EARTHQUAKE MAGNITUDE PREDICTION

TABLE OF CONTENTS

LIST OF FIGURES

Figures Page no

INTRODUCTION TO THE PROJECT

Using the Earthquake Magnitude Prediction we can predict the magnitude of any earthquake. Earthquake is a common natural calamity which occurs due to the collision of the tectonic plates under Earth's surface. Occurrence of any earthquake can certainly be predicted, but it is even more important to predict the impact that earthquake is going to make. This can be done by taking into account various factors such as the location which is considered using the latitude-longitude values and the depth.

The dataset contains 5 attributes namely lat, long, depth, mag, stations. Out of which we have 4 Independent variables (Non Target Attributes) and 1 dependent variables (Target Attributes).

This project also contains Prediction analysis using the multiple regression analysis and the use of various statistical functions to analyze the data.

Description of each attribute is given below:

Independent Variables:

1. lat: numeric Latitude of event
2. long: numeric Longitude
3. depth: numeric Depth (km)
4. stations: numeric Number of stations reporting

Dependent Variables:

1. mag: numeric Richter Magnitude (MB > 4.0)

TOOLS AND TECHNOLOGIES USED

Hardware Requirements:

System type : 64-bit Operating System

RAM : 4.00 GB

Processor : Intel® Core™ i3-4030U CPU @ 1.90GHz

Hard Disk : 1 TB

Software Requirements:

Operating System : Windows 10

Coding Language : R (R Studio)

Documentation : MS Office 2010

RESULTS AND DISCUSSIONS

Data analytics

Data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them which help organizations make informed business decisions. Business intelligence queries answer basic questions about business operations and performance.

Big data analytics

Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information such as hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.

Supervised vs Unsupervised learning models

Supervised Learning models are the models where there is a clear distinction between explanatory and dependent variables. The models are trained to explain dependent variables using explanatory variables. In other words, the model output attributes are known beforehand.

1. Prediction (e.g., linear regression)
2. Classification (e.g., decision trees, k-nearest neighbors)
3. Time-series forecasting (e.g., regression-based)

In unsupervised learning, the model outputs are unknown or there are no target attributes: there is no distinction between explanatory and dependent variables. The models are created to find out the intrinsic structure of data.

1. Association rules
2. Cluster analysis

IMPLEMENTATION USING R

```
setwd("c:/rlab")
d1<-read.csv("quakes.csv")

names(d1)
[1] "S.no"     "lat"      "long"     "depth"    "mag"
[6] "stations"

 dim(d1)
[1] 1000    6

head(d1)
  S.no    lat   long depth mag stations
1    1 -20.42 181.62   562 4.8       41
2    2 -20.62 181.03   650 4.2       15
3    3 -26.00 184.10    42 5.4       43
4    4 -17.97 181.66   626 4.1       19
5    5 -20.42 181.96   649 4.0       11
6    6 -19.68 184.31   195 4.0       12
```

Statistical Functions

Mean

The most common expression for the mean of a statistical distribution with a discrete random variable is the mathematical average of all the terms. To calculate it, add up the values of all the terms and then divide by the number of terms. The mean of a statistical distribution with a continuous random variable, also called the expected value, is obtained by integrating the product of the variable with its probability as defined by the distribution.

Median

The median of a distribution with a discrete random variable depends on whether the number of terms in the distribution is even or odd. If the number of terms is odd, then the median is the value of the term in the middle. This is the value such that the number of terms having values greater than or equal to it is the same as the number of terms having values less than or equal to it. If the number of terms is even, then the median is the average of the two terms in the middle, such that the number of terms having values greater than or equal to it is the same as the number of terms having values less than or equal to it.

Mode

The mode of a distribution with a discrete random variable is the value of the term that occurs the most often. It is not uncommon for a distribution with a discrete random variable to have more than one mode, especially if there are not many terms. This happens when two or more terms occur with equal frequency, and more often than any of the others.

```
mean(d1$mag)
[1] 4.6204
median(d1$mag)
[1] 4.6

summary(d1)
      S.no             lat              long
 Min.   :   1.0   Min.   :-38.59   Min.   :165.7
 1st Qu.: 250.8   1st Qu.:-23.47   1st Qu.:179.6
 Median : 500.5   Median :-20.30   Median :181.4
 Mean   : 500.5   Mean   :-20.64   Mean   :179.5
 3rd Qu.: 750.2   3rd Qu.:-17.64   3rd Qu.:183.2
 Max.   :1000.0   Max.   :-10.72   Max.   :188.1
    depth            mag          stations
 Min.   : 40.0   Min.   :4.00   Min.   : 10.00
 1st Qu.: 99.0   1st Qu.:4.30   1st Qu.: 18.00
 Median :247.0   Median :4.60   Median : 27.00
 Mean   :311.4   Mean   :4.62   Mean   : 33.42
 3rd Qu.:543.0   3rd Qu.:4.90   3rd Qu.: 42.00
 Max.   :680.0   Max.   :6.40   Max.   :132.00
```

Str():

Str is a compact way to display the structure of an R object.
This allows you to use str as a diagnostic function and an alternative to summary. Str will output the information on one line for each basic structure. Str is best for displaying contents of lists. The goals is to get an output for any R object.

```
str(d1)
'data.frame': 1000 obs. of  6 variables:
 $ S.no    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ lat     : num  -20.4 -20.6 -26 -18 -20.4 ...
 $ long    : num  182 181 184 182 182 ...
 $ depth   : int  562 650 42 626 649 195 82 194 211 622 ...
 $ mag     : num  4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...
 $ stations: int  41 15 43 19 11 12 43 15 35 19 ...
```

Creating subsets for the dataset

```
s1<-subset(d1,mag<=4.0)
write.csv(s1,"maglessthan4.csv")
s1
```

	S.no	lat	long	depth	mag	stations
5	5	-20.42	181.96	649	4	11
6	6	-19.68	184.31	195	4	12
26	26	-17.94	181.49	537	4	15
34	34	-23.55	180.80	349	4	10
52	52	-19.26	184.42	223	4	15
58	58	-22.06	180.60	584	4	11
71	71	-15.31	185.80	152	4	11
85	85	-17.70	181.70	450	4	11
96	96	-19.73	182.40	375	4	18
113	113	-19.06	182.45	477	4	16
142	142	-20.65	181.40	582	4	14
150	150	-17.90	181.50	573	4	19
202	202	-17.70	182.20	445	4	12
236	236	-23.54	179.93	574	4	12
284	284	-17.70	185.00	383	4	10
298	298	-17.94	181.51	601	4	16
299	299	-30.64	181.20	175	4	16

362	362	-16.90	185.72	135	4	22
389	389	-10.72	165.99	195	4	14
433	433	-18.55	182.23	563	4	17
483	483	-22.70	183.30	180	4	13
533	533	-21.00	183.20	296	4	16
598	598	-17.02	182.93	406	4	17
637	637	-19.51	183.97	280	4	16
698	698	-15.43	185.19	249	4	11
722	722	-17.91	181.48	555	4	17
598	598	-17.02	182.93	406	4	17
637	637	-19.51	183.97	280	4	16
698	698	-15.43	185.19	249	4	11
722	722	-17.91	181.48	555	4	17
727	727	-17.10	182.80	390	4	14
733	733	-30.30	180.80	275	4	14
750	750	-25.60	180.30	440	4	12
727	727	-17.10	182.80	390	4	14
733	733	-30.30	180.80	275	4	14
750	750	-25.60	180.30	440	4	12
770	770	-20.70	186.30	80	4	10
772	772	-16.40	182.73	391	4	16
775	775	-21.60	180.50	595	4	22
780	780	-17.90	181.50	589	4	12

794	794	-28.00	182.00	199	4	16
816	816	-22.12	180.49	532	4	14
826	826	-21.62	182.40	350	4	12
834	834	-19.70	182.44	397	4	12
856	856	-18.50	185.40	243	4	11
861	861	-21.03	180.78	638	4	14
875	875	-17.80	181.20	530	4	15
900	900	-17.82	181.27	538	4	33
919	919	-25.06	182.80	133	4	14
937	937	-18.14	181.71	574	4	20
955	955	-23.49	180.06	530	4	23
989	989	-17.86	181.30	614	4	12
994	994	-17.95	181.37	642	4	17

```
s2<-subset(d1,stations<=10)
write.csv(s2,"stationslessthan10.csv")
s2
```

	S.no	lat	long	depth	mag	stations
14	14	-21.00	181.66	600	4.4	10
34	34	-23.55	180.80	349	4.0	10
35	35	-16.30	186.00	48	4.5	10
146	146	-20.10	184.40	186	4.2	10
175	175	-15.03	182.29	399	4.1	10
263	263	-19.06	169.01	158	4.4	10
284	284	-17.70	185.00	383	4.0	10
327	327	-21.04	181.20	483	4.2	10
350	350	-27.21	182.43	55	4.6	10
431	431	-18.40	183.40	343	4.1	10
438	438	-20.30	182.30	476	4.5	10
482	482	-14.85	184.87	294	4.1	10
598	598	-17.02	182.93	406	4	10
637	637	-19.51	183.97	280	4	10
698	698	-15.43	185.19	249	4	10
722	722	-17.91	181.48	555	4	10
727	727	-17.10	182.80	390	4	10
733	733	-30.30	180.80	275	4	10
750	750	-25.60	180.30	440	4	10

690	690	-17.60	181.50	548	4.1	10
693	693	-20.61	182.44	518	4.2	10
704	704	-25.00	180.00	488	4.5	10
763	763	-17.78	185.33	223	4.1	10
770	770	-20.70	186.30	80	4.0	10
776	776	-21.77	181.00	618	4.1	10
778	778	-21.05	180.90	616	4.3	10
995	995	-17.70	188.10	45	4.2	10

Regression

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. Regression analysis provides detailed insight that can be applied to further improve products and services.

Linear Regression

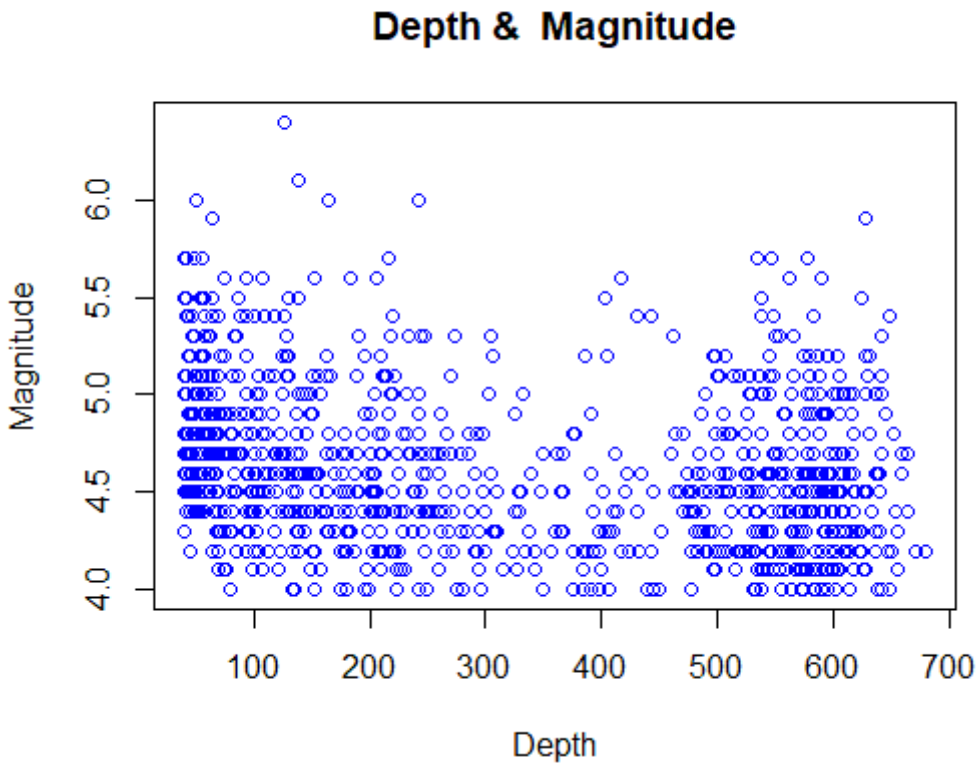
In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

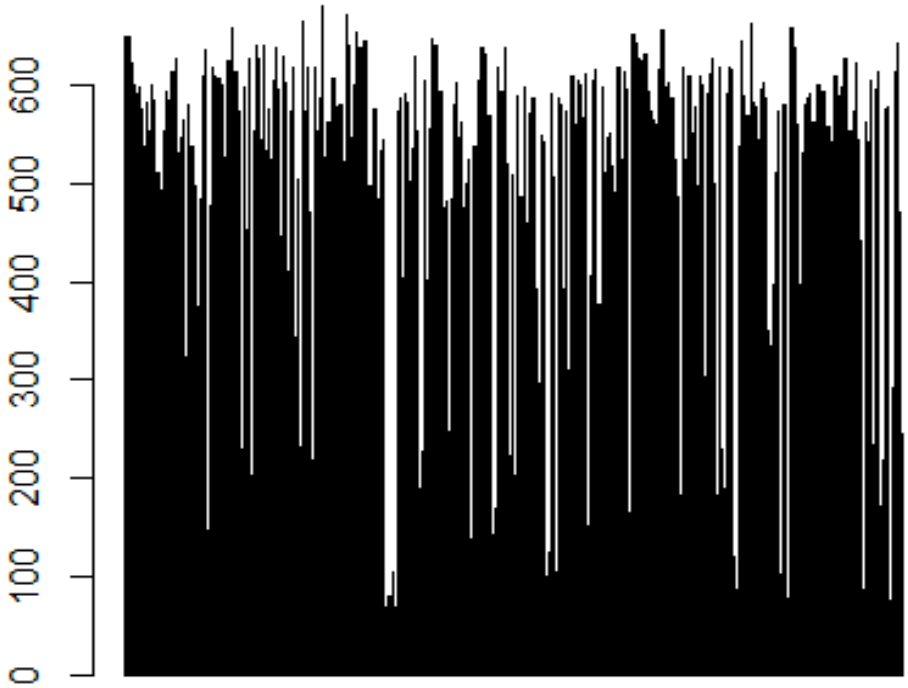
LINEAR REGRESSION ANALYSIS

Depth Vs Magnitude

```
t1 <- d1[,4]

t2 <- d1[,5]

relation <- lm(t2~t1)

png(file = "linearregression.png")

dev.off()

plot(t1,t2,col = "blue",main = "Depth &  Magnitude",

   xlab = "Depth",ylab = "Magnitude"
```



barplot(t1,t2)



Histogram

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson. It differs from a bar graph, in the sense that a bar graph relates two variables, but a histogram relates only one. To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable.

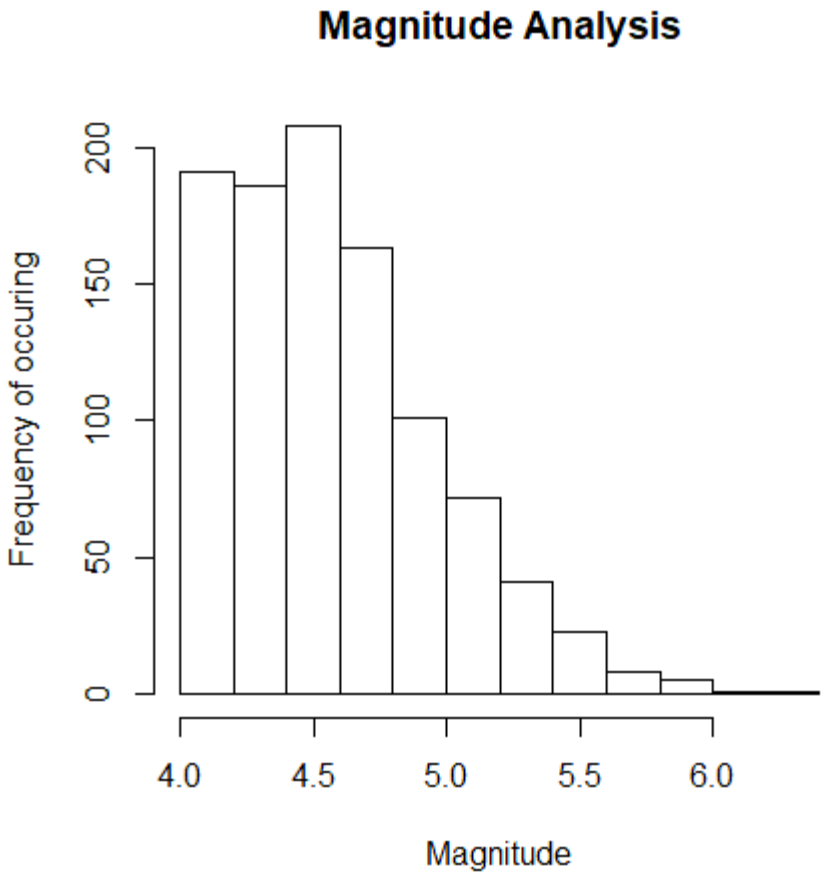
If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency—the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency of cases in the bin. The vertical axis is then not the frequency but frequency density—the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

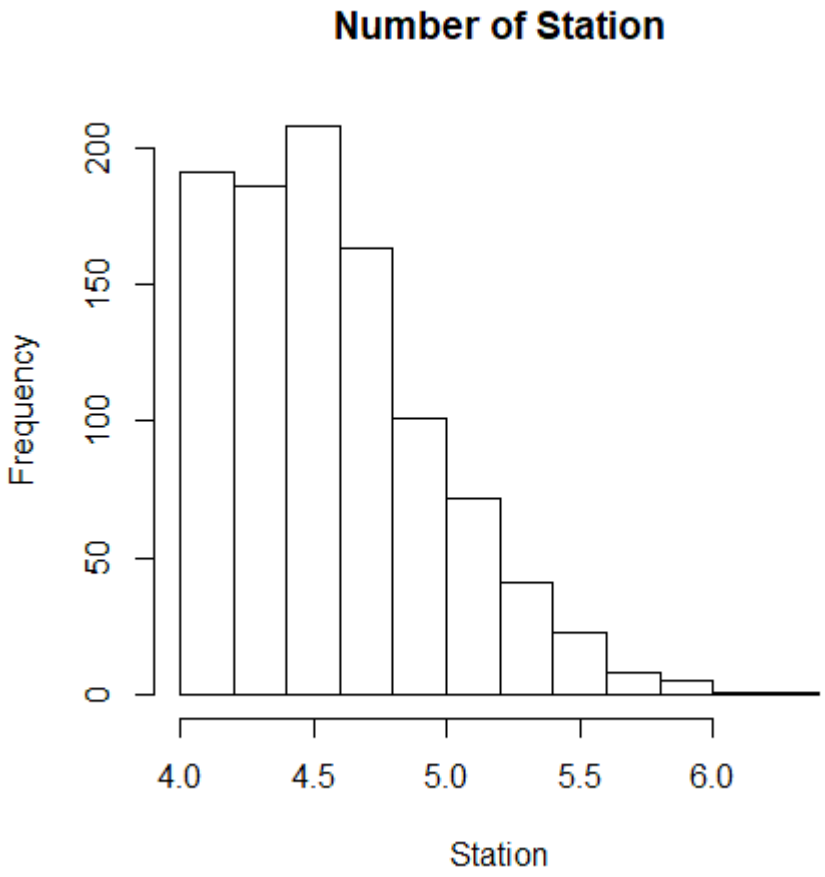
As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.

HISTOGRAM

hist(d1$mag,xlab="Magnitude",ylab="Frequency of occuring",main="Magnitude Analysis")



```
hist(d1$mag,xlab="Station",ylab="Frequency",main="Number of Station")
```



Regression

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. Regression analysis provides detailed insight that can be applied to further improve products and services.

Multiple Regression

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term. Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.

The multiple regression equation explained above takes the following form:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c.$$

Here, b_i's (i=1,2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes.

MULTIPLE REGRESSION ANALYSIS

```
> model <- lm(mag~lat+long+depth, data = d1)
> print(model)

Call:
lm(formula = mag ~ lat + long + depth, data = d1)

Coefficients:
(Intercept)         lat        long       depth
  6.7532716   -0.0089390   -0.0122630   -0.0003746



> cat("# # # # The Coefficient Values # # # ","\n")
# # # # The Coefficient Values # # #

> a <- coef(model)[1]
> print(a)
(Intercept)
   6.753272
>
> Xlat <- coef(model)[2]
> Xlong <- coef(model)[3]
> Xdepth <- coef(model)[4]
> print(Xlat)
 > lat
```

```
-0.008939008
>
> print(Xlong)
      long
-0.01226301
>
> print(Xdepth)
       depth
-0.0003746435
```

Scatter plot

A scatter plot is a set of points plotted on horizontal and vertical axis.

Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables). If no correlation exists between the variables, the points appear randomly scattered on the coordinate plane. If a large correlation exists, the points concentrate near a straight line. Scatter plots are useful data visualization tools for illustrating a trend.

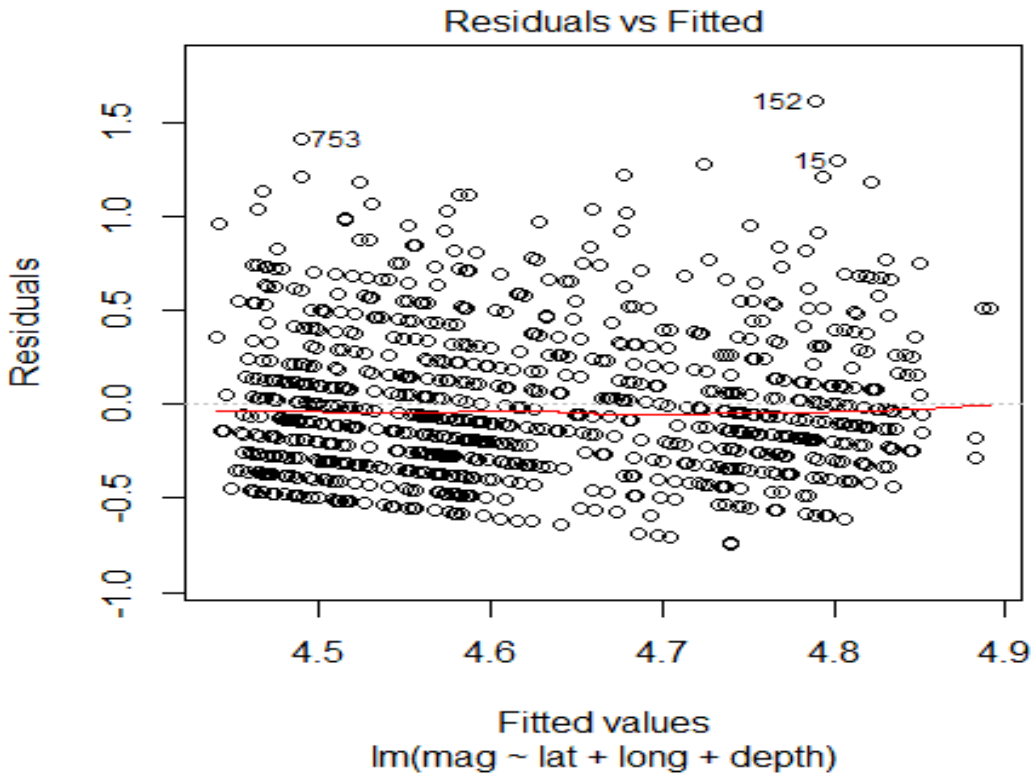
Besides showing the extent of correlation, a scatter plot shows the sense of the correlation:

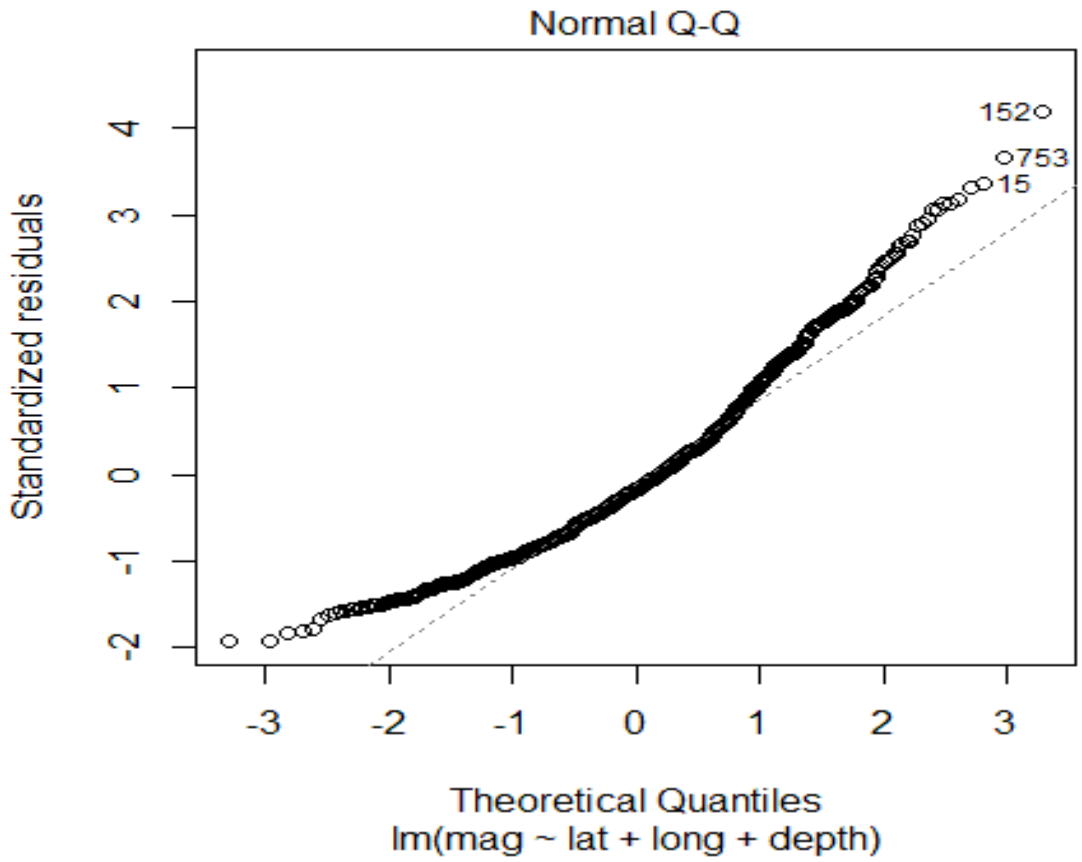
- If the vertical (or y-axis) variable increases as the horizontal (or x-axis) variable increases, the correlation is positive.
- If the y-axis variable decreases as the x-axis variable increases or vice-versa, the correlation is negative.
- If it is impossible to establish either of the above criteria, then the correlation is zero.

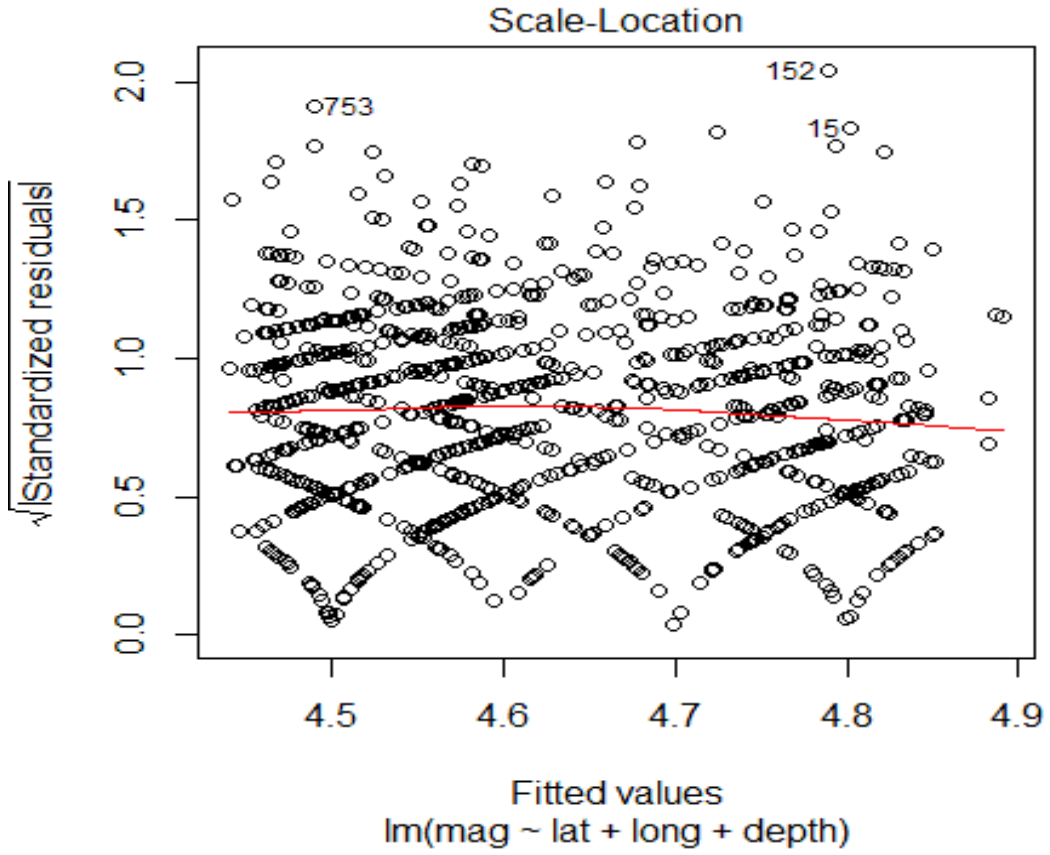
The maximum possible positive correlation is +1 or +100%, when all the points in a scatter plot lie exactly along a straight line with a positive slope. The maximum possible negative correlation is -1 or -100%, in which case all the points lie exactly along a straight line with a negative slope.

SCATTER PLOTS

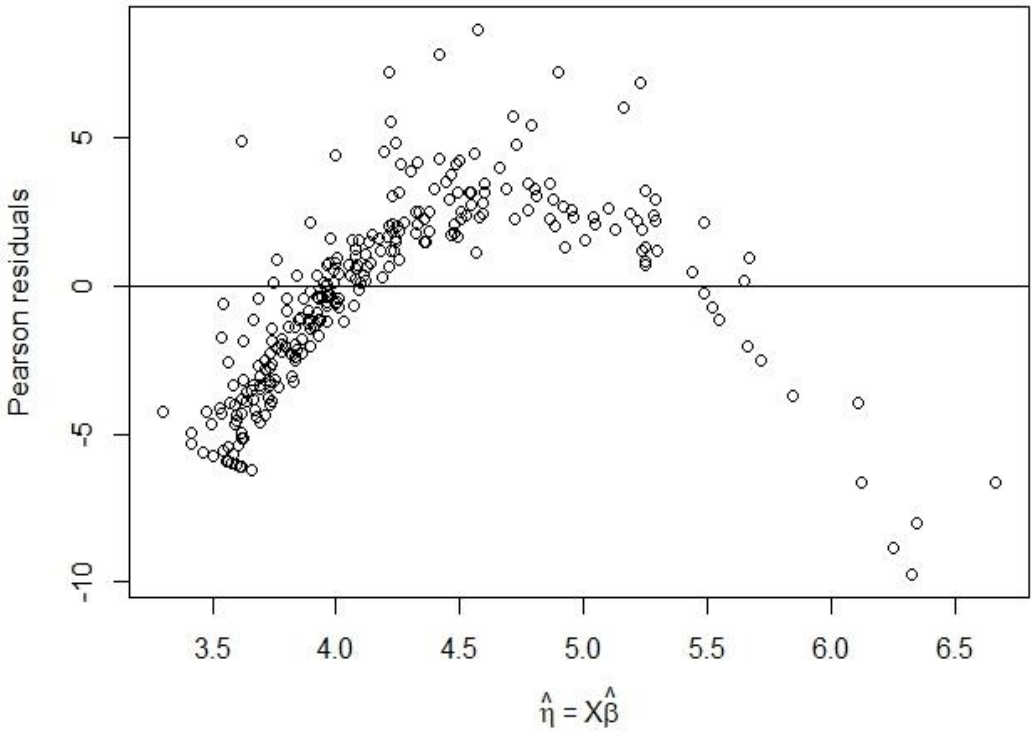
plot(model)




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mag ~ lat + long + depth)
152
753
15
-2
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
753
152
150
4.5
4.6
4.7
4.8
4.9
Fitted values
lm(mag ~ lat + long + depth)

(d)



Predictive analytics

Predictive analytics is a form of advanced analytics that uses both new and historical data to forecast activity, behavior and trends. It involves applying statistical analysis techniques, analytical queries and automated machine learning algorithms to data sets to create predictive models that place a numerical value, or score, on the likelihood of a particular event happening.

Multiple variables are combined into a predictive model capable of assessing future probabilities with an acceptable level of reliability.

PREDICTION OF MAGNITUDE

x1=-20.99

x2=181.99

x3=599

Y = a+Xlat*x1+Xlong*x2+Xdepth*x3

Y

(Intercept)

4.484745

x1=-31.99

x2=111.64

x3=450

Y = a+Xlat*x1+Xlong*x2+Xdepth*x3

Y

(Intercept)

5.501599

CONCLUSION

The analysis of this study clearly showed how the magnitude of an earthquake can be predicted, by taking into account the latitude, longitude and depth values. It uses the Multiple Regression Algorithm for this prediction. **Regression analysis** examines the influence of one or more independent variables on a dependent variable. By predicting the magnitude of an earthquake we can scale the amount of damage it could cause to man and property, so preventive measures can be taken to prevent any major losses.

FUTURE SCOPE

1. Research could be carried on to scale the damage caused by an earthquake by taking into account the earthquake size, distance from fault, site and regional geology etc.
2. Efforts could be made to study the probability of an earthquake by considering the frequency and vibrations of the tectonic plates at that site.
3. A comparative analysis on earthquake occurrence can be made to predict the magnitude of any earthquake, by applying the neural network methods.