**Statistical Learning Lab**

**Assignment - 1**

**Linear Regression Assignment**

***Submitted by,***

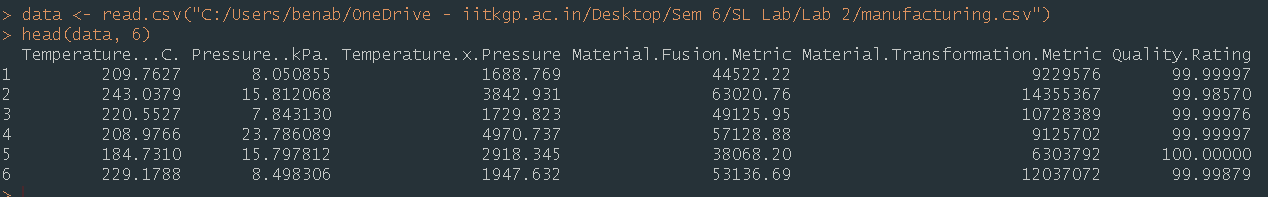
***Ben Abraham Biju, 22IM10048***

**Show the code snippets and the corresponding output for the following:**

1. **Load the dataset “manufacturing.csv”. Display first few rows of the dataset.**

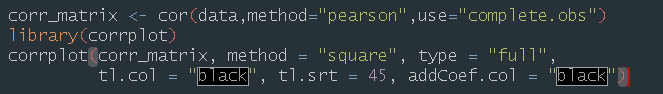
The csv file is read using the ***read.csv(..)*** function by giving the path name of the file and the first 6 rows are displayed.



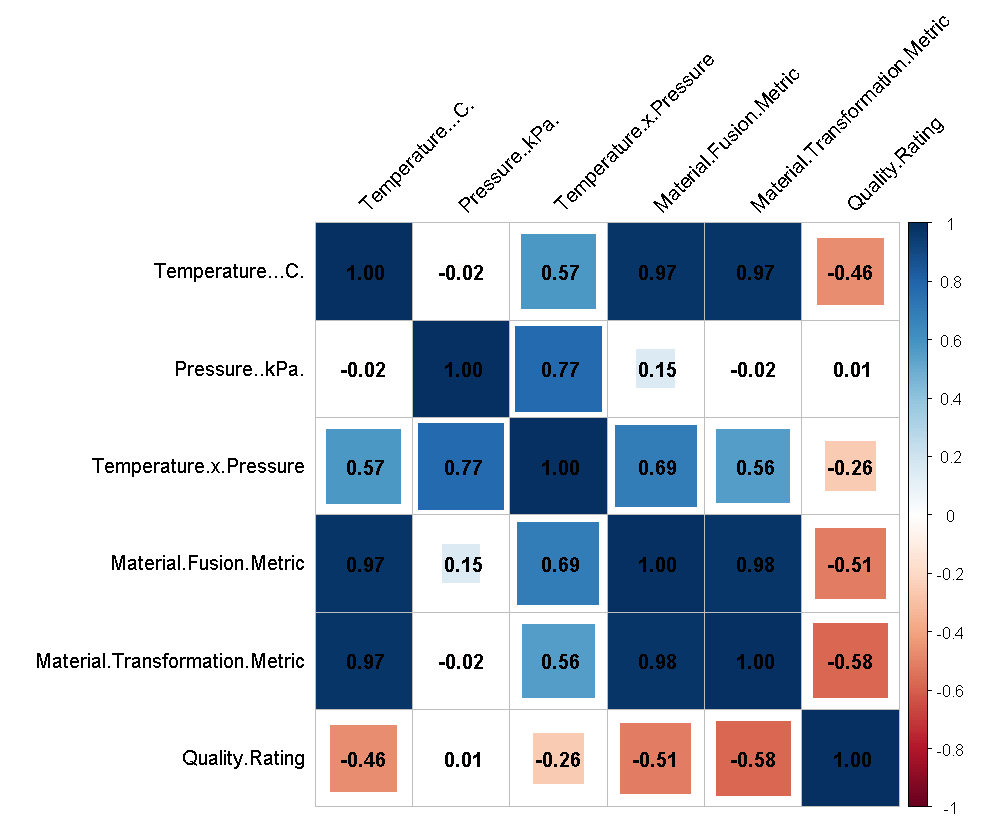


1. **Perform matrix plot and correlation analysis indicate if there is any correlation among the predictors**

The correlation matrix is generated using the ***cor(…)*** function and is plotted using the ***corrplot*** library.



The following plot was observed:



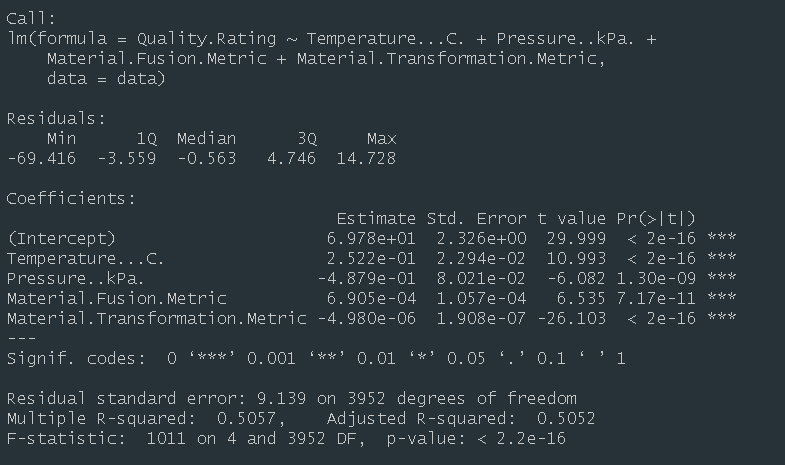
From this correlation matrix plot, we see the following pairs to have considerable correlation:

1. Temperature…C. and Material.Fusion.Metric : 0.97
2. Temperature…C. and Material.Transformation.Metric : 0.97
3. Material.Transformation.Metric and Material.Fusion.Metric : 0.98
4. **Fit a Linear Regression model without the interaction term. From the linear regression summary which factors seem to be significant?**

A model was trained to predict the Quality Rating based on Temperature, Pressure, Material Fusion Metric and Material Transformation Metric.

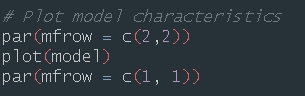


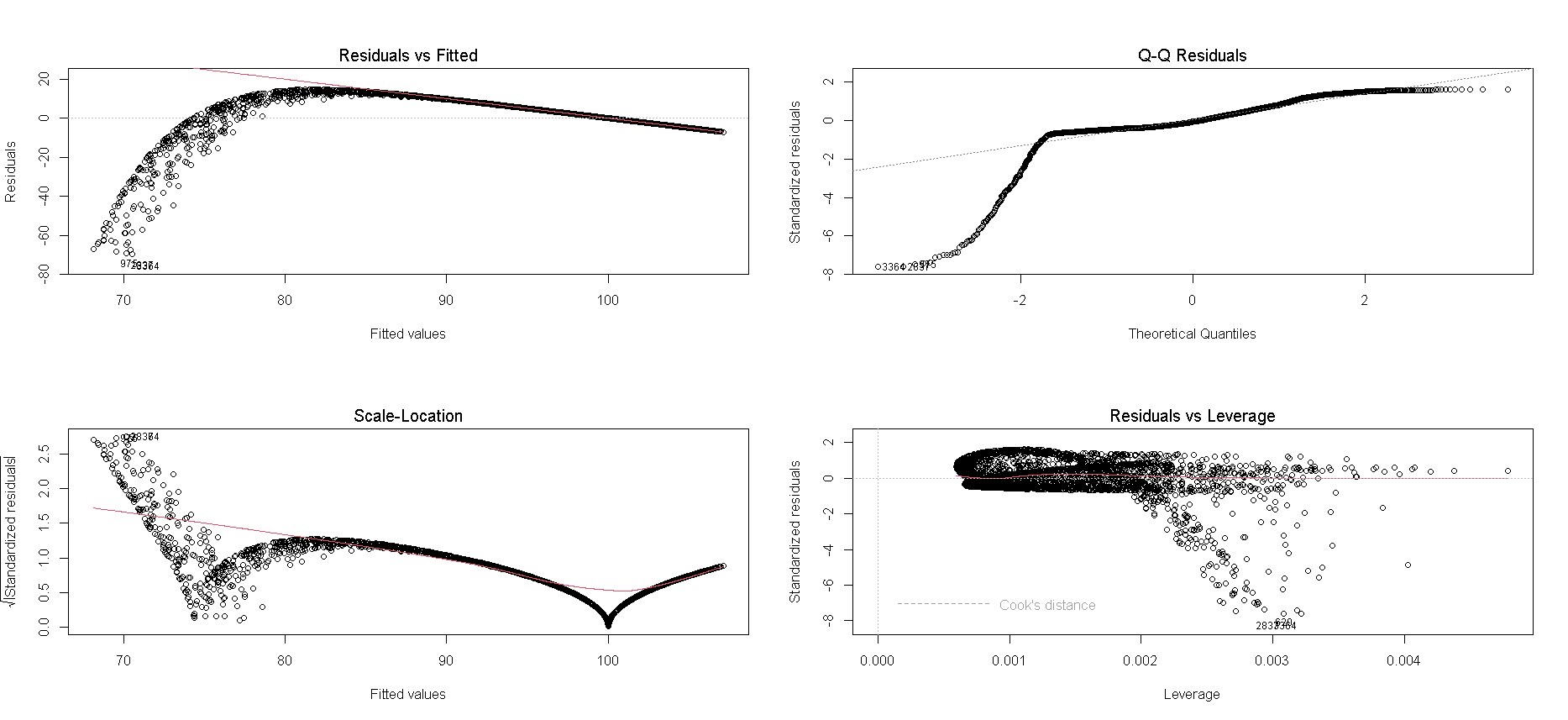
Summary of model:



From the summary, all the factors seem to be significant, since the p-value is significantly less than 0.05, at 5% significance level.

On plotting the model characteristics, the following graphs were observed.





The residuals vs fitted graph shows a curve, indicating that the relationship might be non-linear.

1. **What is your interpretation R-sq and R-sq adjusted?**

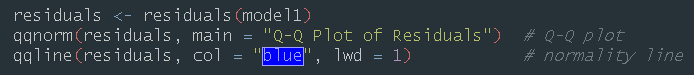
The **R-squared value = 0.5057**. This means that 50.57% of the variance in the target variable is explained by the predictor variables. The rest of the variance remains unexplained.

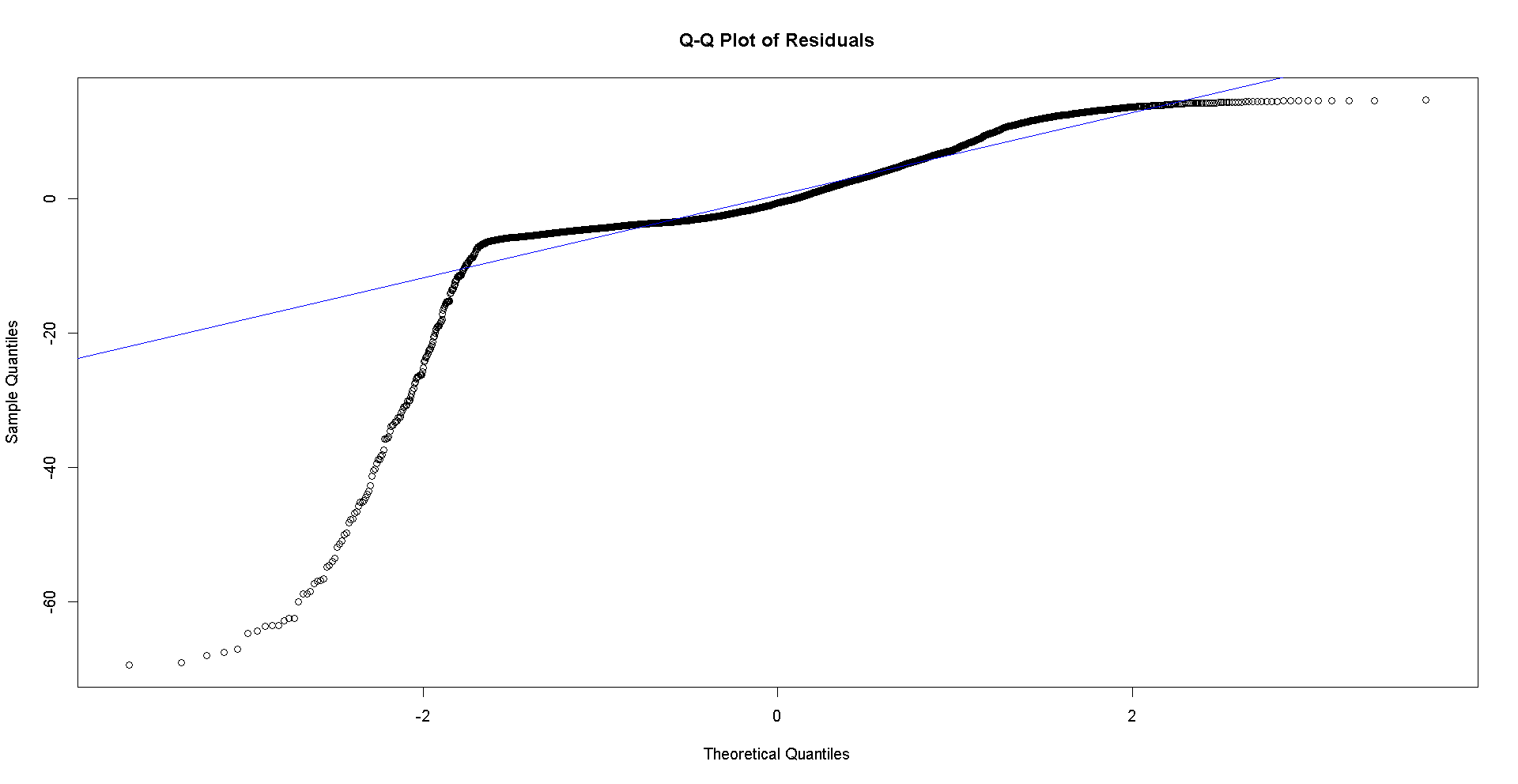
**Adjusted R-squared = 0.5052**. This shows the actual contribution of the predictor variables in the model, and is only slightly less than the R-squared value.

The R-squared and Adjusted R-squared values show a **moderate fit** of the data by the model.

Since the difference between R-squared and Adjusted R-squared is very small, the predictor variables seem to be appropriate and risk of overfitting will be low.

1. **Perform normal probability plot of residuals and comment on model adequacy.**

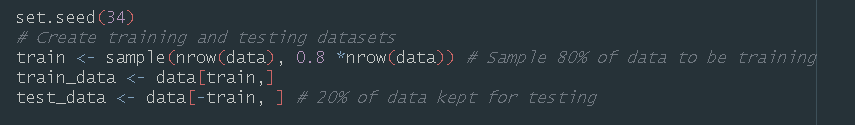




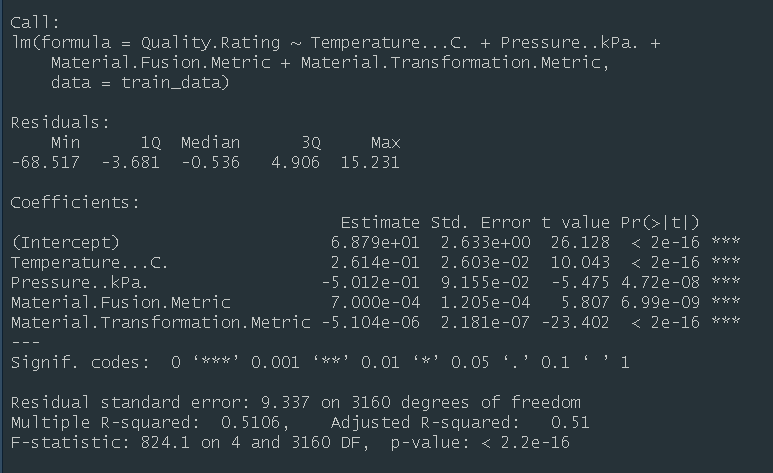
The normality plot of residuals showed a **significant deviation** from the reference line, forming a curve-like pattern. This shows that the residuals are not normally distributed and hence **indicates the inadequacy** of the model.

This can be due to a non-linear relationship of the target variable with the predictor variables.

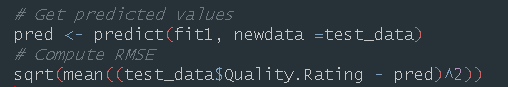
1. **Randomly sample 20% of the data and keep it as test data. Use rest of the 80% data to train the linear model. What is the RMSE on test data?**



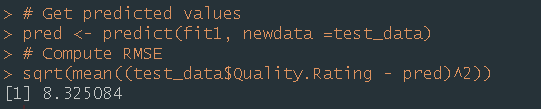
A linear model was trained with the training data.



The model was tested on the testing dataset



The **RMSE** value obtained was **8.325084**



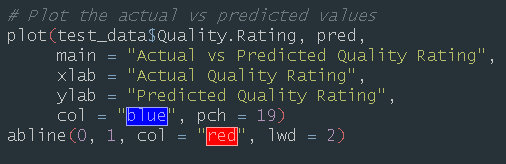
To assess the goodness of the model, the range of the Quality Rating field in the data was checked, which varied from 1 to 100.

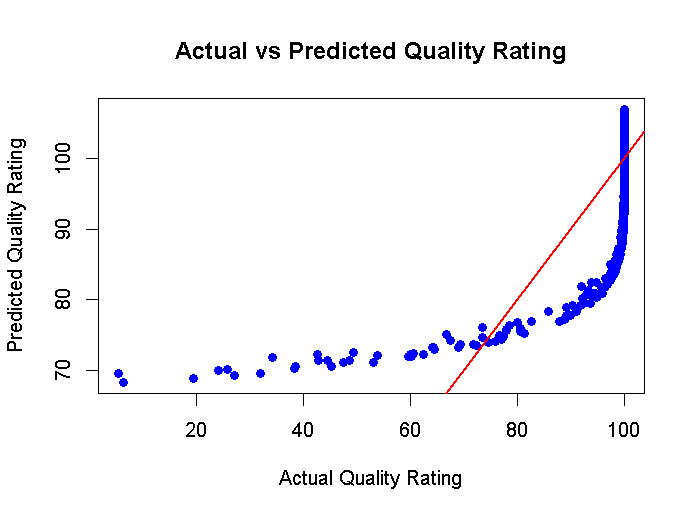




This means the **model error** is approximately **8.32%** of the target range. It shows a reasonable level of optimization, however **further investigation of other models might be required** for better accuracy.

The predicted and actual values were plotted, which showed deviation from the ideal line.





**Conclusion**

The quality rating field appears to have a non-linear relationship with the predictor variables. A linear model was trained on 4 variables. The RMSE is 8.32% of the target range which might be a large error rate in this particular context. Further more investigation might slightly improve the model, like removing variables which have high correlation amongst themselves and conducting ANOVA to test significance of the predictor variables.