Regression

Alper Duru

02/18/2023

For more information on Machine Learning tests, follow here: https://alper-enes-duru.github.io/Machine_Learning_Portfolio/ (https://alper-enes-duru.github.io/Machine_Learning_Portfolio/)

Overview of Linear Regression

Linear regression is a popular and widely used statistical method for modeling the relationship between a dependent variable (or response variable) and one or more independent variables (or predictor variables). The goal of linear regression is to estimate the coefficients of a linear function that best fits the observed data and can be used to make predictions or infer the relationship between the variables.

Some of its strengths are: simplicity, interpretability, flexibility, robustness. Some of its weaknesses are: linearity assumption, independence assumption, overfitting, non-robustness to outliers.

Import the data

I used “SignificantEarthquakeDataset1900-2023.csv” data set from Kaggle to observe the data of “https://www.kaggle.com/datasets/jahaidulislam/significant-earthquake-dataset-1900-2023 (https://www.kaggle.com/datasets/jahaidulislam/significant-earthquake-dataset-1900-2023)”

```
df <- read.csv("SignificantEarthquakeDataset1900-2023.csv")
```

Train and Split the Data

We need to split the data into 2 for training and testing:

```
i <- sample(1:nrow(df), nrow(df)*0.8, replace=FALSE)
trainData <- df[i,]
testData <- df[-i,]
```

Explore the Data

This dataset is a valuable resource for researchers interested in studying major earthquakes that have taken place globally between the years 1900 and 2023. The dataset provides a vast collection of information on over 37,000 earthquakes that have occurred during this time period, and is meticulously curated and maintained by the National Earthquake Information Center (NEIC), which is a part of the United States Geological Survey (USGS). The NEIC consistently updates the dataset to ensure that the information is current and accurate, and each entry in the dataset contains important details such as the date, time, location, magnitude, and depth of each earthquake.

R functions:

Summary

```
summary(trainData)
```

```
##      Time               Place              Latitude         Longitude
##  Length:29864       Length:29864       Min.   :-77.080   Min.   :-180.00
##  Class :character   Class :character   1st Qu.:-16.590   1st Qu.: -75.41
##  Mode  :character   Mode  :character   Median :  1.194   Median :  98.62
##                                        Mean   :  5.473   Mean   :  39.12
##                                        3rd Qu.: 33.793   3rd Qu.: 143.31
##                                        Max.   : 87.199   Max.   : 180.00
##
##      Depth              Mag           MagType               nst
##  Min.   : -4.00   Min.   :5.500   Length:29864       Min.   :  0.0
##  1st Qu.: 15.00   1st Qu.:5.600   Class :character   1st Qu.:136.0
##  Median : 28.40   Median :5.800   Mode  :character   Median :242.0
##  Mean   : 58.89   Mean   :5.951                      Mean   :265.9
##  3rd Qu.: 41.10   3rd Qu.:6.150                      3rd Qu.:372.0
##  Max.   :700.00   Max.   :9.500                      Max.   :934.0
##  NA's   :103                                         NA's   :23911
##       gap              dmin             rms              net
##  Min.   :  8.00   Min.   : 0.005   Min.   : 0.040   Length:29864
##  1st Qu.: 24.10   1st Qu.: 1.173   1st Qu.: 0.880   Class :character
##  Median : 36.00   Median : 2.532   Median : 1.000   Mode  :character
##  Mean   : 44.91   Mean   : 4.354   Mean   : 1.001
##  3rd Qu.: 55.00   3rd Qu.: 5.168   3rd Qu.: 1.110
##  Max.   :344.00   Max.   :39.730   Max.   :42.410
##  NA's   :21801    NA's   :26331    NA's   :13726
##       ID              Updated             X               Type
##  Length:29864       Length:29864       Mode:logical   Length:29864
##  Class :character   Class :character   NA's:29864     Class :character
##  Mode  :character   Mode  :character                  Mode  :character
##
##
##
##
##  horizontalError    depthError        magError         magNst
##  Min.   : 0.085   Min.   :  0.00   Min.   :0.000   Min.   :  0.00
##  1st Qu.: 5.700   1st Qu.:  3.60   1st Qu.:0.200   1st Qu.: 18.00
##  Median : 7.100   Median :  6.10   Median :0.200   Median : 32.00
##  Mean   : 7.355   Mean   : 10.66   Mean   :0.262   Mean   : 47.92
##  3rd Qu.: 8.500   3rd Qu.: 16.20   3rd Qu.:0.330   3rd Qu.: 56.00
##  Max.   :99.000   Max.   :569.20   Max.   :1.840   Max.   :941.00
##  NA's   :26671    NA's   :13101    NA's   :16533   NA's   :25545
##     status          locationSource      magSource
##  Length:29864       Length:29864       Length:29864
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##
```

str

```
str(trainData)
```

```
## 'data.frame':    29864 obs. of  23 variables:
##  $ Time           : chr  "2021-08-12T18:41:58.470Z" "1926-03-13T19:35:54.180Z" "1946-
05-08T05:20:27.130Z" "2004-12-26T03:08:44.210Z" ...
##  $ Place          : chr  "South Sandwich Islands region" "256 km SSE of Hirara, Japa
n" "122 km W of Pariaman, Indonesia" "228 km N of Bamboo Flat, India" ...
##  $ Latitude       : num  -60.292 22.886 -0.729 13.745 35.056 ...
##  $ Longitude      : num  -26.5 126.7 99 93 50.2 ...
##  $ Depth          : num  35 15 35 30 15 10.5 11.4 120 33 97.2 ...
##  $ Mag            : num  5.8 5.89 6.99 5.9 5.52 5.9 6.4 6.3 5.7 5.77 ...
##  $ MagType        : chr  "mb" "mw" "mw" "mb" ...
##  $ nst            : int  NA NA NA 289 NA 361 NA NA 183 NA ...
##  $ gap            : num  57 NA NA 64 NA 22.6 NA 29 NA NA ...
##  $ dmin           : num  8.07 NA NA NA NA ...
##  $ rms            : num  0.63 NA NA 0.87 NA 0.84 1.2 1.22 0.71 NA ...
##  $ net            : chr  "us" "iscgem" "iscgem" "us" ...
##  $ ID             : chr  "us6000f903" "iscgem909763" "iscgem898378" "usp000dbfc" ...
##  $ Updated        : chr  "2021-10-23T15:47:02.040Z" "2022-04-25T23:14:29.944Z" "2022-
04-26T18:55:29.459Z" "2022-07-14T19:04:28.882Z" ...
##  $ X              : logi  NA NA NA NA NA NA ...
##  $ Type           : chr  "earthquake" "earthquake" "earthquake" "earthquake" ...
##  $ horizontalError: num  12 NA NA NA NA NA NA 7.2 NA NA ...
##  $ depthError     : num  1.4 25 9.9 NA 3.9 NA NA 1.9 NA 5 ...
##  $ magError       : num  0.12 0.2 0.39 NA 0.4 NA NA 0.045 NA 0.48 ...
##  $ magNst         : int  25 NA NA 97 NA NA NA 47 NA NA ...
##  $ status         : chr  "reviewed" "reviewed" "reviewed" "reviewed" ...
##  $ locationSource : chr  "us" "iscgem" "iscgem" "us" ...
##  $ magSource      : chr  "us" "iscgem" "iscgem" "us" ...
```

Head Function for Dataset

```
head(trainData$Mag, n=10)
```

```
##  [1] 5.80 5.89 6.99 5.90 5.52 5.90 6.40 6.30 5.70 5.77
```

Dimensions of the Dataset

```
dim(trainData)
```

```
## [1] 29864    23
```

Are there any NA's in the dataset?

```
sapply(trainData, function(x) sum(is.na(x)))
```

```
##             Time            Place         Latitude        Longitude            Depth 
##                0                0                0                0              103 
##              Mag          MagType              nst              gap             dmin 
##                0                0            23911            21801            26331 
##              rms              net               ID          Updated                X 
##            13726                0                0                0            29864 
##             Type horizontalError       depthError         magError           magNst 
##                0            26671            13101            16533            25545 
##           status   locationSource       magSource 
##                0                0                0
```

Informative Graphs

Plot the Data

```
plot(trainData$Mag, trainData$Depth, pch=15, col="red", cex=0.5, main="Magnitude vs Dept
h", xlab="Magnitude", ylab="Depth")
```

Magnitude greater than 7

```
magnitudeGreaterThan7 <- subset(trainData, Mag > 7)
hist(magnitudeGreaterThan7$Mag, labels=TRUE)
```

Linear Regression Model

```
lrm <- lm(Mag~Depth, data = trainData)
summary (lrm)
```

```
##
## Call:
## lm(formula = Mag ~ Depth, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4522 -0.3513 -0.1497  0.1982  3.5484
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  5.952e+00  3.002e-03 1982.534   <2e-16 ***
## Depth       -2.374e-05  2.403e-05   -0.988    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4568 on 29759 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  3.279e-05,  Adjusted R-squared:  -8.098e-07
## F-statistic: 0.9759 on 1 and 29759 DF,  p-value: 0.3232
```

Plot the Residuals

```
plot (lrm)
```

Multiple Linear Regression Model

```
mlrm <- lm(Mag~Longitude+Latitude, data = trainData)
summary (mlrm)
```

```
##
## Call:
## lm(formula = Mag ~ Longitude + Latitude, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4966 -0.3467 -0.1319  0.1965  3.5831
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 5.944e+00  2.793e-03 2127.816  < 2e-16 ***
## Longitude   1.134e-04  2.178e-05    5.205 1.96e-07 ***
## Latitude    4.823e-04  8.690e-05    5.550 2.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4566 on 29861 degrees of freedom
## Multiple R-squared:  0.002304,   Adjusted R-squared:  0.002237
## F-statistic: 34.48 on 2 and 29861 DF,  p-value: 1.106e-15
```

Plot the Residuals

```
plot(mlrm)
```

Third Linear Regression Model using different combinations of predictors

```
mlrm2 <- lm(Mag~Depth+locationSource, data = trainData)
summary (mlrm2)
```

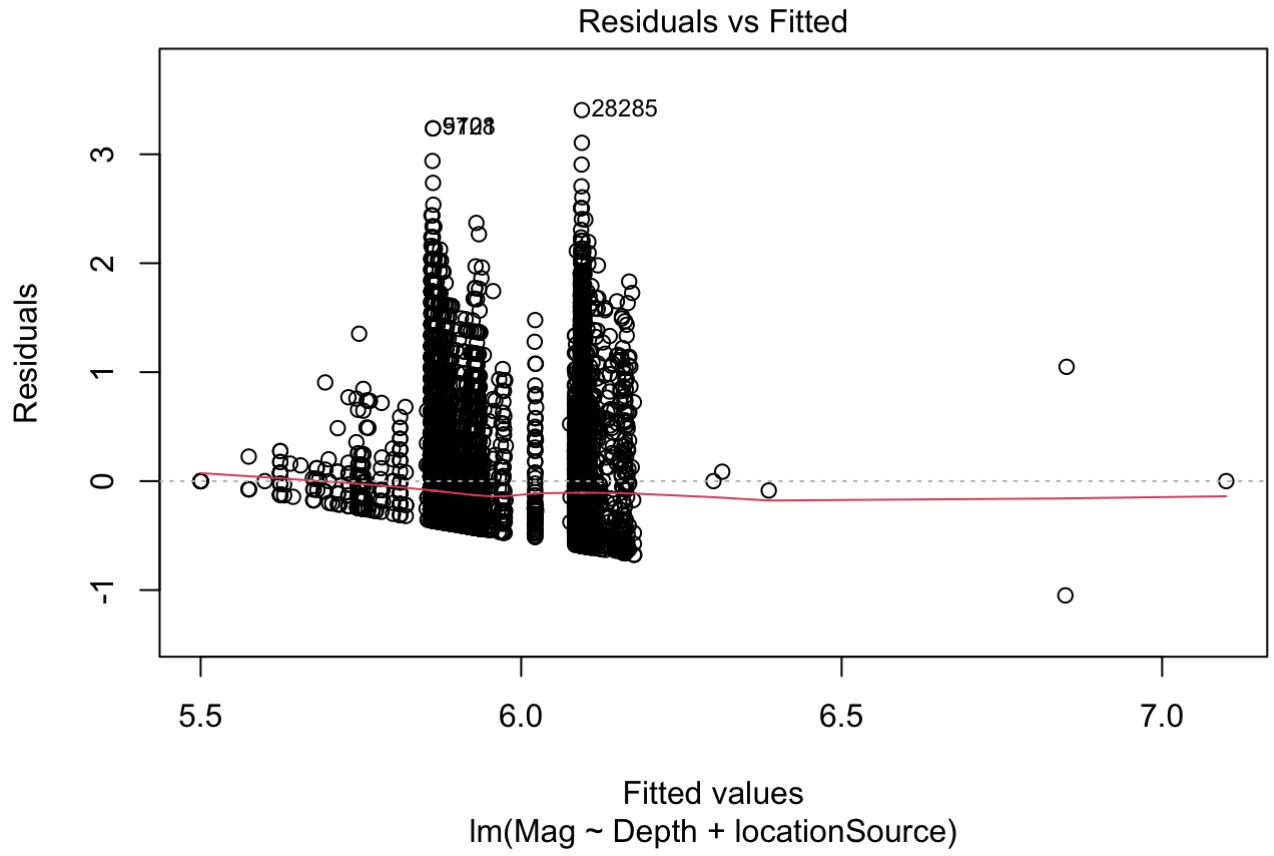
```
##
## Call:
## lm(formula = Mag ~ Depth + locationSource, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0491 -0.3135 -0.1335  0.1759  3.4053
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.811e+00  7.384e-02  78.701  < 2e-16 ***
## Depth                    1.191e-04  2.359e-05   5.048 4.48e-07 ***
## locationSourceaeic      -3.165e-02  1.526e-01  -0.207 0.835742
## locationSourceag         1.287e+00  4.491e-01   2.866 0.004164 **
## locationSourceags        1.163e-01  1.830e-01   0.636 0.525011
## locationSourceak         1.156e-01  8.928e-02   1.294 0.195548
## locationSourceath        4.067e-02  1.526e-01   0.266 0.789860
## locationSourcebeo       -3.112e-01  4.491e-01  -0.693 0.488364
## locationSourcebrk       -3.121e-01  4.491e-01  -0.695 0.487183
## locationSourcecar       -3.118e-01  4.491e-01  -0.694 0.487499
## locationSourcecasc       1.616e-01  2.335e-01   0.692 0.488898
## locationSourceci         2.100e-01  9.023e-02   2.327 0.019948 *
## locationSourcecsem      -2.123e-01  4.491e-01  -0.473 0.636440
## locationSourcedoe       -1.870e-01  1.105e-01  -1.691 0.090774 .
## locationSourcee         -3.111e-01  4.491e-01  -0.693 0.488513
## locationSourceg          8.806e-02  4.491e-01   0.196 0.844570
## locationSourcegcmt      -1.137e-01  3.219e-01  -0.353 0.723776
## locationSourceguc       -6.877e-02  1.030e-01  -0.667 0.504552
## locationSourceh         -1.266e-02  4.491e-01  -0.028 0.977514
## locationSourcehv         1.440e-01  1.304e-01   1.104 0.269397
## locationSourceiscgem     2.806e-01  7.397e-02   3.793 0.000149 ***
## locationSourceiscgemsup  2.715e-01  7.661e-02   3.544 0.000395 ***
## locationSourceisk       -8.120e-02  1.434e-01  -0.566 0.571084
## locationSourcejma        1.536e-01  2.662e-01   0.577 0.563950
## locationSourcelim       -3.236e-01  4.491e-01  -0.721 0.471214
## locationSourcemdd        5.026e-01  3.219e-01   1.561 0.118505
## locationSourcenc         1.600e-01  9.906e-02   1.615 0.106377
## locationSourcenn         7.867e-03  2.114e-01   0.037 0.970320
## locationSourceofficial   1.038e+00  3.219e-01   3.226 0.001255 **
## locationSourcepgc        1.570e-01  1.434e-01   1.095 0.273475
## locationSourcepr        -1.210e-02  2.662e-01  -0.045 0.963738
## locationSourcept        -3.080e-02  4.492e-01  -0.069 0.945335
## locationSourceren       -1.120e-01  4.491e-01  -0.249 0.803051
## locationSourcerom       -2.374e-01  2.335e-01  -1.017 0.309368
## locationSourcerspr      -1.687e-01  3.219e-01  -0.524 0.600172
## locationSourcese        -1.183e-02  4.491e-01  -0.026 0.978995
## locationSourcesja        4.877e-01  4.491e-01   1.086 0.277554
## locationSourcespe        3.608e-01  1.732e-01   2.083 0.037215 *
## locationSourcetap       -3.129e-01  4.491e-01  -0.697 0.485987
## locationSourceteh       -9.849e-02  1.830e-01  -0.538 0.590460
## locationSourcethe       -1.810e-01  2.662e-01  -0.680 0.496627
## locationSourcethr       -1.305e-02  3.219e-01  -0.041 0.967652
```

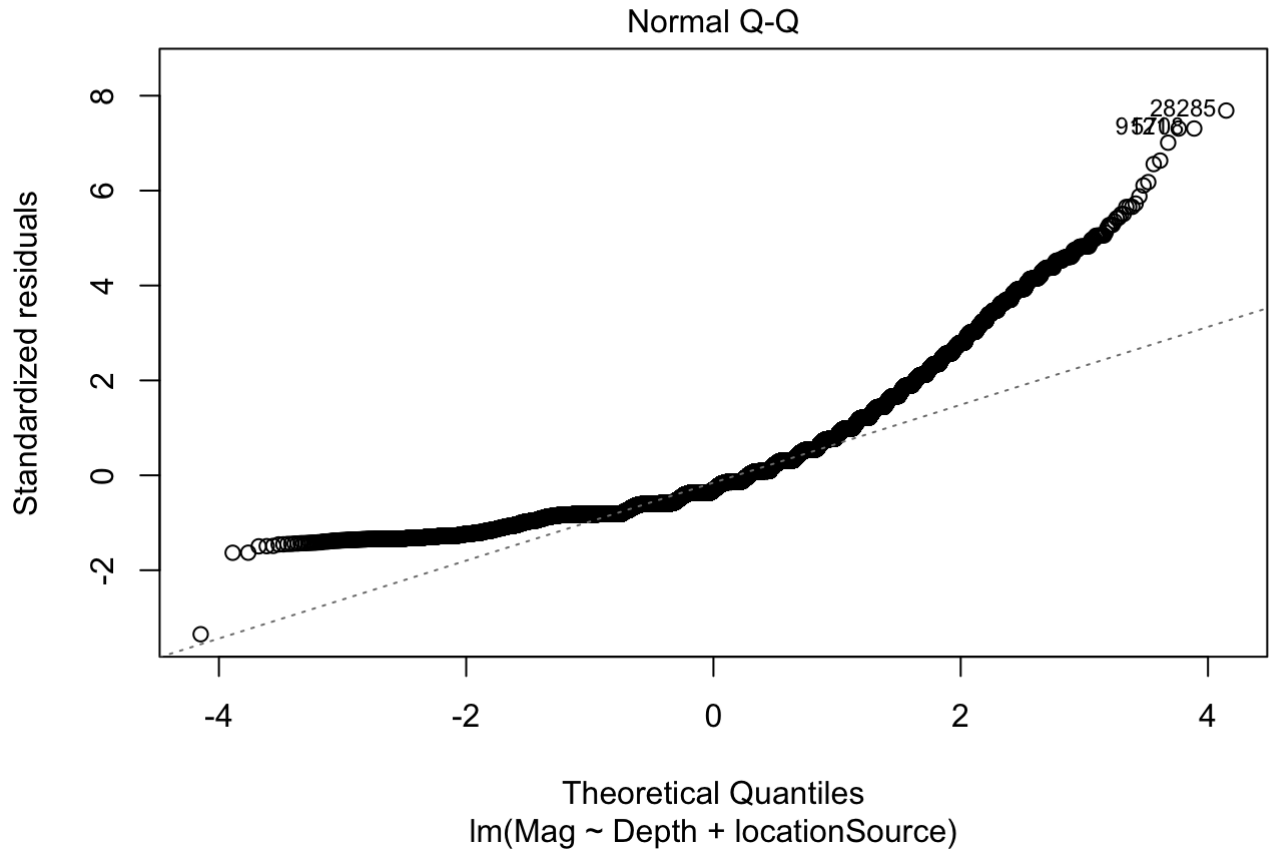
```
## locationSourcetul       -6.175e-02  3.219e-01  -0.192 0.847851
## locationSourceucr        8.022e-02  4.491e-01   0.179 0.858249
## locationSourceunm       -1.366e-01  1.279e-01  -1.068 0.285401
## locationSourceus         4.775e-02  7.393e-02   0.646 0.518334
## locationSourceus_wel     1.378e-01  2.335e-01   0.590 0.555104
## locationSourceushis     -6.133e-02  1.001e-01  -0.613 0.540014
## locationSourceuu         2.642e-01  2.662e-01   0.992 0.321029
## locationSourceuw         1.530e-01  1.954e-01   0.783 0.433604
## locationSourcewel        4.715e-02  1.331e-01   0.354 0.723203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.443 on 29710 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.06092,    Adjusted R-squared:  0.05934
## F-statistic: 38.55 on 50 and 29710 DF,  p-value: < 2.2e-16
```

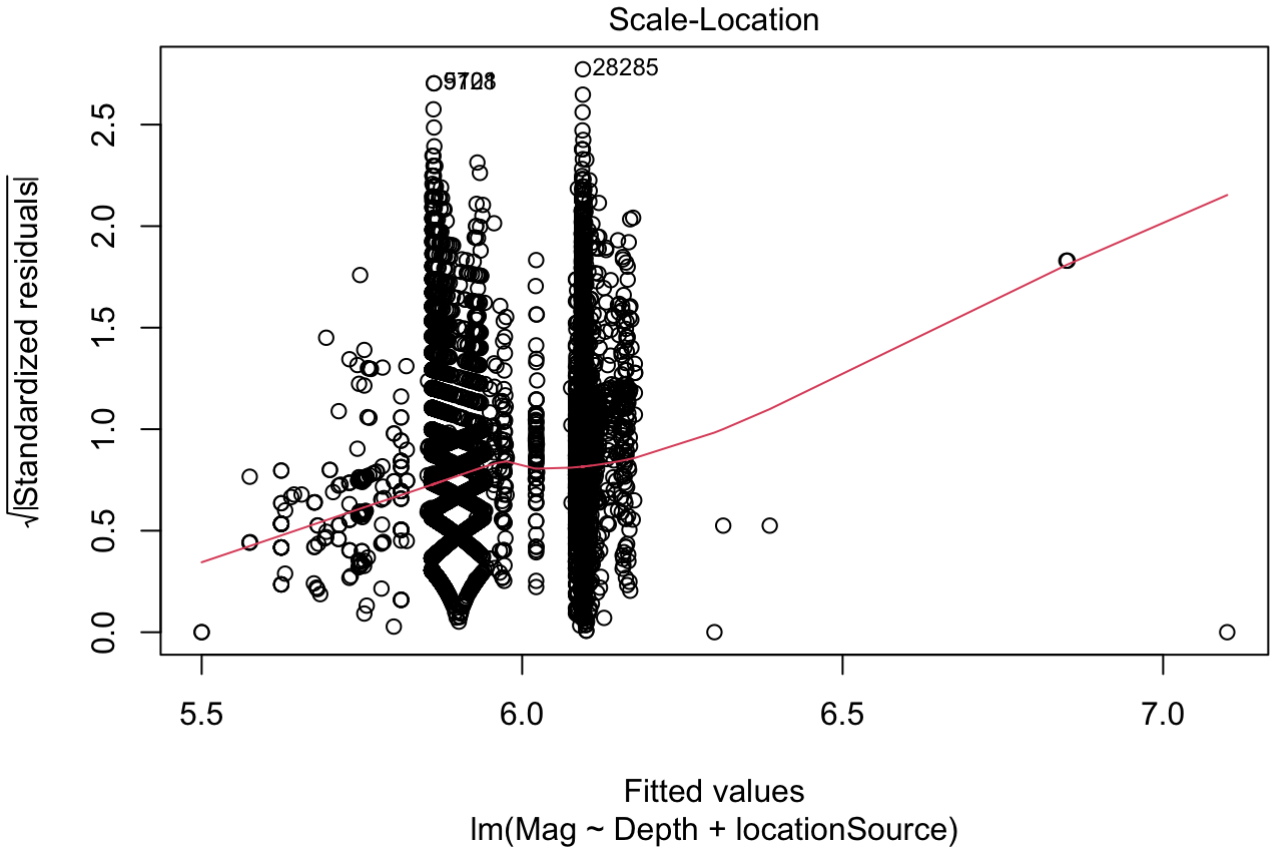
Plot the Data

```
plot(mlrm2)
```

```
## Warning: not plotting observations with leverage one:
##   714, 4518, 5719, 6832, 7337, 7872, 8256, 10659, 10761, 17796, 27908
```

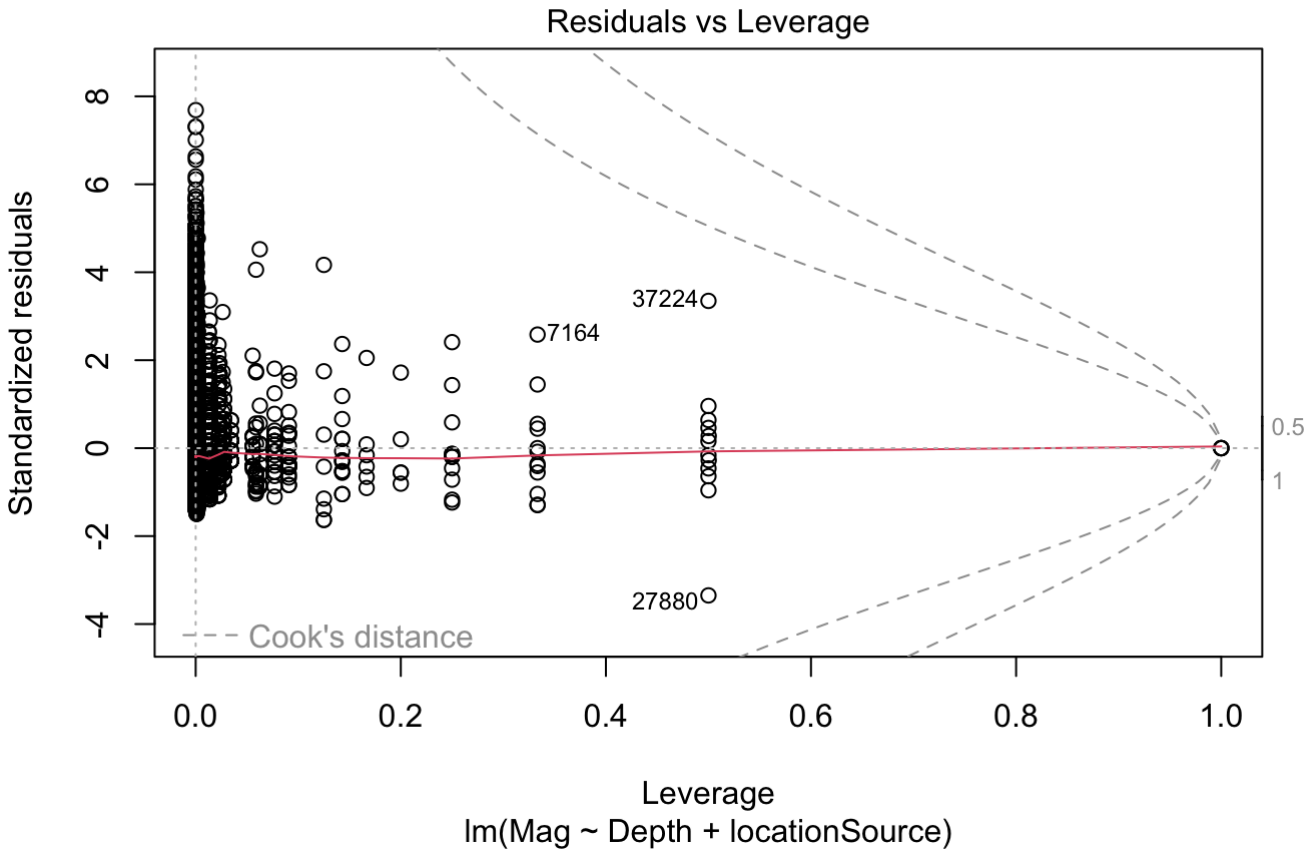

Residuals vs Fitted
Residuals
Fitted values
lm(Mag ~ Depth + locationSource)
28285
5.5
6.0
6.5
7.0
-1
0
1
2
3


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Mag ~ Depth + locationSource)
28285
-4
-2
0
2
4
-2
0
2
4
6
8




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Compare the Results

First Model lrm

```
prediction1 <- predict(lrm, newdata = testData)
correlation1 <- cor(prediction1, testData$Mag)
mse1 <- mean((prediction1-testData$Mag)^1)
rmse1 <- sqrt(mse1)
print(paste('Correlation: ', correlation1))
```

```
## [1] "Correlation:  NA"
```

```
print(paste('MSE: ', mse1))
```

```
## [1] "MSE:  NA"
```

```
print(paste('RMSE: ', rmse1))
```

```
## [1] "RMSE:  NA"
```

Second Model mlrm

```
prediction2 <- predict(mlrm, newdata = testData)
correlation2 <- cor(prediction2, testData$Mag)
mse2 <- mean((prediction2-testData$Mag)^2)
rmse2 <- sqrt(mse2)
print(paste('Correlation: ', correlation2))
```

```
## [1] "Correlation:  0.0565482896126271"
```

```
print(paste('MSE: ', mse2))
```

```
## [1] "MSE:  0.199514807949824"
```

```
print(paste('RMSE: ', rmse2))
```

```
## [1] "RMSE:  0.446670804899788"
```

Third Model mlrm2

```
prediction3 <- predict(mlrm, newdata = testData)
correlation3 <- cor(prediction3, testData$Mag)
mse3 <- mean((prediction3-testData$Mag)^2)
rmse3 <- sqrt(mse3)
print(paste('Correlation: ', correlation3))
```

```
## [1] "Correlation:  0.0565482896126271"
```

```
print(paste('MSE: ', mse3))
```

```
## [1] "MSE:  0.199514807949824"
```

```
print(paste('RMSE: ', rmse3))
```

```
## [1] "RMSE:  0.446670804899788"
```