Assignment 1

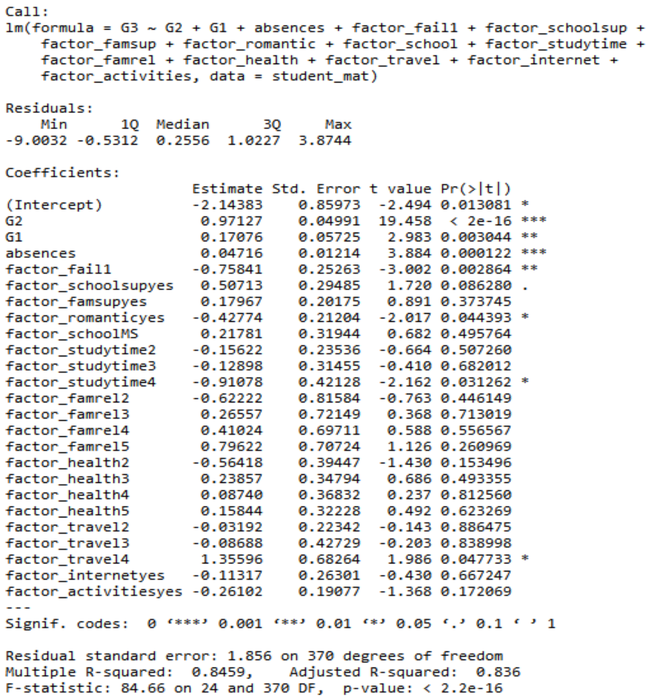
Group Members: Varun Parthasarathy, Shanmukh Kali Prasad, Siddarth Suresh Gopalakrishnan, Rahul Shevade, Siddharth Sampath, Vamshi Kasam, Dev Gupta, SLN Vashist, Pranav Itapu

Q1. Run an econometric regression model on your dataset giving proper justification for selection of the variables. Interpret the coefficients of your variable appropriately. [Marks will only be given if you give correct justification for variable and interpretation for variables]

The results of the OLS Regression are given below. We used the command given below to run the required regression.

```
ols = lm(G3 ~ G2 + G1 + absences + factor_fail1 + factor_schoolsup +
factor_famsup + factor_romantic + factor_school + factor_studytime +
factor_famrel + factor_health + factor_travel + factor_internet +
factor_activities, data = student_mat)
summary(ols)
```

Results:



```
Call:
lm(formula = G3 ~ G2 + G1 + absences + factor_fail1 + factor_schoolsup +
    factor_famsup + factor_romantic + factor_school + factor_studytime +
    factor_famrel + factor_health + factor_travel + factor_internet +
    factor_activities, data = student_mat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0032 -0.5312  0.2556  1.0227  3.8744

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.14383    0.85973  -2.494 0.013081 *
G2                    0.97127    0.04991  19.458  < 2e-16 ***
G1                    0.17076    0.05725   2.983 0.003044 **
absences              0.04716    0.01214   3.884 0.000122 ***
factor_fail1         -0.75841    0.25263  -3.002 0.002864 **
factor_schoolsupyes   0.50713    0.29485   1.720 0.086280 .
factor_famsupyes      0.17967    0.20175   0.891 0.373745
factor_romanticyes   -0.42774    0.21204  -2.017 0.044393 *
factor_schoolMS       0.21781    0.31944   0.682 0.495764
factor_studytime2    -0.15622    0.23536  -0.664 0.507260
factor_studytime3    -0.12898    0.31455  -0.410 0.682012
factor_studytime4    -0.91078    0.42128  -2.162 0.031262 *
factor_famrel2       -0.62222    0.81584  -0.763 0.446149
factor_famrel3        0.26557    0.72149   0.368 0.713019
factor_famrel4        0.41024    0.69711   0.588 0.556567
factor_famrel5        0.79622    0.70724   1.126 0.260969
factor_health2       -0.56418    0.39447  -1.430 0.153496
factor_health3        0.23857    0.34794   0.686 0.493355
factor_health4        0.08740    0.36832   0.237 0.812560
factor_health5        0.15844    0.32228   0.492 0.623269
factor_travel2       -0.03192    0.22342  -0.143 0.886475
factor_travel3       -0.08688    0.42729  -0.203 0.838998
factor_travel4        1.35596    0.68264   1.986 0.047733 *
factor_internetyes   -0.11317    0.26301  -0.430 0.667247
factor_activitiesyes -0.26102    0.19077  -1.368 0.172069
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.856 on 370 degrees of freedom
Multiple R-squared:  0.8459,    Adjusted R-squared:  0.836
F-statistic: 84.66 on 24 and 370 DF,  p-value: < 2.2e-16
```

The variables significant at the 95% confidence interval are:

G2, G1, absences, factor_fail1, factor_schoolsupyes, factor_romanticyes, factor_studytime4, factor_travel4.

G2 having a significant coefficient of 0.97 implies if a student increases his/her score by 1 mark in the second quiz, they will do better in their third quiz with a factor of 0.97 marks in G3 for each extra mark scored in G2.

G1 having a significant coefficient of 0.17076 implies if a student increases his/her score by 1 mark in the second quiz, they will do better in their third quiz with a factor of 0.17076 marks in G3 for each extra mark scored in G2.

The variable ‘absences’ is shown to have a positive coefficient of 0.047 in the results. In actual terms, it would mean for every extra day of leave taken by the student, he/she will score 0.04 marks more in the third quiz compared to those who go to class. However, this goes against the general idea based on intuition.

The variable ‘factor_fail1’ is a custom dummy which takes value 0 is a student hasn’t failed and 1 if a student has failed. The results show that if a student has failed earlier, he/she will get 0.75 marks lesser in the third quiz compared to those who have never failed before.

The results of the OLS regression also show a significant positive coefficient of 0.507 for the variable ‘factor_schoolsupyes’. This implies students who have extra educational support will score 0.5 marks more in the third quiz compared to those who don’t.

The dummy variable which shows if a student has a romantic interest or not, factor_romanticyes, has a negative significant coefficient of -0.427. It implies that those involved in a romantic relationship will be scoring 0.4 marks less in the third quiz compared to single students.

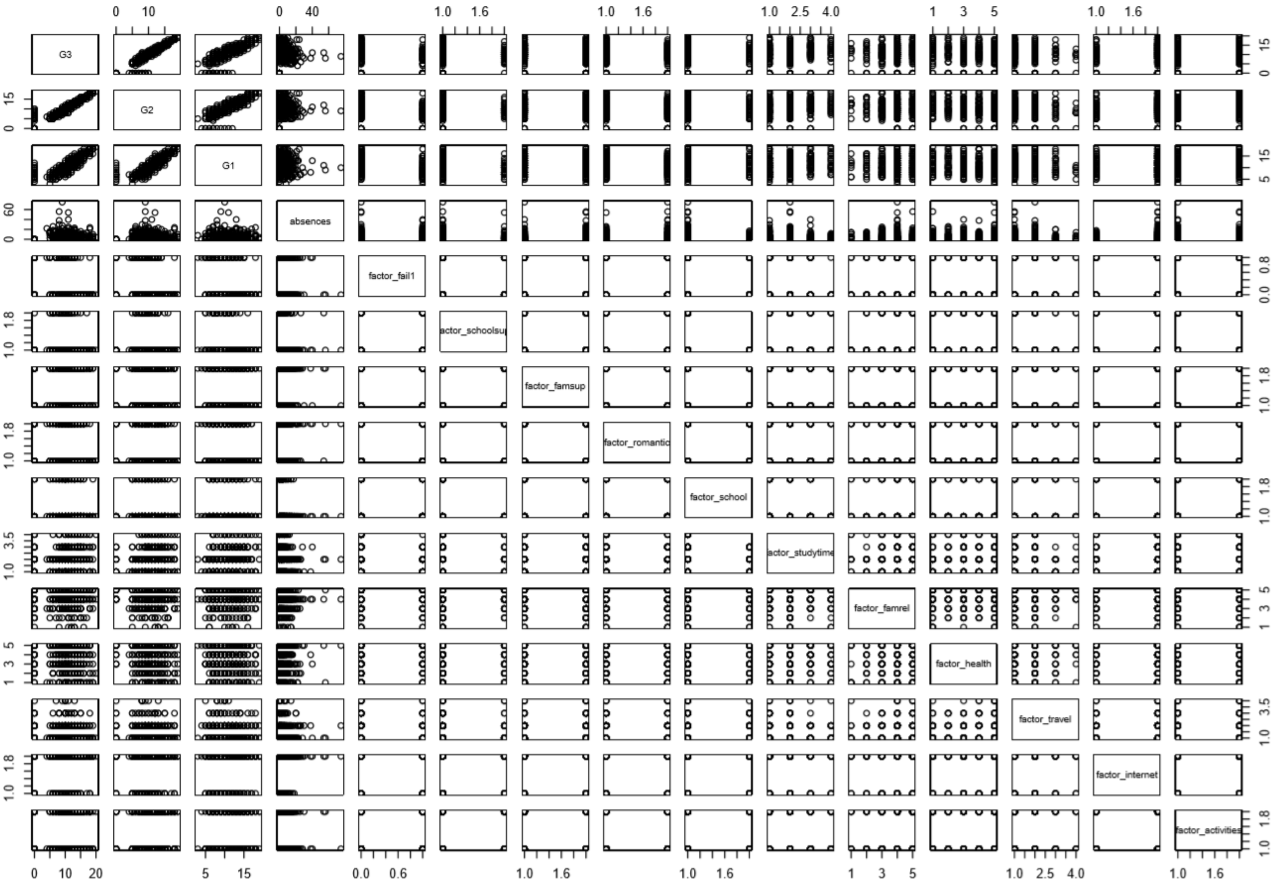
The next variable having a significant coefficient is factor_studytime4. This is a dummy which takes value 0 if the study time is less than 4 and 1 if the study time is 4. The negative coefficient of -0.9 would indicate a student would score less in quiz 3 if he studies for 4 hours compared to his score if he had studied for one hour only. Although this goes against general intuition, the results gave positive insignificant coefficients which showed students who study for 2 or 3 hours will score more than those who study for one hour only.

The last significant variable in the result is factor_travel4. This is a dummy variable which takes the value 1 for students who have to travel for more than one hour to get to school. The positive significant coefficient of 1.35 says that the student will score on an average 1.35 more marks in quiz 3 if he travels for more than 1 hour to go to school compared to those who travel for less than 15 minutes. One explanation could be that people who travel more would be more willing to work hard and thus do well in their quiz.

Q2. Prepare Graph Matrix for your dataset. Comment on the association between dependent variable & independent variables of your dataset.

To prepare the graph matrix, we used the following command(s).

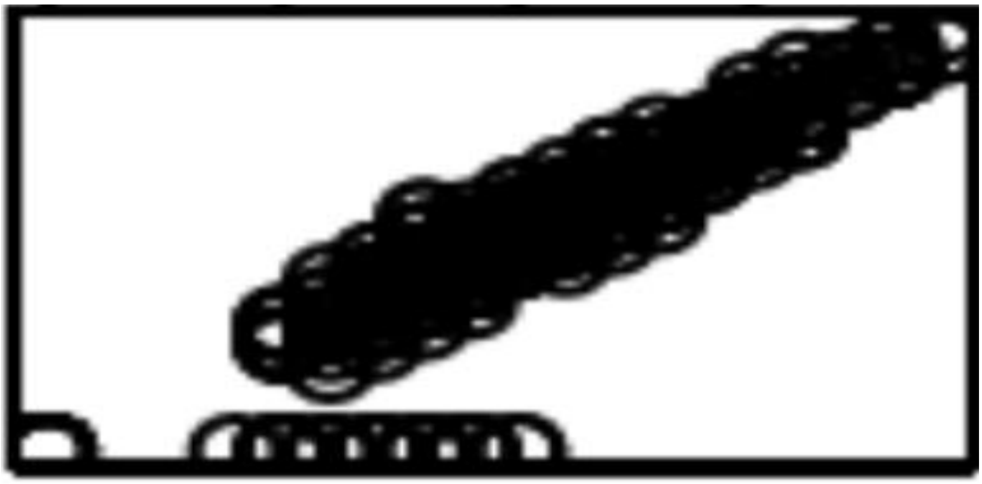
```
columns = c("G3", "G2", "G1", "absences", "factor_fail1", "factor_schoolsup",
"factor_famsup", "factor_romantic", "factor_school", "factor_studytime",
"factor_famrel", "factor_health", "factor_travel", "factor_internet",
"factor_activities")

pairs(student_mat[,columns])

summary(ols)
```



The remaining variables are dummy variables due to which we can observe that the observations are more or less uniformly distributed among certain strata/levels.

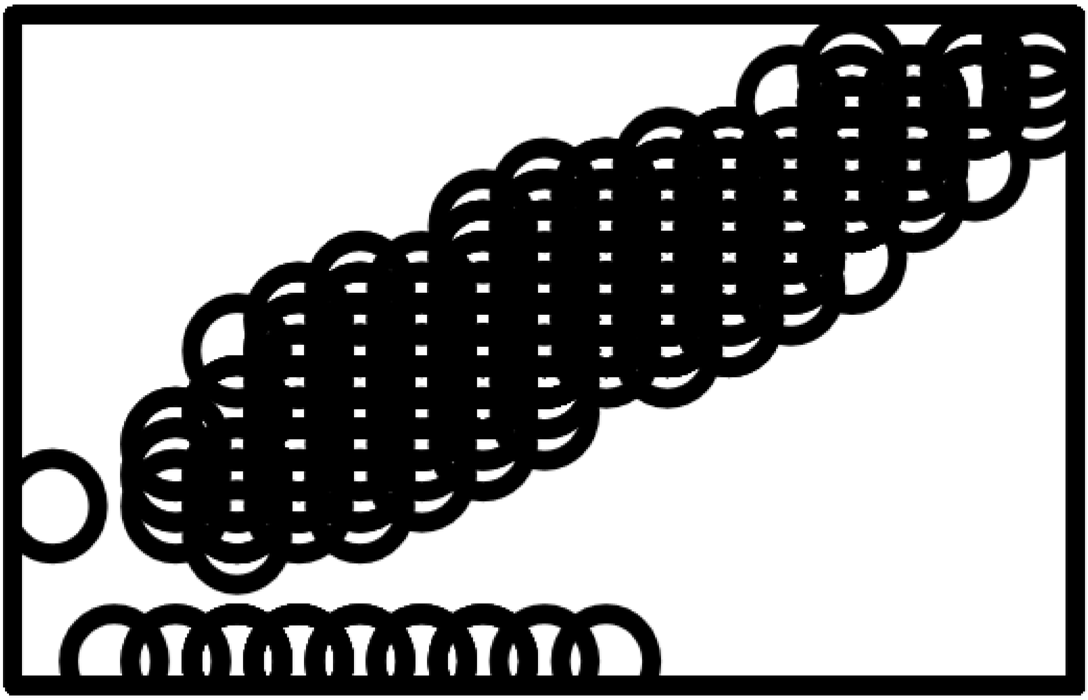
Relation between continuous independent variables G1, G2 and dependent variable G3.

G3 (Y-axis) vs G2 (X-axis).



We can see that G3 has a positive relation with G2, i.e. as a student's G2 scores increase the G3 scores are also projected to increase.

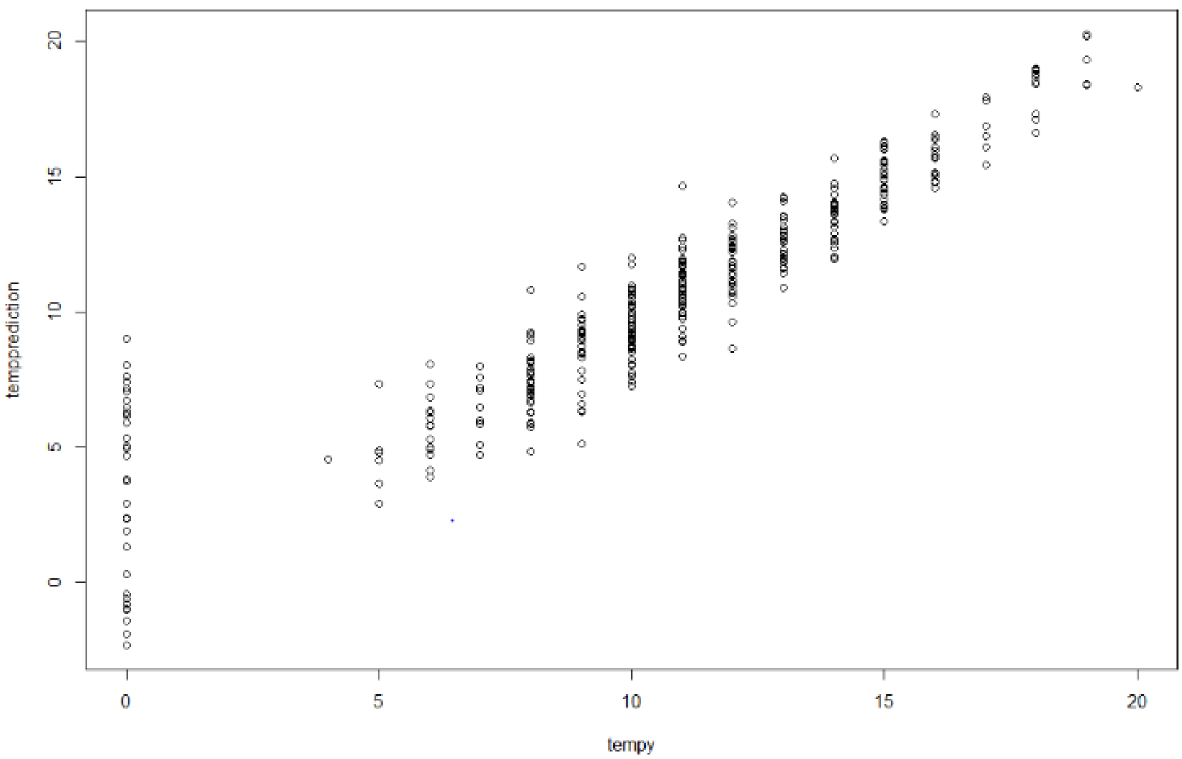
G3 (Y-axis) vs G1 (X-axis)



Similar to the previous results, we have a positive relationship between a student's G3 scores and G1 scores, i.e. as a student's G1 score increases G3 scores are also predicted to increase.

Q3. Plot the predict Y and discuss the accuracy of your model.

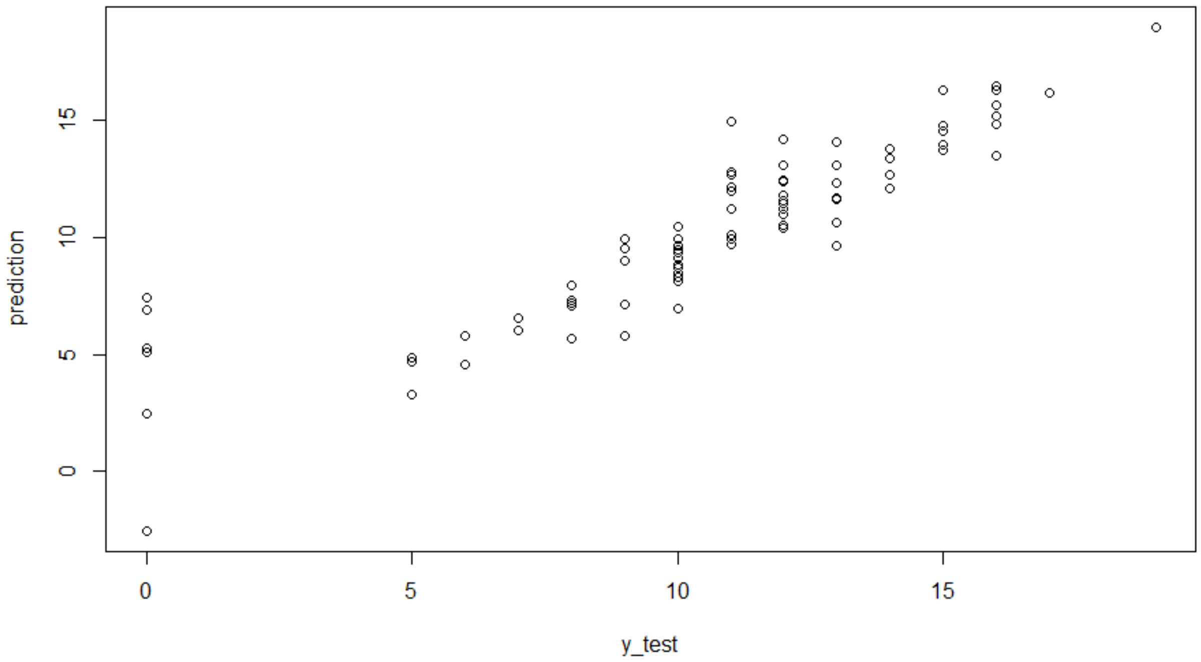
The R squared value of our regression is 0.84. The R squared value represents the variance of the dependent variable explained by our model and ranges from zero to 1. In essence, it represents a measure of how well the model fits the given data.



OLS Estimates v/s Actual Dependent Variables

Another way of interpreting accuracy is prediction accuracy, where a model is used to estimate unknown data based on pre-existing data. We approached this by splitting the data into training (80%) and testing sets (20%) and calculating the RMSE (root mean squared errors).

RMSE: 1.930958



Testing set predictions vs Actual Values

Q4. For regression model fit in Question 1, run the tests for checking following OLS assumptions and interpret your results.

a. Heteroscedasticity

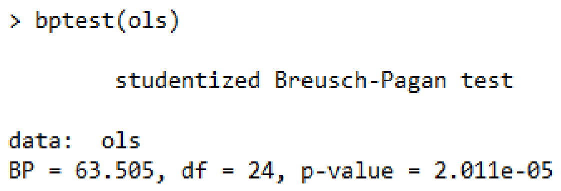
b. Multicollinearity

c. Normality of the error term

d. Omitted-Variable Bias

a) BP (Breusch Pagan)

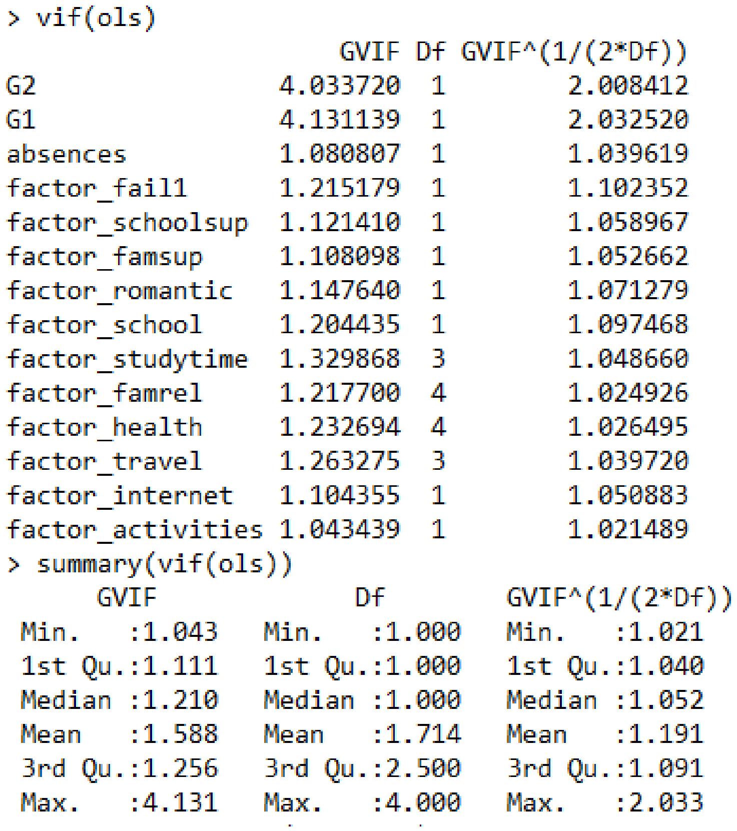
Null Hypothesis: Homoscedasticity



```
> bptest(ols)

        studentized Breusch-Pagan test

data:  ols
BP = 63.505, df = 24, p-value = 2.011e-05
```

Very low P value indicates that null hypothesis is rejected, therefore heteroscedasticity.

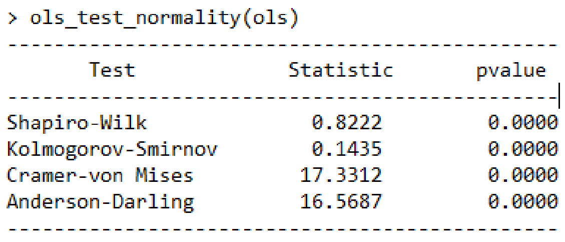
b) VIF test (Variance Inflationary Factor)



```
> vif(ols)
                       GVIF Df GVIF^(1/(2*Df))
G2                 4.033720  1        2.008412
G1                 4.131139  1        2.032520
absences           1.080807  1        1.039619
factor_fail1       1.215179  1        1.102352
factor_schoolsup   1.121410  1        1.058967
factor_famsup      1.108098  1        1.052662
factor_romantic    1.147640  1        1.071279
factor_school      1.204435  1        1.097468
factor_studytime   1.329868  3        1.048660
factor_famrel      1.217700  4        1.024926
factor_health      1.232694  4        1.026495
factor_travel      1.263275  3        1.039720
factor_internet    1.104355  1        1.050883
factor_activities  1.043439  1        1.021489
> summary(vif(ols))
      GVIF              Df        GVIF^(1/(2*Df))
 Min.   :1.043   Min.   :1.000   Min.   :1.021  
 1st Qu.:1.111   1st Qu.:1.000   1st Qu.:1.040  
 Median :1.210   Median :1.000   Median :1.052  
 Mean   :1.588   Mean   :1.714   Mean   :1.191  
 3rd Qu.:1.256   3rd Qu.:2.500   3rd Qu.:1.091  
 Max.   :4.131   Max.   :4.000   Max.   :2.033  
```

All the VIFs are significantly less than 10, which happens to be our threshold for multicollinearity. Therefore, there’s no multicollinearity.

c)

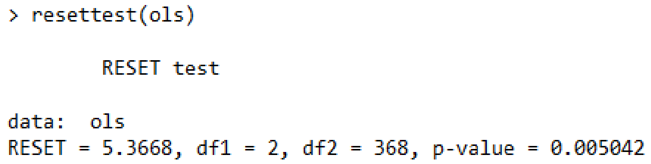


```
> ols_test_normality(ols)
-----------------------------------------------
       Test             Statistic       pvalue  
-----------------------------------------------
Shapiro-Wilk              0.8222         0.0000 
Kolmogorov-Smirnov        0.1435         0.0000 
Cramer-von Mises         17.3312         0.0000 
Anderson-Darling         16.5687         0.0000 
-----------------------------------------------
```

Null hypothesis: Normality of error term holds.

Very low p-value indicates rejection of the null hypothesis. Normality condition does not hold.

d)



```
> resettest(ols)

	RESET test

data:  ols
RESET = 5.3668, df1 = 2, df2 = 368, p-value = 0.005042
```

Null hypothesis: No omitted variables.

P-value indicates rejection of the null hypothesis

5. Based on results of Question 4, use the remedies to address the issues identified and alter your model suitably.

a) Multicollinearity: No multicollinearity.

b) Omitted Variable Bias: Our model does suffer from this. Some possible variables that could be included are student motivation index, IQ level, student’s major etc.

c) Heteroscedasticity: Our model suffers from this. To rectify this, we used the command:

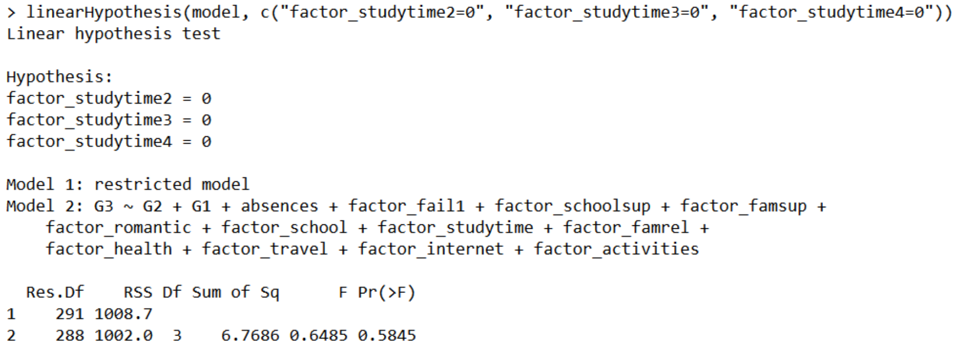
coeftest(ols, vcov = vcovHC(ols, "HC0")) from package ***“sandwich”.***

This command runs the regression assuming the variances of the error terms are constant.

d) Normality: Our model failed all four tests of normality. (refer above) In order to rectify this, we ran our model by considering logarithmic transformations of our dependent variable and certain permutations of the independent ones, but they failed the test as well. One reason for the failure of this test could be the vast majority of categorical variables in a linear regression or the limited amount of data available. Some other possible solutions could be to run a logistic regression model or an ordinal regression model.

6. For your model, run two joint tests (F-test) giving justification for the same. Interpret the results.

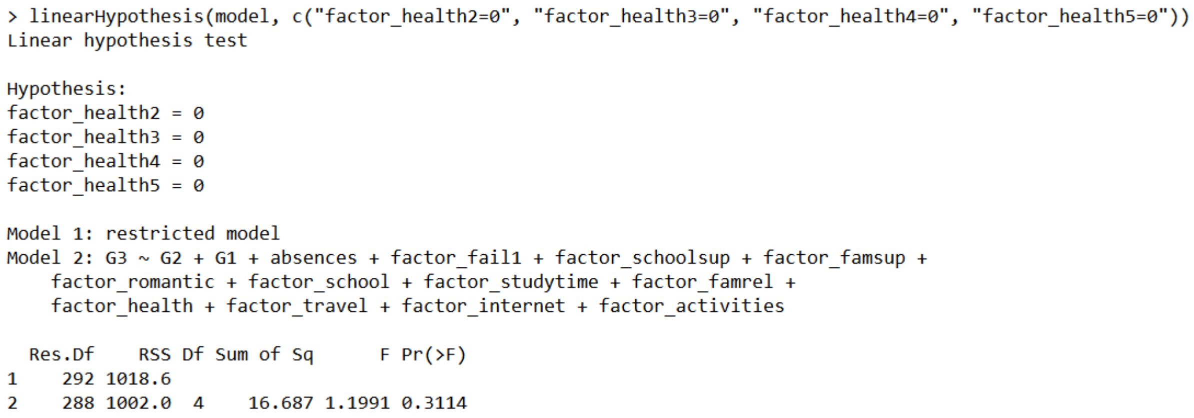
First F test was to see if the studytime factor variables were all zero or not.



```
> linearHypothesis(model, c("factor_studytime2=0", "factor_studytime3=0", "factor_studytime4=0"))
Linear hypothesis test

Hypothesis:
factor_studytime2 = 0
factor_studytime3 = 0
factor_studytime4 = 0

Model 1: restricted model
Model 2: G3 ~ G2 + G1 + absences + factor_fail1 + factor_schoolsup + factor_famsup +
    factor_romantic + factor_school + factor_studytime + factor_famrel +
    factor_health + factor_travel + factor_internet + factor_activities

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    291 1008.7
2    288 1002.0  3    6.7686 0.6485 0.5845
```

Looking at the P value of the F test we fail to reject the null hypothesis at a 95% confidence interval as it is greater than 0.05, which implies that the coefficients of the studytime factor variables could be all zero jointly.

Second F test was to see if the health factor variables were all zero or not.



```
> linearHypothesis(model, c("factor_health2=0", "factor_health3=0", "factor_health4=0", "factor_health5=0"))
Linear hypothesis test

Hypothesis:
factor_health2 = 0
factor_health3 = 0
factor_health4 = 0
factor_health5 = 0

Model 1: restricted model
Model 2: G3 ~ G2 + G1 + absences + factor_fail1 + factor_schoolsup + factor_famsup +
    factor_romantic + factor_school + factor_studytime + factor_famrel +
    factor_health + factor_travel + factor_internet + factor_activities

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    292 1018.6
2    288 1002.0  4    16.687 1.1991 0.3114
```

Looking at the p value of the F test we fail to reject the null hypothesis at a 95% confidence interval as it is greater than 0.05, which could mean that the factor health variables are all zero jointly.