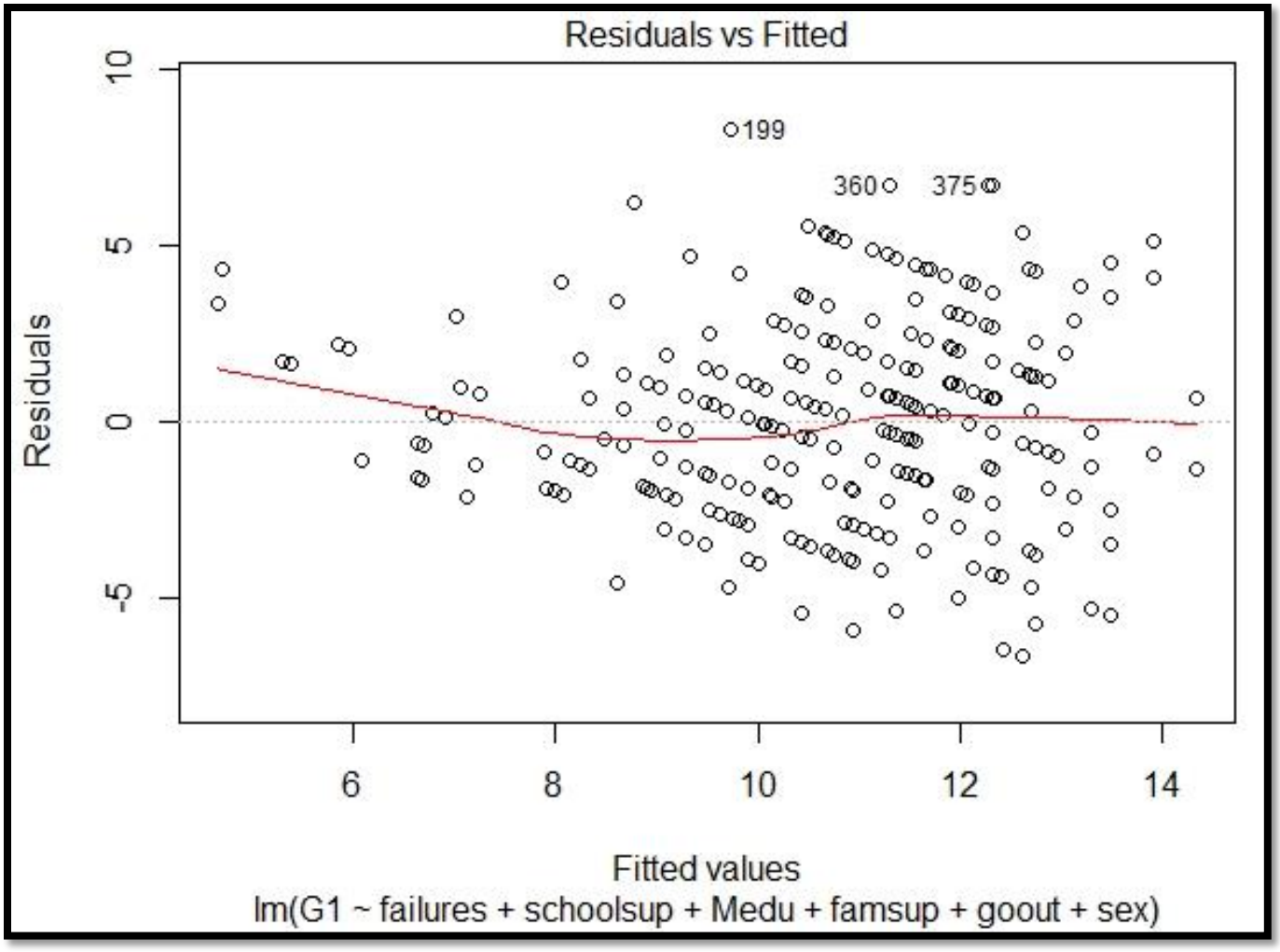
Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.848	0.2847	0.2714	21.41	<2.2E-16	269
schoolsup***						
Medu***						
famsup***						
goout*						



In this model, all variables except goout are 99.9% significant, goout is 95% significant. Further improving our model, by adding sex as the next most significant variable which we got for the correlation matrix resulting in the maximum increase in the adjusted R^2 value, the resulting model is as below.

Model_5:

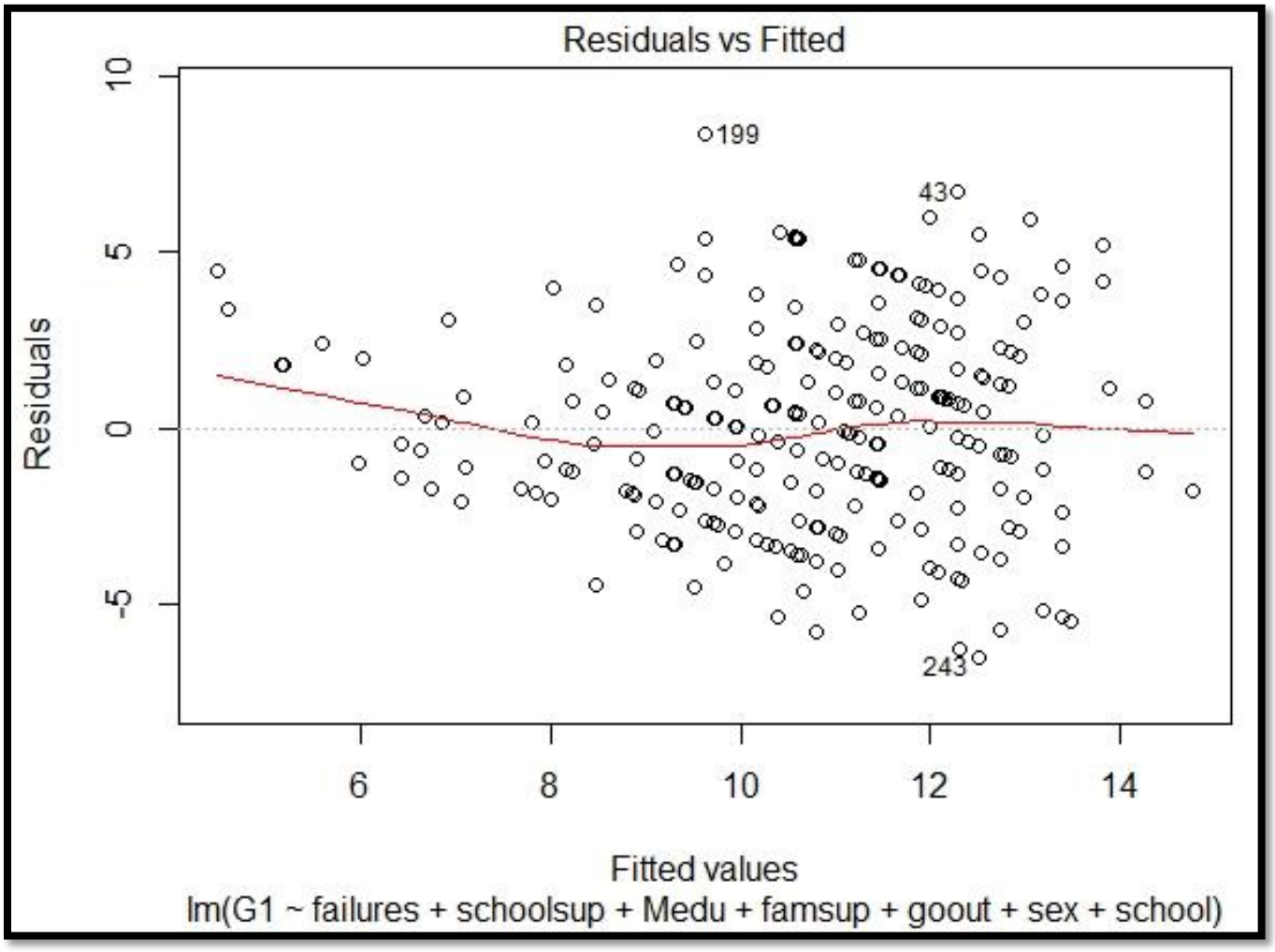
Marks = 13.7765 – 1.3876*failures – 1.8469*schoolsup + 0.6149*Medu – 1.1575*famsup – 0.4295*goout + 0.7723*sex



Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.828	0.2975	0.2818	18.92	< 2.2e-16	268
schoolsup***						
Medu***						
famsup**						
goout**						
sex*						

Variables failures, schoolsup, and Medu are 99.9% significant, famsup and goout are 99% significant, and sex, which is the latest added attribute, is 95% significant. Further improving the model by adding school to model_5, which results in the maximum increase in the adjusted R^2 value; thus, the model is given below.

Model_6:
Marks = 12.3044 – 1.3794*failures – 1.7081*schoolsup + 0.6449*Medu – 1.0928*famsup – 0.4372*goout + 0.8373*sex + 0.9364*school



Attribute	RSE	Multiple R^2	Adjusted R^2	F-stat	p-value	DOF
failures***	2.817	0.3051	0.2869	16.75	<2.2e-16	267
schoolsup***						
Medu***						
famsup**						
goout**						
sex*						
school.						

Variables failures, schoolsup and Medu are 99.9% significant, whereas famsup and gout are 99% significant, and school, i.e., the last added variable to is only 90% significant, but still an increase of ~2% is observed in adjusted R^2 which is good enough to consider in our model. Also, Model_6 explains ~31% of the total data.

On analysis, we found that adding variable age to Model_6 will still improve adjusted R^2, so the final model is as depicted in the report.

R CODE:

```
library(car)
library(corrplot)
library(psych)
library(Hmisc)
library(tidyverse)
library(caret)
library(ggthemes)
library(plotly)
library(lmtest)
library(VIF)

getwd()
setwd("E:/IIT Kanpur/IME 2nd Sem Courses/MBA652-SMBA/Project/R script")
d1 <- read.table("student-mat.csv", sep = ";", header = TRUE)
summary(d1)
attach(d1)

d1$school <- as.numeric(d1$school)
class(d1$school)

d1$sex <- as.numeric(d1$sex) # male = 2, female = 1
class(d1$sex)

d1$address <- as.numeric(d1$address) # R(rural) = 1, U(urban) = 2
class(d1$address)

d1$famsize <- as.numeric(d1$famsize) # LE3 = 1, GT3 = 2
class(d1$famsize)

d1$Pstatus <- as.numeric(d1$Pstatus) #A=1,T=2
class(d1$Pstatus)
```

```
d1$Mjob <- as.numeric(d1$Mjob) #mother's job (nominal: "teacher"=5, "health"=2 care related, civil "services"=4 (e.g. administrative or police), "at_home"=1 or "other"=3)
class(d1$Mjob)

d1$Fjob <- as.numeric(d1$Fjob) #father's job (nominal: "teacher"=5, "health" care related, civil "services"=4 (e.g. administrative or police), "at_home"=1 or "other"=3)
class(d1$Fjob)

d1$reason <- as.numeric(d1$reason) #reason to choose this school (nominal: close to "home"=2, school "reputation"=4, "course"=1 preference or "other"=3)
class(d1$reason)

d1$guardian <- as.numeric(d1$guardian) #student's guardian (nominal: "mother"=2, "father"=1 or "other"=3)
class(d1$guardian)

d1$schoolsup <- as.numeric(d1$schoolsup) #student's guardian (nominal: "mother"=2, "father"=1 or "other"=3)
class(d1$schoolsup)

d1$famsup <- as.numeric(d1$famsup) #family educational support (binary: yes=2 or no=1)
class(d1$famsup)

d1$paid <- as.numeric(d1$paid) #extra paid classes within the course subject (Math) NO=1 , YES=2
class(d1$paid)

d1$activities <- as.numeric(d1$activities) #extra-curricular activities (binary: yes=2 or no=1)
class(d1$activities)

d1$nursery <- as.numeric(d1$nursery) #attended nursery school (binary: yes=2 or no=1)
class(d1$nursery)

d1$higher <- as.numeric(d1$higher) #wants to take higher education (binary: yes=2 or no=1)
class(d1$higher)
```

```
d1$internet <- as.numeric(d1$internet) #Internet access at home (binary: yes=2 or no=1)
class(d1$internet)

d1$romantic <- as.numeric(d1$romantic) #with a romantic relationship (binary: yes=2 or no=1)
class(d1$romantic)

summary(d1)

# write.csv(d1, "dataset.csv")
# ?write.table

data_train_1 <- d1[1:243, 1:31]
data_train_2 <- d1[350:381, 1:31]
data_train <- rbind(data_train_2, data_train_1)

data_val_1 <- d1[244:278, 1:31]
data_val_2 <- d1[382:386, 1:31]
data_val <- rbind(data_val_1, data_val_2)

data_test_1 <- d1[279:349, 1:31]
data_test_2 <- d1[387:395, 1:31]
data_test <- rbind(data_test_1, data_test_2)
#rm(data_train, data_val, data_test)
rm(data_test_1, data_test_2, data_val_1, data_val_2, data_train_1, data_train_2)


regressor <- lm(formula = G1 ~ .- G1, data = data_train)
data_val$y_pred <- predict(regressor, newdata = data_val)
data_val$diff <- data_val$y_pred - data_val$G1
summary(regressor)

class(data_train$school)

correlation <- cor(data_train)
correlation_1 <- round(correlation, 2)


corrplot(correlation_1, method = "number")
?corrplot
```

```
summary(data_train)
class(school)

#########################
# scatter plots

ggplot(data_train, aes(x = G1, y = failures)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs failures", x = "Marks", y = "Failures")

ggplot(data_train, aes(x = G1, y = school)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs School", x = "Marks", y = "School")

ggplot(data_train, aes(x = G1, y = sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Sex", x = "Marks", y = "Sex")

ggplot(data_train, aes(x = G1, y = age)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Age", x = "Marks", y = "Age")

ggplot(data_train, aes(x = G1, y = address)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Address", x = "Marks", y = "Address")

ggplot(data_train, aes(x = G1, y = famsize)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs family size", x = "Marks", y = "family size")

ggplot(data_train, aes(x = G1, y = Pstatus)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```
  labs(title = "Marks vs Pstatus", x = "Marks", y = "Pstatus")

ggplot(data_train, aes(x = G1, y = Medu)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Mother's Education", x = "Marks", y = "Mother's Education")

ggplot(data_train, aes(x = G1, y = Fedu)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Father's Education", x = "Marks", y = "Father's Education")

ggplot(data_train, aes(x = G1, y = Mjob)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Mother's job type", x = "Marks", y = "Mother's job type")

ggplot(data_train, aes(x = G1, y = Fjob)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Father's job type", x = "Marks", y = "Father's job type")

ggplot(data_train, aes(x = G1, y = reason)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Reason", x = "Marks", y = "Reason")

ggplot(data_train, aes(x = G1, y = guardian)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs guardian", x = "Marks", y = "guardian")

ggplot(data_train, aes(x = G1, y = traveltime)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs traveltime", x = "Marks", y = "traveltime")

ggplot(data_train, aes(x = G1, y = studytime)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```
  labs(title = "Marks vs studytime", x = "Marks", y = "studytime")

ggplot(data_train, aes(x = G1, y = schoolsup)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs schoolsup", x = "Marks", y = "schoolsup")

ggplot(data_train, aes(x = G1, y = famsup)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs famsup", x = "Marks", y = "famsup")

ggplot(data_train, aes(x = G1, y = paid)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs paid", x = "Marks", y = "paid")

ggplot(data_train, aes(x = G1, y = activities)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs activities", x = "Marks", y = "activities")

ggplot(data_train, aes(x = G1, y = nursery)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs nursery", x = "Marks", y = "nursery")

ggplot(data_train, aes(x = G1, y = higher)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs higher", x = "Marks", y = "higher")

ggplot(data_train, aes(x = G1, y = internet)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs internet", x = "Marks", y = "internet")

ggplot(data_train, aes(x = G1, y = romantic)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```
  labs(title = "Marks vs romantic", x = "Marks", y = "romantic")

ggplot(data_train, aes(x = G1, y = famrel)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs famrel", x = "Marks", y = "famrel")

ggplot(data_train, aes(x = G1, y = freetime)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs freetime", x = "Marks", y = "freetime")

ggplot(data_train, aes(x = G1, y = goout)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs goout", x = "Marks", y = "goout")

ggplot(data_train, aes(x = G1, y = Dalc)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Dalc", x = "Marks", y = "Dalc")

ggplot(data_train, aes(x = G1, y = Walc)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs Walc", x = "Marks", y = "Walc")

ggplot(data_train, aes(x = G1, y = health)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs health", x = "Marks", y = "health")

ggplot(data_train, aes(x = G1, y = absences)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Marks vs absences", x = "Marks", y = "absences")
```

```
#model 15 - failures
model_15 <- lm(G1 ~ failures, data_train)
summary(model_15)
plot(model_15)
abline(model_15)
bptest(model_15)

# ggplot(full_data[1:3000,], aes(x = budget, y = revenue, color = budget)) +
#   geom_point() +
#
#   scale_color_viridis(begin = 0, end = .95, option = 'D') +
#   geom_smooth(method = 'lm', color = 'red3', fill = 'red3') +
#   scale_y_continuous(labels = c('$0', '$500', '$1000', '$1500')) +
#   labs(title = 'Revenue by budget', x = 'Budget', y = 'Revenue (Millions)')

#model A5 - failures + schoolsup
model_A5 <- lm(G1 ~ failures + schoolsup, data_train)
summary(model_A5)
plot(model_A5)
abline(model_A5)
bptest(model_A5)

#model A2 - failures + Medu + schoolsup
model_A2 <- lm(G1 ~ failures + Medu + schoolsup, data_train)
summary(model_A2)
plot(model_A2)
abline(model_A2)
bptest(model_A2)

# model C6 - failures + schoolsup + Medu + famsup
model_C6 <- lm(G1 ~ failures + schoolsup + Medu + famsup, data_train)
summary(model_C6)
plot(model_C6)
abline(model_C6)
bptest(model_C6)

# model D8 - failures + schoolsup + Medu + famsup + goout
model_D8 <- lm(G1 ~ failures + schoolsup + Medu + famsup + goout, data_train)
summary(model_D8)
plot(model_D8)
```

```
abline(model_D8)
bptest(model_D8)

# model E2 - failures + schoolsup + Medu + famsup + goout + sex
model_E2 <- lm(G1 ~ failures + schoolsup + Medu + famsup + goout + sex, data_train)
summary(model_E2)
plot(model_E2)
abline(model_E2)
bptest(model_E2)

# model F1 - failures + schoolsup + Medu + famsup + goout + sex + school
model_F1 <- lm(G1 ~ failures + schoolsup + Medu + famsup + goout + sex + school,
data_train)
summary(model_F1)
plot(model_F1)
abline(model_F1)
bptest(model_F1)

# model G1 - failures + schoolsup + Medu + famsup + goout + sex + school + age
model_G1 <- lm(G1 ~ failures + schoolsup + Medu + famsup + goout + sex + school +
age, data_train)
summary(model_G1)
plot(model_G1)
abline(model_G1)
bptest(model_G1)

### check for multicollinearity
car::vif(model_G1)

### heteroskedasticity check
bptest(model_G1)

############
library(olsrr)
test_model <- lm(d1$G1~.,data = d1)
olsrr::ols_step_all_possible(test_model)

ols_step_both_p(test_model, pent = 0.1, prem = 0.1, details = FALSE)
```