Lab 8: Normality Assumptions in R

Jeevan Koshy

September 4, 2019

**Aim**

To check whether the error term follows normality with constant variance or not using R, interpret the results and comment about heteroscedasticity.

**Procedure**

**1. Importing the data**

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.2

data <- read\_excel("C:/Users/Jeevan/Desktop/Christ University/Statistics/Linear Regression/lab-data8.xlsx")  
attach(data)

**2. Linear Regression Model**

mlmodel=lm(DeliveryTimeY~NumberOfCasesX1+DistanceX2, data=data)

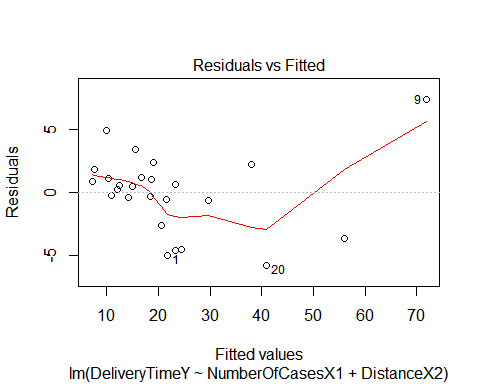
summary(mlmodel)

##   
## Call:  
## lm(formula = DeliveryTimeY ~ NumberOfCasesX1 + DistanceX2, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7860 -0.6649 0.4509 1.1713 7.4063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.323473 1.095005 2.122 0.045348 \*   
## NumberOfCasesX1 1.615374 0.170466 9.476 3.18e-09 \*\*\*  
## DistanceX2 0.014417 0.003607 3.997 0.000608 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.254 on 22 degrees of freedom  
## Multiple R-squared: 0.9597, Adjusted R-squared: 0.9561   
## F-statistic: 262.3 on 2 and 22 DF, p-value: 4.497e-16

**3. Plots**

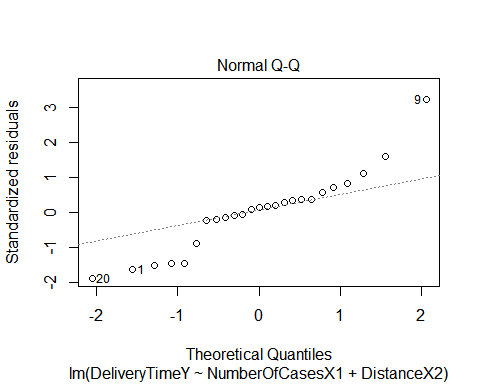
**1. Residual v/s Fitted Plot**

plot(mlmodel)



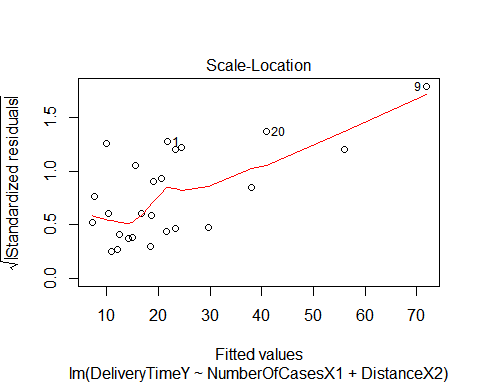
**Interpretation**

From the graph, it is clear that the data points are not equally spread across the line, i.e., the datapoints are not randomly scattered. Hence, it suggests as that the data is non-normal.

**2. Normal Q-Q Plot**

**Interpretation**

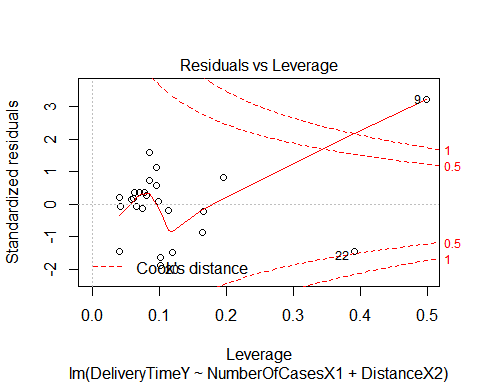
From the graph, it is clear that the data points are increasing on the extreme ends, i.e., the datapoints are deviating from normal distribution. Hence, it can be concluded that the distribution is heavy-tailed.

**3. Scale Location Plot**

**Interpretation**

From the graph, it is clear that the data points are not equally spread across the line, i.e., the datapoints are not randomly scattered. Hence, it suggests as that the data is non-normal.

**4. Residuals vs Leverage Plot**



**Interpretation**

From the graph, it can be seen that the data is near and below the Cook’s distance and hence the observations are to be deleted in order to obtain better results. The data points 22 and 9 are most influential. 22 is not affecting as is it below the line and may be include. However, the data point 9 lies on the series and is affecting it and so is best not to include it.

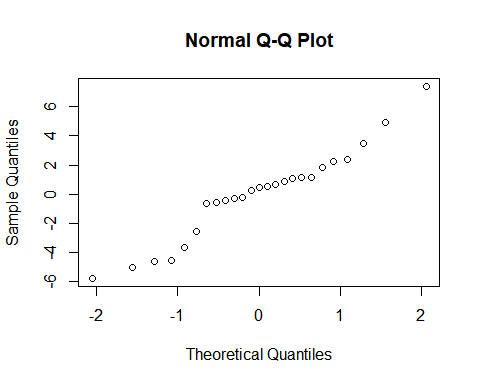
r=resid(mlmodel)   
r

## 1 2 3 4 5 6   
## -5.0246957 1.1586320 0.2285735 4.9416599 -0.4282885 -0.2787501   
## 7 8 9 10 11 12   
## 0.8598928 1.1713084 7.4062681 2.3772799 2.2415445 -0.5768987   
## 13 14 15 16 17 18   
## 0.5386578 1.0735592 0.6792781 -0.6649225 0.4508538 3.4658465   
## 19 20 21 22 23 24   
## 1.8113885 -5.7860360 -2.5956121 -3.6810862 -4.5995562 -4.5713561   
## 25   
## -0.1975409

p=fitted(mlmodel)  
p

## 1 2 3 4 5 6 7   
## 21.704696 10.341368 12.071427 9.938340 14.178288 18.388750 7.140107   
## 8 9 10 11 12 13 14   
## 16.658692 71.833732 19.122720 38.088456 21.576899 12.461342 18.676441   
## 15 16 17 18 19 20 21   
## 23.320722 29.664922 14.899146 15.534153 7.688612 40.886036 20.495612   
## 22 23 24 25   
## 56.001086 23.349556 24.401356 10.947541

qqnorm(r) # to plot the q-q plot



**Conclusion**

The error term has been checked to see if it follows normality with constant variance or not using R. The results have been interpreted with comments about heteroscedasticity.