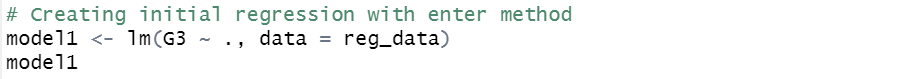
Q.1>

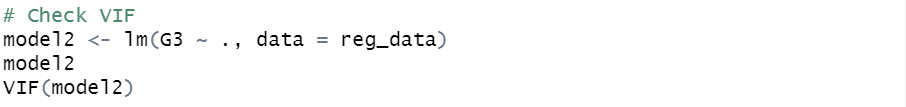
1. There are 30 variables in the dataset. Out of which G3 is a dependent and a numeric variable. So, if we exclude the dependent variables, then 29 independent variables are left in the dataset. The requirement for regression is that the variables should be numeric irrespective of the fact whether the variables are categorical or continuous. In this case, there are 16 numeric variables within the dataset which are age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2. Hence, only the above-mentioned variables are usable at the initial step. But after some preprocessing the variables and converting the categorical variables into numeric variables, those variables can also be used in the regression model. The main reason why a string variable needs to be converted into numeric is that linear regression uses a 2d plane to represent the line of regression. On this plane, only numbers can be represented, and strings cannot be represented. Hence, the linear multiple regression required the variables to be numerical.
2. The assumptions that I made are there are three or more variables involved, there should be no major outliers, relationships between variables are linear and additive, there should be no autocorrelation, there should be no multicollinearity, data has equal variances which is also known as homoscedasticity, and residual variables should have a normal distribution
3. Creating initial regression with the enter method



A picture containing table

Description automatically generated

Check VIF



Text

Description automatically generated

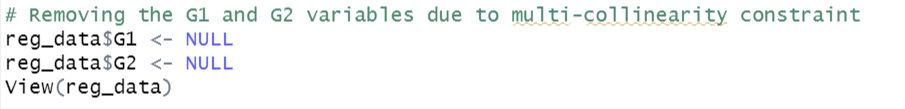
Correlation Plot



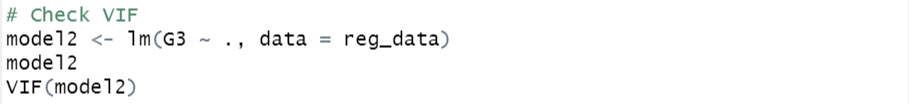
Chart

Description automatically generated

Removing the G1 and G2 variables due to the multi-collinearity constraint



Check VIF for model2



A screenshot of a computer

Description automatically generated with medium confidence

1. Creating the diagnostic plot of model1



Chart

Description automatically generated

Creating the diagnostic plot of model2



Chart, scatter chart

Description automatically generated

Summarizing both the models



Table

Description automatically generated

Table

Description automatically generated



1. Stepwise Multiple Linear Regression

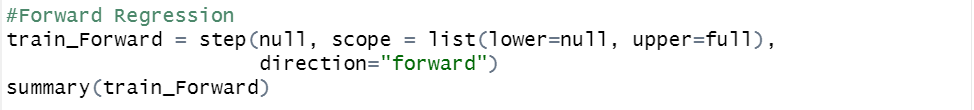
Text

Description automatically generated with low confidence

Table

Description automatically generated

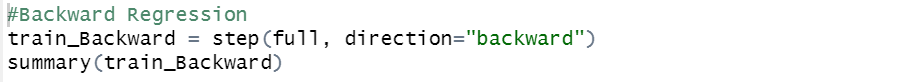
Forward regression



Table

Description automatically generated

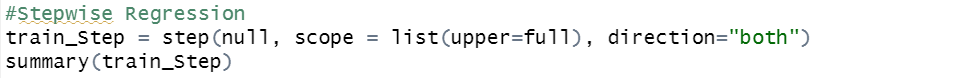
Backward regression



Text, table

Description automatically generated with medium confidence

Stepwise regression



A screenshot of a computer

Description automatically generated with medium confidence

Q.2>

1. LASSO (Least Absolute Shrinkage and Selection Operation) regression is one of three methods of regularized regression used to correct multicollinearity and overfitting. Lasso regression optimizes the variable selection as well as the manual/automatic regression. Unlike manual or automatic regression models, it reduces residual error, seeks sparse solutions, and does feature selection.
2. Preparing the lasso regression model

A picture containing chart

Description automatically generated

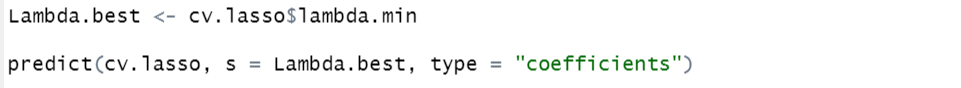
Text

Description automatically generated

A picture containing chart

Description automatically generated

The lasso regression plot shows the confidence interval of 13 various models. The confidence interval is represented by two vertical dotted lines which are between 1 and 11.



Graphical user interface, text, application

Description automatically generated

Only variables are selected for the regression model which are age, Medu, traveltime, failures, gout.

1. On this dataset, we discover that: Not all predictors are required to locate the final period marks when using the Lasso Regression model. The problem of overfitting the model can be mitigated via regularization. The following are the five elements on which the final period grades are heavily reliant: The age of the student, Education of the mother, Time spent traveling from home to school, Number of failed classes in the past, Going out with pals regularly.
2. The result of the lasso regression is as follows



Text

Description automatically generated

Yes, the results of the lasso regression are different from the results of step-wise regression. The results of the lasso regression are more significant because the p-value of lasso regression is greater than the p-value of stepwise regression.