Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #5: (REGRESSION MODELLING)

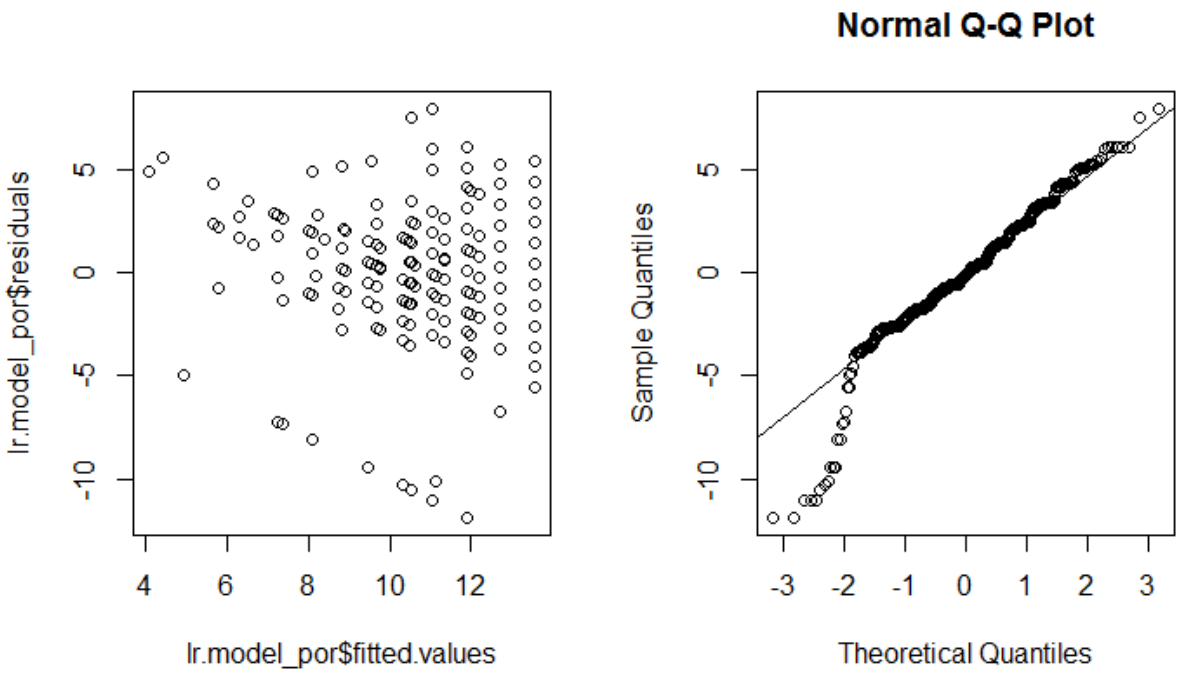
***PURPOSE*:** In this assignment, we use the regression modelling tools to build predictive models for Portuguese student dataset and Automobile mpg dataset.

DATASETS: Automobile MPG Dataset (Link) and Student Performance Dataset (Link).

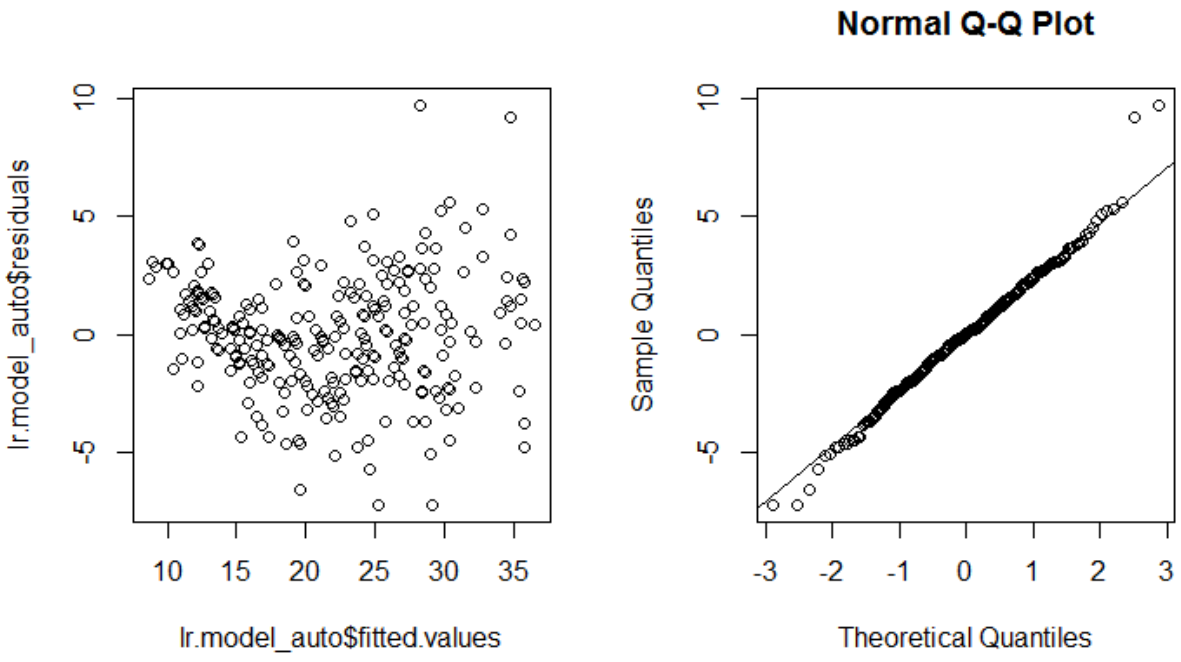
APPROACH:

- ***Portuguese language - Student Dataset –*** This dataset gives the information about the student achievement in the secondary education. In this approach, we are using Portuguese language dataset.
- The data set has 33 variables, where G3 is the target attribute and we find the regression model from other attributes and see how well the model will predict the G3 (final year grade).
- We see that there are many of character variables. So, we need to change all into factors.
- First, we build a linear model for G3 with all other variables. The P-values for G1 and G2 are very small (<0.001) compared to others who P-value is big. The R^2 value is 0.86. The qqnorm plot of the residuals looks good but has some outliers that is there are only few students who got low and high scores.
- From the previous model, we seen that there are many useless variables and we removed them and again build a linear model. Now it seems slightly improved but not much to say a good model. Now R^2 value is 0.8512.
- Now, new model is built with few variables (school, sex, failures, schoolsup, higher) and without first and second period grades (G1 and G2), we observed that all variables have very small P-value but the R^2 value is very small (0.2808)(see graph 1).
- ***Auto MPG Dataset –*** This dataset has the information about city cycle fuel consumption in miles per gallon. This has 9 variables where 'mpg' is the target attribute. We try to build predictive model for it.
- The variable 'car name' is a string which is a unique value. So, we build a model without taking it into the account.
- When we build the predictive model for 'mpg' with the other variable removing 'name car'. The model seems good but some variables have a big p-value but R^2 value is 0.85.
- The variables 'horsepower' and 'acceleration' are not good for predicting mpg. So, we build model without these.
- The plot between the residuals and fitted values are shown in graph 2 and the qqnorm of the residuals are shown.

GRAPHS:



(Graph 1)



(Graph 2)

SUMMARY:

- In this assignment, we do the regression analysis to find the relationship among the other variables and how good are they in predicting the target attribute.
- For the Student Dataset, G3 is the response variable. By applying the lm function with G3 as response variable and other variables as predictors, we seen that some of the variables are not good (with p-value very high). So, some of the variables are only predict well with G3.
- For the Automobile dataset, mpg is the response variable. The car name variable has many names so it may create more dummy variables. So, we tried using other variables in predicting mpg of an automobile.