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.

```
data <- read.csv("C:/Users/benab/OneDrive - iitkgp.ac.in/Desktop/Sem 6/SL Lab/Lab 2/manufacturing.csv")
head(data, 6)
```

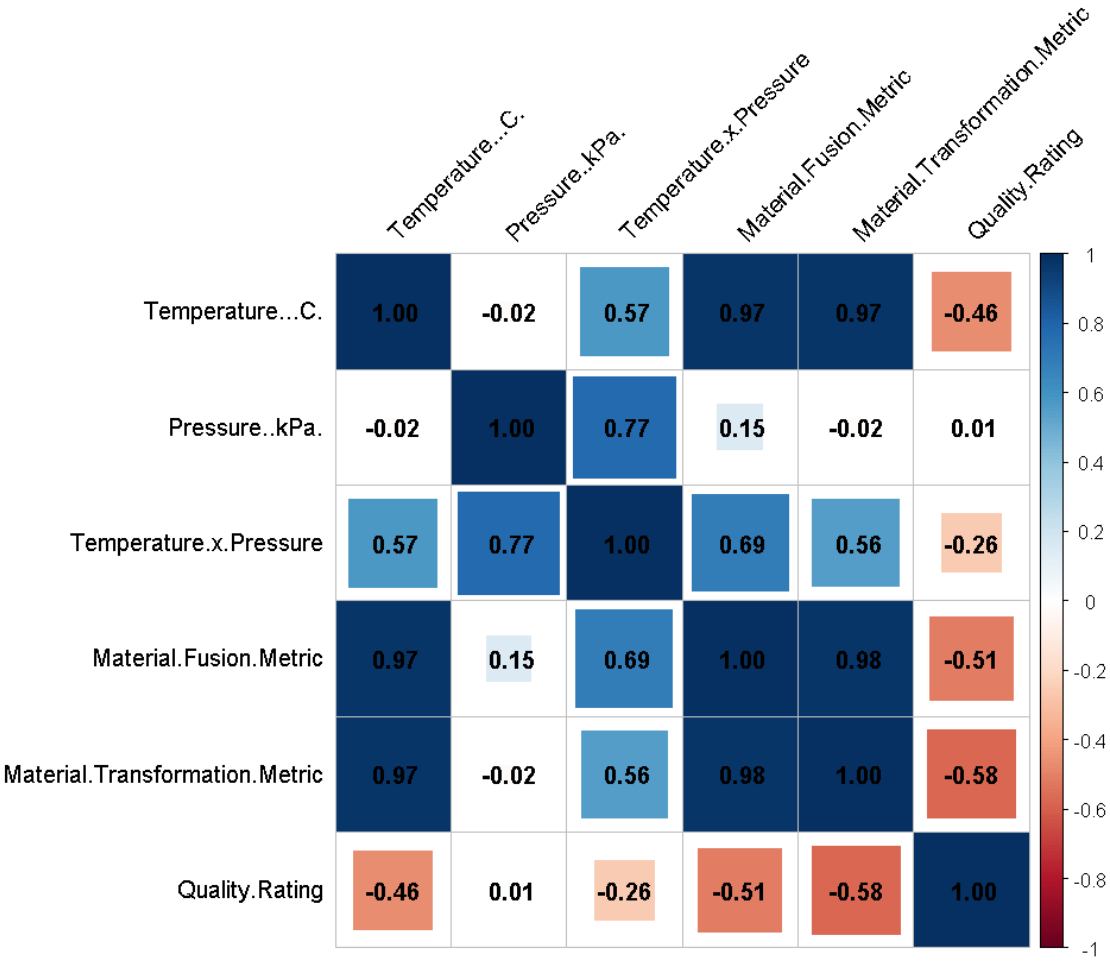
```
> data <- read.csv("C:/Users/benab/OneDrive - iitkgp.ac.in/Desktop/Sem 6/SL Lab/Lab 2/manufacturing.csv")
> head(data, 6)
  Temperature...C. Pressure..kPa. Temperature.x.Pressure Material.Fusion.Metric Material.Transformation.Metric Quality.Rating
1         209.7627       8.050855              1688.769              44522.22                        9229576       99.99997
2         243.0379      15.812068              3842.931              63020.76                       14355367       99.98570
3         220.5527       7.843130              1729.823              49125.95                       10728389       99.99976
4         208.9766      23.786089              4970.737              57128.88                        9125702       99.99997
5         184.7310      15.797812              2918.345              38068.20                        6303792      100.00000
6         229.1788       8.498306              1947.632              53136.69                       12037072       99.99879
```

2. **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.

```
corr_matrix <- cor(data,method="pearson",use="complete.obs")
library(corrplot)
corrplot(corr_matrix, method = "square", type = "full",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```

The following plot was observed:



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

a. Temperature…C. and Material.Fusion.Metric : 0.97
b. Temperature…C. and Material.Transformation.Metric : 0.97
c. Material.Transformation.Metric and Material.Fusion.Metric : 0.98

3. 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.

```
model1 <- lm(Quality.Rating ~ Temperature...C. + Pressure..kPa. +Material.Fusion.Metric +Material.Transformation.Metric , data =data)
summary(model1)
```

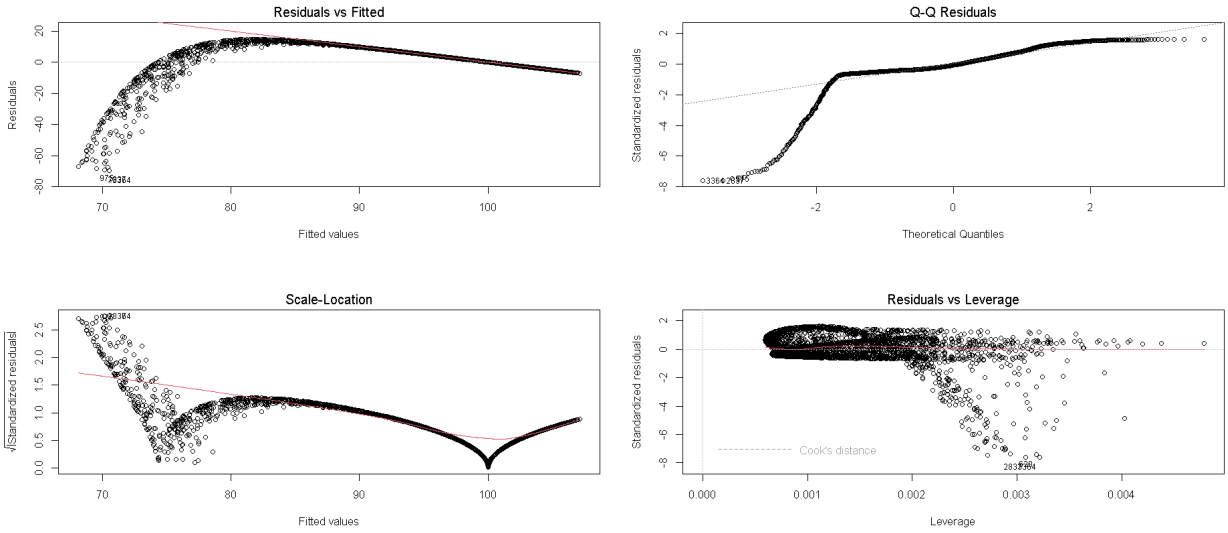
Summary of model:

```
Call:
lm(formula = Quality.Rating ~ Temperature...C. + Pressure..kPa. +
    Material.Fusion.Metric + Material.Transformation.Metric,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-69.416  -3.559  -0.563   4.746  14.728

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.978e+01  2.326e+00  29.999  < 2e-16 ***
Temperature...C.              2.522e-01  2.294e-02  10.993  < 2e-16 ***
Pressure..kPa.               -4.879e-01  8.021e-02  -6.082 1.30e-09 ***
Material.Fusion.Metric        6.905e-04  1.057e-04   6.535 7.17e-11 ***
Material.Transformation.Metric -4.980e-06  1.908e-07 -26.103  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.139 on 3952 degrees of freedom
Multiple R-squared:  0.5057,    Adjusted R-squared:  0.5052
F-statistic:  1011 on 4 and 3952 DF,  p-value: < 2.2e-16
```

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.

```
# Plot model characteristics
par(mfrow = c(2,2))
plot(model)
par(mfrow = c(1, 1))
```



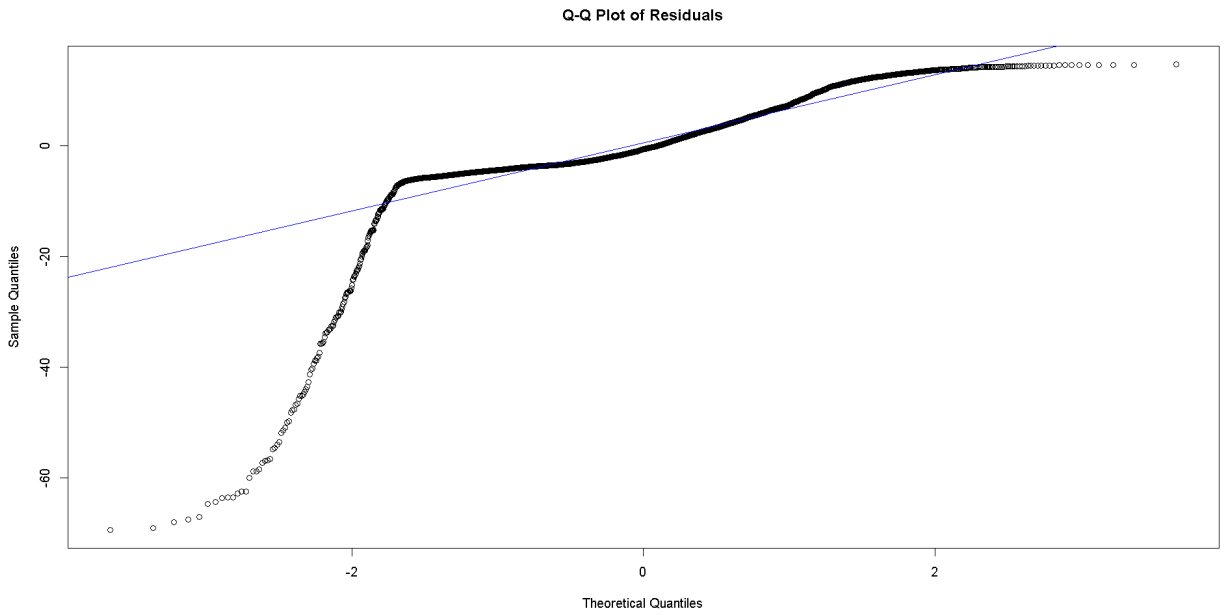
The residuals vs fitted graph shows a curve, indicating that the relationship might be non-linear.

4. 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.

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

```
residuals <- residuals(model1)
qqnorm(residuals, main = "Q-Q Plot of Residuals")  # Q-Q plot
qqline(residuals, col = "blue", lwd = 1)           # normality line
```



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.

6. **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?**

```r
set.seed(34)
# Create training and testing datasets
train <- sample(nrow(data), 0.8 *nrow(data)) # Sample 80% of data to be training
train_data <- data[train,]
test_data <- data[-train, ] # 20% of data kept for testing
```

A linear model was trained with the training data.

```
Call:
lm(formula = Quality.Rating ~ Temperature...C. + Pressure..kPa. +
    Material.Fusion.Metric + Material.Transformation.Metric,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-68.517  -3.681  -0.536   4.906  15.231

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.879e+01  2.633e+00  26.128  < 2e-16 ***
Temperature...C.              2.614e-01  2.603e-02  10.043  < 2e-16 ***
Pressure..kPa.               -5.012e-01  9.155e-02  -5.475 4.72e-08 ***
Material.Fusion.Metric        7.000e-04  1.205e-04   5.807 6.99e-09 ***
Material.Transformation.Metric -5.104e-06  2.181e-07 -23.402  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.337 on 3160 degrees of freedom
Multiple R-squared:  0.5106,    Adjusted R-squared:   0.51
F-statistic: 824.1 on 4 and 3160 DF,  p-value: < 2.2e-16
```

The model was tested on the testing dataset

```r
# Get predicted values
pred <- predict(fit1, newdata =test_data)
# Compute RMSE
sqrt(mean((test_data$Quality.Rating - pred)^2))
```

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

```
> # Get predicted values
> pred <- predict(fit1, newdata =test_data)
> # Compute RMSE
> sqrt(mean((test_data$Quality.Rating - pred)^2))
[1] 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.

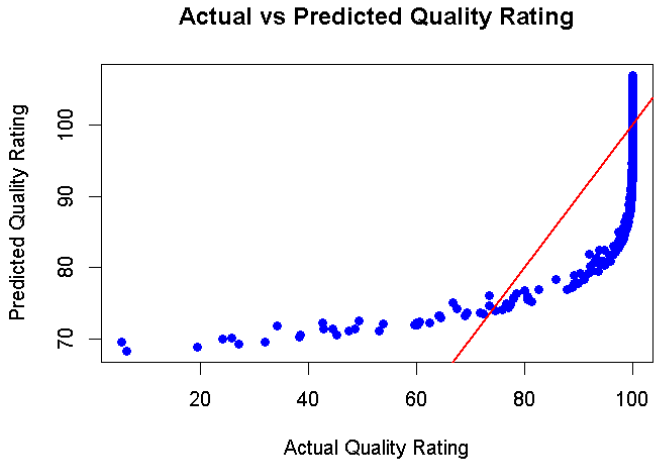
```
# Get range of the Quality rating in the data
range(data$Quality.Rating)
```

```
> # Get range of the Quality rating in the data
> range(data$Quality.Rating)
[1]   1 100
```

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.

```
# Plot the actual vs predicted values
plot(test_data$Quality.Rating, pred,
     main = "Actual vs Predicted Quality Rating",
     xlab = "Actual Quality Rating",
     ylab = "Predicted Quality Rating",
     col = "blue", pch = 19)
abline(0, 1, col = "red", lwd = 2)
```



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.