4) Residual plots and Detection of Outliers

FOSSEE (R Team)

28 January, 2020

What are Residual plots and Outliers?

1. **Residual plots** are plots involving residuals of the generated regression model. These plots are used to identify outliers and non-linearity of the respective model.
2. **Outliers** are extreme observations — points which are typically further than three or four standard deviations from the mean.

Introduction

The purpose of this experiment is to create a multivariate linear regression model fitted over the dataset "quakes". This dataset gives the locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964. After successfully creating the model, we plot various residual plots and detect outliers.

Procedure

Step by step procedure to conduct the required experiment -

1. Normalizing data
2. Creating a linear regression model for predicting Richter Magnitude
3. Plotting residual plots
4. Detecting outliers

Note : Please make sure that the following package is already installed -

- *car*

Code and Results

Data used for analysis

```
## R has a predefined dataset with the name "quakes"
  # To know more about the dataset type "?quakes" in the console
```

Normalizing data

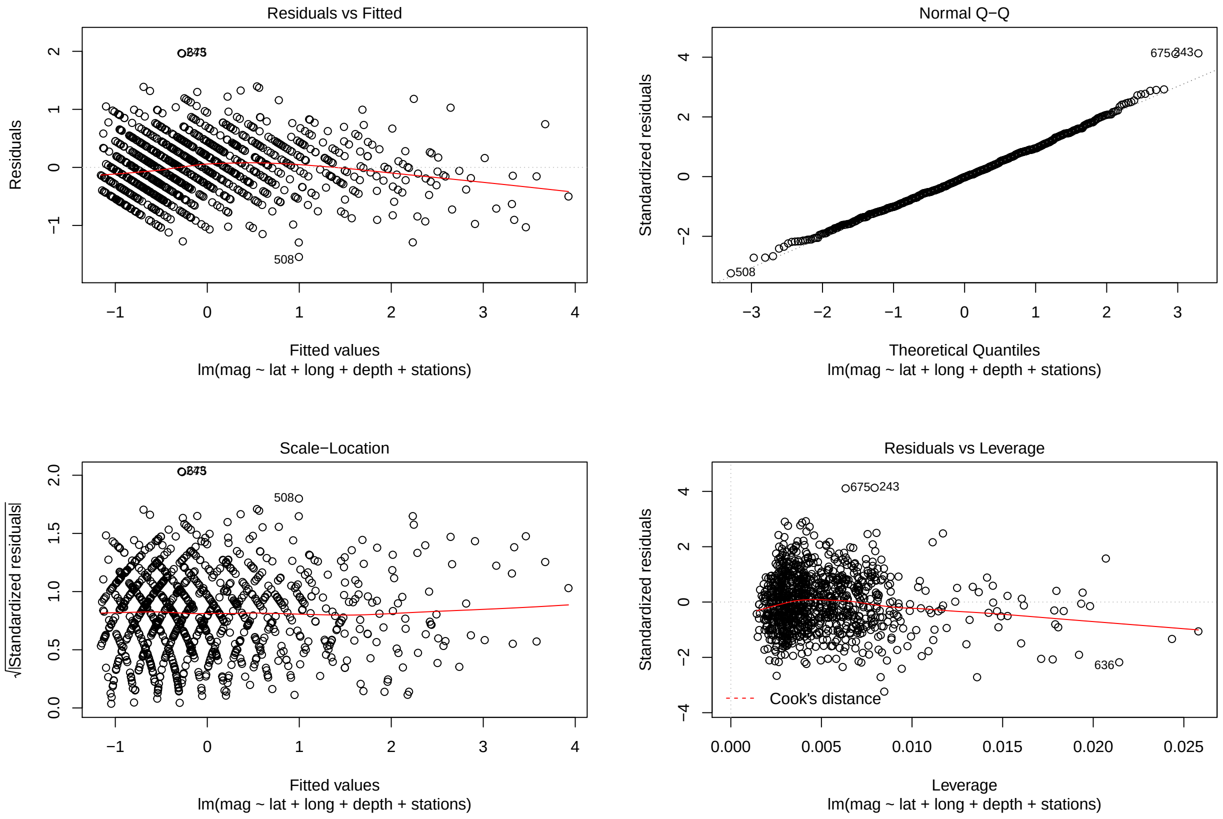
```
# 1) Normalizing data
  Data <- as.data.frame(scale(quakes))
```

Creating a linear regression model for predicting Richter Magnitude

```
# 2) Creating a Linear Regression model and predicting values
  Model <- lm(mag~lat+long+depth+stations,data = Data)
```

Plotting residual plots

```
# 3) Plotting Residuals and other model information
  plot(Model)
```



Detecting outliers

```
# 4) Outlier Test
  # Run the following command after removing "#" if "car" library is not installed
  # install.packages("car")
  library(car)
  outlierTest(Model)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 243 4.163793         3.4016e-05     0.034016
## 675 4.142554         3.7267e-05     0.037267
```

Conclusion

By observing the residual plots we can conclude the following from each plot -

1) **Residuals vs Fitted** - It indicates if there are any non-linear patterns. As there is no clear pattern in the above plot and the red line is almost linear and close to the grey dashed line. We can conclude that there are no non-linear patterns. Also, the red line starts diving towards the negative end of the residual axis. We may ignore it as the dive is not significant and was due to the points at the right-hand side of the plot being slightly negatively skewed.

2) **Normal Q-Q** - Residuals should be normally distributed. Normal Q-Q plot shows whether there is a normal distribution among the residuals or not. If the points closely follow the grey dashed line, then we may conclude that the residuals are normally distributed. In the above plot, the points closely follow the grey dashed line except for a few points on the extreme ends. Therefore, we can conclude that there is a normal distribution among the residuals.

3) **Scale Location** - It checks homoscedasticity or whether the residuals have equal variance along the regression line or not. In the above plot, the red line is almost straight, which shows that there is little change in the variance along the regression line.

4) **Residuals vs Leverage** - It identifies influential cases, which includes such extreme values that when included or excluded, may influence the regression results. In the above plot, there are no influential cases, but three extreme data points are labelled with their respective row number as present in the data set.

Outlier test outputs the row number of the outlier data points as present in the data set. For the above-generated model, the outliers are at the row numbers 243 and 675.