Naresh Gajendran

Predective model on Student marks

Table of Contents

[Introduction 2](#_Toc510729838)

[Objective 2](#_Toc510729839)

[Steps to Follow 2](#_Toc510729840)

[Descriptive statistics 3](#_Toc510729841)

[Data cleaning 3](#_Toc510729842)

[Define Dummies 3](#_Toc510729843)

[Normalisation 3](#_Toc510729844)

[Data visualisation 3](#_Toc510729845)

[Chart from top left 4](#_Toc510729846)

[Why Linear regression 4](#_Toc510729847)

[Model with all variables 5](#_Toc510729848)

[output of linear model run with significant variables 6](#_Toc510729849)

[Testing of Assumptions 6](#_Toc510729850)

[Parsimony 8](#_Toc510729851)

[Split the data 9](#_Toc510729852)

[Summary 10](#_Toc510729853)

# Introduction

This project is to work on a predictive machine learning model to predict and find the grade of the student. The reference was from the data provided along with the problem statement. The data set has an information on various information (attributes) that were recorded in association to students. We will be setting up a linear regression model to predict the total marks using other variables in the data.

# Objective

The student grades data set has information on attributes of the gender classification, this has also recorded few data from parents, ownership from student, attendance and then the topics that they were graded with the marks.

Since this is a continuous data we shall be performing a linear regression and then split the data by 70:30 then we will be training the model and then predict the values in the validation data set. We will be performing a set of steps that are part of this regression model. Post this we will be conducting the results of the model using R tool.

# Steps to Follow

We shall perform Regression Analysis in the following sequence:

1. Data Loading and Descriptive Statistics
2. Visualization
3. Running the Regression and Interpretation
4. Testing of Assumptions
   * Mean of the Residuals is zero
   * Homoscedasticity of Residuals
   * Correlation
5. Multivariate Regression
6. Robust Regression (Optional)
7. Parsimony

After the csv file is imported in R, the following steps are followed to perform as a part of creating ML in this project.

# 

# Descriptive statistics

The basic view and the outcome from the data are 16 columns and 480 observations of the data. Of these ten factors and four integers are shown.

# Data cleaning

There were three changes that were performed as

* Removing comma form mm column
* Replaced a text ‘F’ as First in Semester
* Replaced mother from ‘mom’ in parents responsible for student

Three columns were unique and they were not added in the data as ‘nationality’, ‘educational stages’ and ‘grade levels’.

# Define Dummies

Post checking the data had seven other factors in the data are to be converted to dummies as a part of regression. So the dummies were created.

# Normalisation

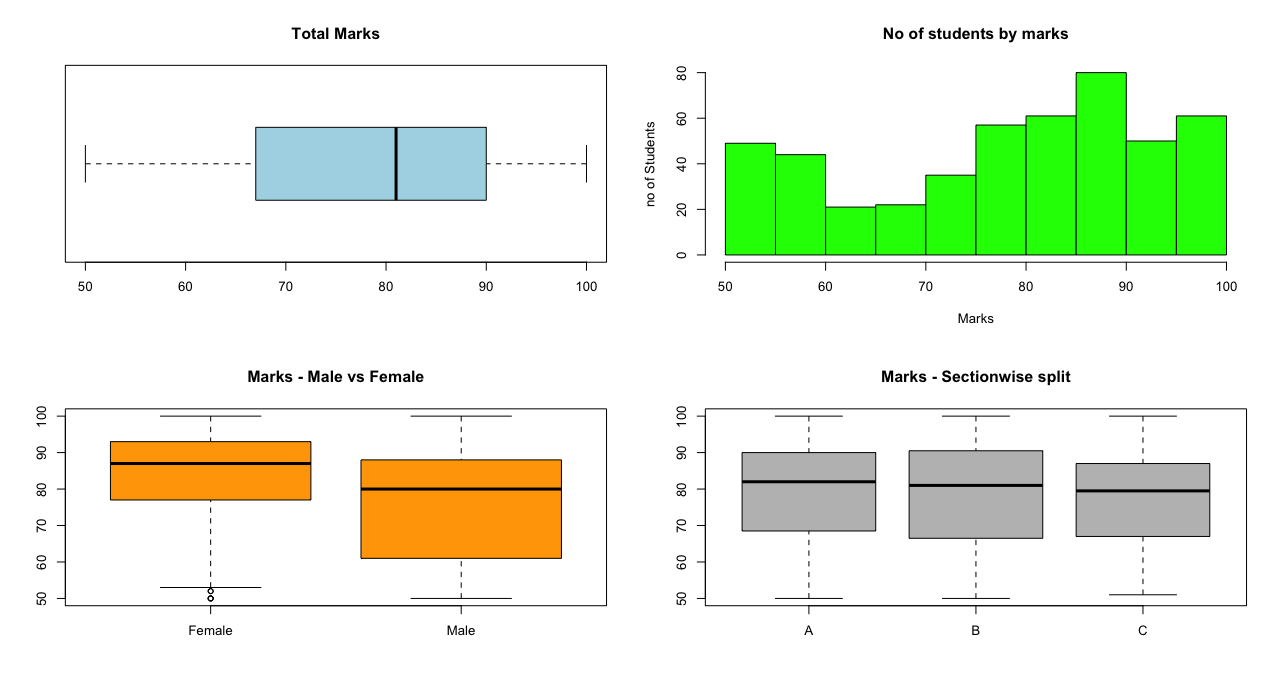
There were four columns of the data were in integer and carry a separate weightage, hence normalisation has to be performed.

All dummies and normalisation were added twenty additional columns. This will be the final data that will be used to perform linear regression.

# 

# Data visualisation

A few important charts were extracted to visualize the data.



# Chart from top left

* Chart 1 - Total marks - one of the important component and a box plot indicates between 67 to 90 as the 1st and 3rd quartile whilst the median is 80 marks.
* Chart 2 - Most of students are between 70 and 100 marks and this is not a normally distributed data
* Chart 3 - Male vs female students, most of the female students fall between 78 and 94 marks category and most of the male students fall between 60 and 85. Female students have the higher marks.
* Chart 4 – section wise split, Section B students are better than the other two sections
  + - Section A has most of them between 68 and 90 marks
    - Section B has most of them between 66 and 91 marks
    - Section C has most of them between 68 and 87 marks

# Why Linear regression

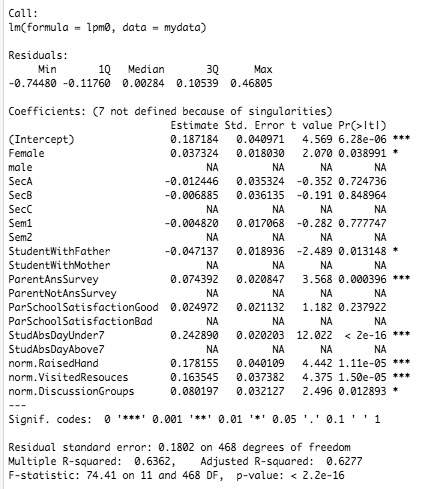
As the project target to predict the scores for the students and this is derived by total marks. This total marks are of continuous data hence linear regression is the only possibility with this option.

## Model with all variables

The below is the output of adding all the variables to the lm model and the model output is shown the below

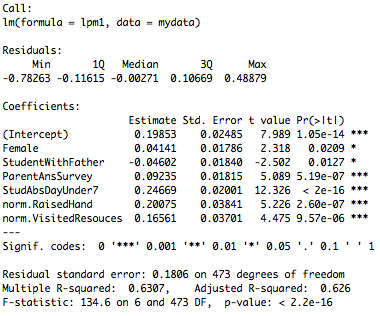
Seven of the coefficients are not defined because of singularities and this is because the variables are perfectly collinear. So these will be removed.

The output also shows the variables that are significant and do contribute strongly for the total marks. Hence the model is run again with the variables which have more than one star at the end of each variable output.



## output of linear model run with significant variables

This is the output which is run to get the optimal output from the significant variables.



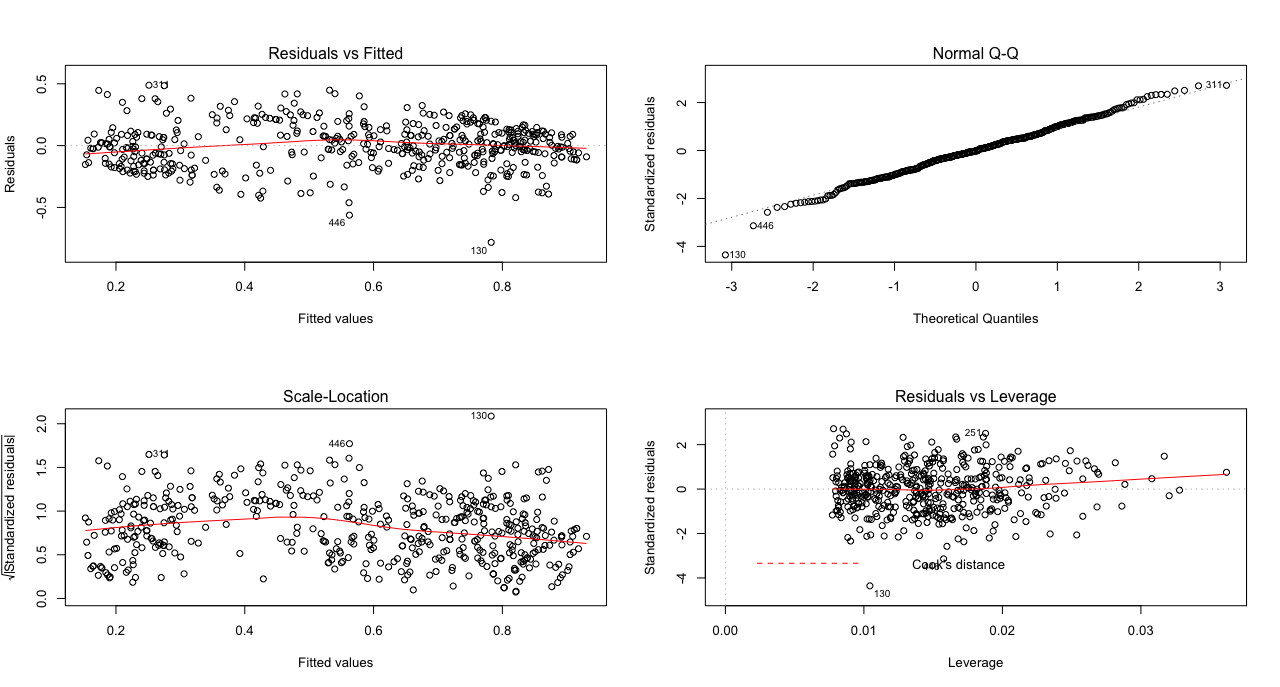
This is a multivariate analysis which explains close to 62.6% variance in the total marks. And as single variables are positive but not seem to have a major impact.

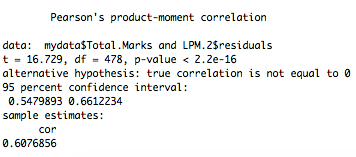
## 

## Testing of Assumptions

> mean(LPM.2$residuals)

[1] 4.050173e-18



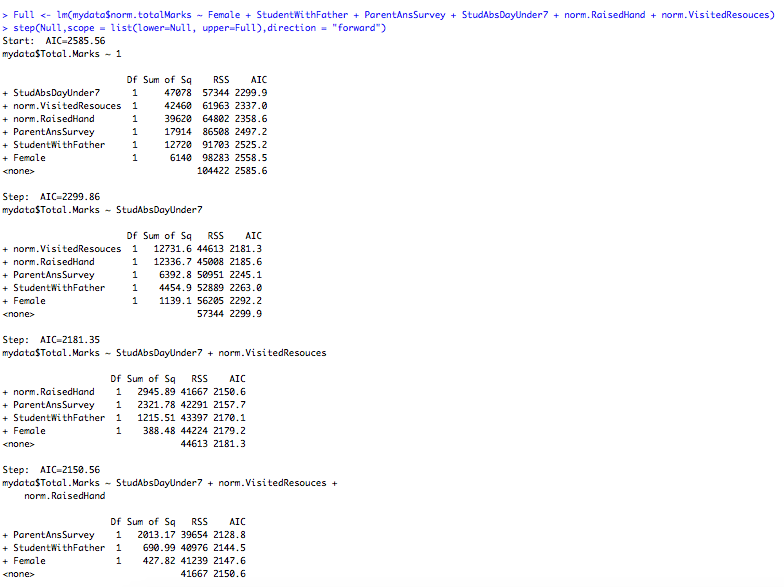


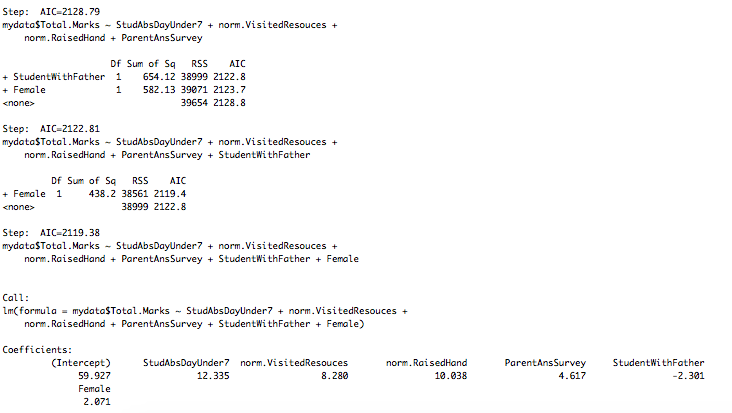
The above charts demonstrate the model has pushed it to optimal scale to maximum fit the line straight.

Interpretation:

* The assumption of mean of Residuals is zero holds True.
* The Residuals plot confirms normal linear regression model.
* The Q-Q Plot confirms that our dependent variable is normally distributed.
* The Scale Location graph confirms homoscedasticity in the data.
* The Residuals Vs Leverage graphs shows some extreme Cooks distance lines, which suggests impactful outliers present in the data.
* This recommends to go for Robust Regression Model.
* Correlation between Errors and Explanatory Variables:
  + P Value is greater than 0.05, which suggests that Errors and Explanatory Variables are not correlated.

# Parsimony





so the coefficients are

Coefficients:

(Intercept) StudAbsDayUnder7 norm.VisitedResouces norm.RaisedHand

59.927 12.335 8.280 10.038

ParentAnsSurvey StudentWithFather Female

4.617 -2.301 2.071

now that the model coefficients are set, these are inputted in the model by training and modelling the data

# Split the data

The data is split as 70:30 now and the model is trained and the trained model is validated in the validation data set.

The prediction is done on the logic that was built on linear regression.

The scores that were predicted are matched against the actual total marks and a MAPE is calculated on the differences. The model was run to find the optimistic score and this score is arrived to be 73%

# Summary

The following regression model can be set up to explain student marks using the contributing factors and this model can be used to predict the marks in the future as well.