

Prediction of Earthquake Magnitude

Sharmin Akhter

2022-11-30

1. Introduction

This project mainly purposed to predict earthquake magnitude based on earthquake depth by using Simple Linear Regression. Later I predict magnitude based on other variables. The data set give the locations of 1000 seismic events of $MB > 4.0$. The events occurred in a cube near Fiji since 1964. This is one of the Harvard PRIM-H project data sets. They in turn obtained it from Dr. John Woodhouse, Dept. of Geophysics, Harvard University.

A data frame with 1000 observations on 5 variables.

- **Mag:** predictor: Richer Magnitude
- **lat:** Numeric: Latitude of event
- **long:** Numeric: Longitude
- **depth:** Numeric: depth(km)
- **Stations** Numeric: Number of stations reporting

Simple linear Regression assumption.

The Linear Regression Model is based on several assumptions which are listed below:-

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

2. Preliminary Analysis - Data Structure, Summary and Exploratory Analysis

Import libraries

```
library(tidyverse)
library(ggthemes)
library(ggrepel)
library(dplyr)
library(corrplot)
library(MASS)
library(olsrr)
```

Importing Data

```
data_quakes<- read.csv("dataset-45892.csv")
head(data_quakes)
```

```
##      lat    long depth mag stations
## 1 -20.42 181.62   562 4.8        41
## 2 -20.62 181.03   650 4.2        15
## 3 -26.00 184.10    42 5.4        43
## 4 -17.97 181.66   626 4.1        19
## 5 -20.42 181.96   649 4.0        11
## 6 -19.68 184.31   195 4.0        12
```

```
attach(data_quakes)
```

Relocate Predictor

```
data_quakes <- data_quakes %>% relocate(mag,.after= stations)
head(data_quakes)
```

```
##      lat    long depth stations mag
## 1 -20.42 181.62   562        41 4.8
## 2 -20.62 181.03   650        15 4.2
## 3 -26.00 184.10    42        43 5.4
## 4 -17.97 181.66   626        19 4.1
## 5 -20.42 181.96   649        11 4.0
## 6 -19.68 184.31   195        12 4.0
```

Dimension of Data

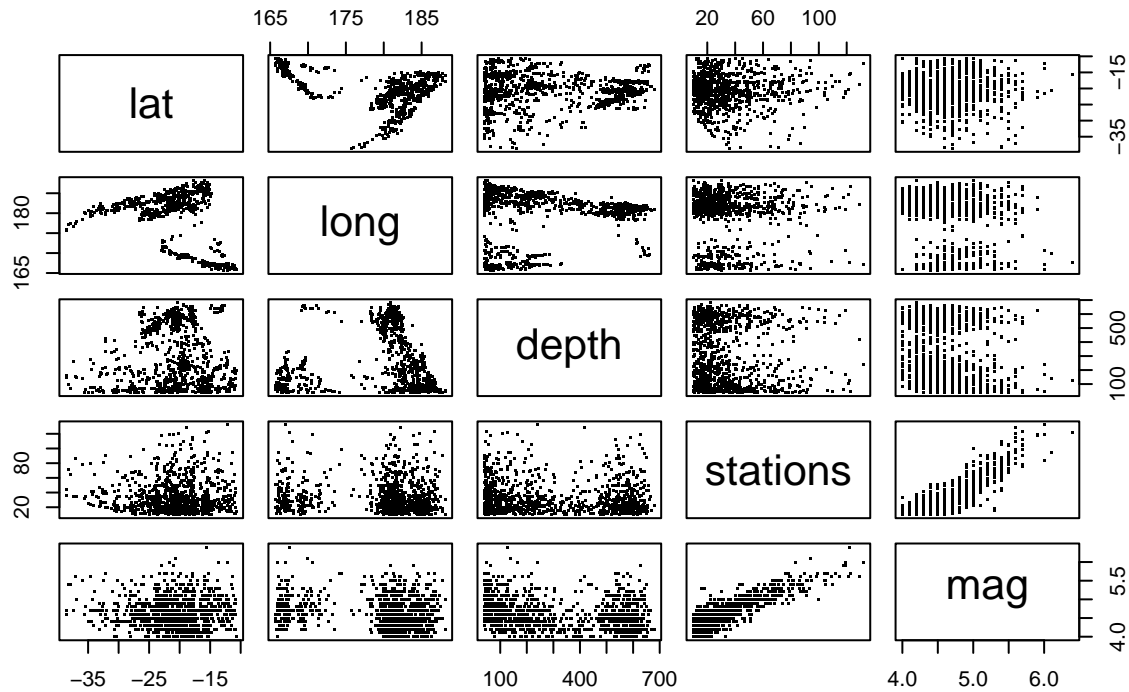
```
dim(data_quakes)
```

```
## [1] 1000    5
```

Scatterplot of the datasets

```
pairs(data_quakes, main = "Fiji Earthquakes, N = 1000", cex.main = 1.2, pch = ".")
```

Fiji Earthquakes, N = 1000



Structure of the data

```
#View(data_quakes)
str(data_quakes)

## 'data.frame':  1000 obs. of  5 variables:
## $ 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 ...
## $ stations: int    41 15 43 19 11 12 43 15 35 19 ...
## $ mag      : num   4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...

sum(is.na(data_quakes))

## [1] 0
```

From the structure we can see that the all variables values are numeric.

Now to get the inside idea we will look summary of the data

```
summary(data_quakes)
```

	lat	long	depth	stations
## Min.	-38.59	Min. :165.7	Min. : 40.0	Min. : 10.00
## 1st Qu.	-23.47	1st Qu.:179.6	1st Qu.: 99.0	1st Qu.: 18.00
## Median	-20.30	Median :181.4	Median :247.0	Median : 27.00
## Mean	-20.64	Mean :179.5	Mean :311.4	Mean : 33.42
## 3rd Qu.	-17.64	3rd Qu.:183.2	3rd Qu.:543.0	3rd Qu.: 42.00
## Max.	-10.72	Max. :188.1	Max. :680.0	Max. :132.00

```
##      mag
## Min.   :4.00
## 1st Qu.:4.30
## Median :4.60
## Mean   :4.62
## 3rd Qu.:4.90
## Max.   :6.40
```

Check for relationship between mag and depth

```
mean(data_quakes$mag[data_quakes$depth>median(data_quakes$depth)])
```

```
## [1] 4.5232
```

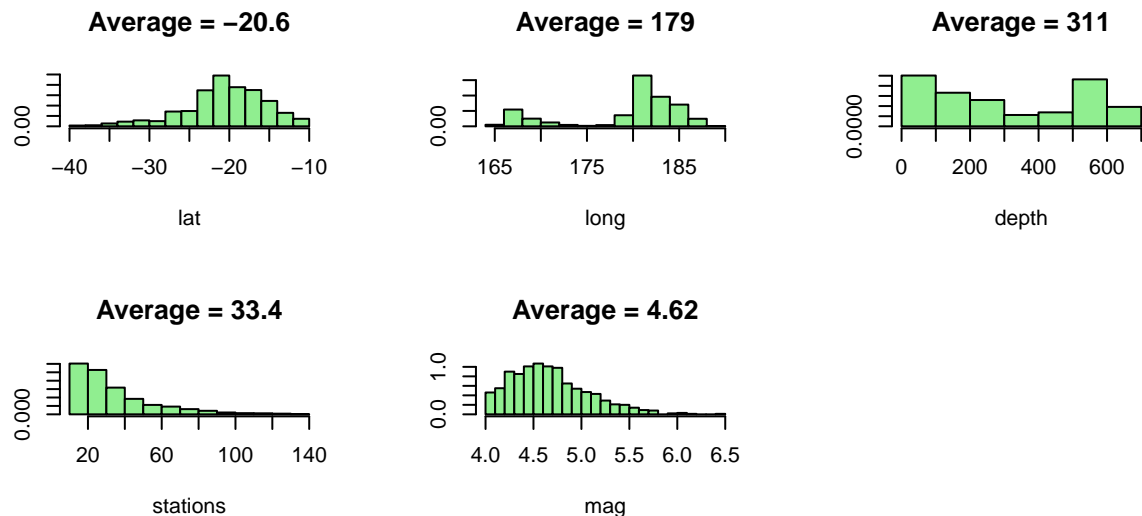
```
mean(data_quakes$mag[data_quakes$depth<median(data_quakes$depth)])
```

```
## [1] 4.7176
```

From the above mean we can see there is an inverse relationship between an earthquake's depth and its magnitude.

Now we look at the variables distribution:

```
pred = par(mfrow = c(3,3))
for ( i in 1:5 ) {
  truehist(data_quakes[[i]], xlab = names(data_quakes)[i], col = 'lightgreen', main = paste("Average = ")
}
#pred
```

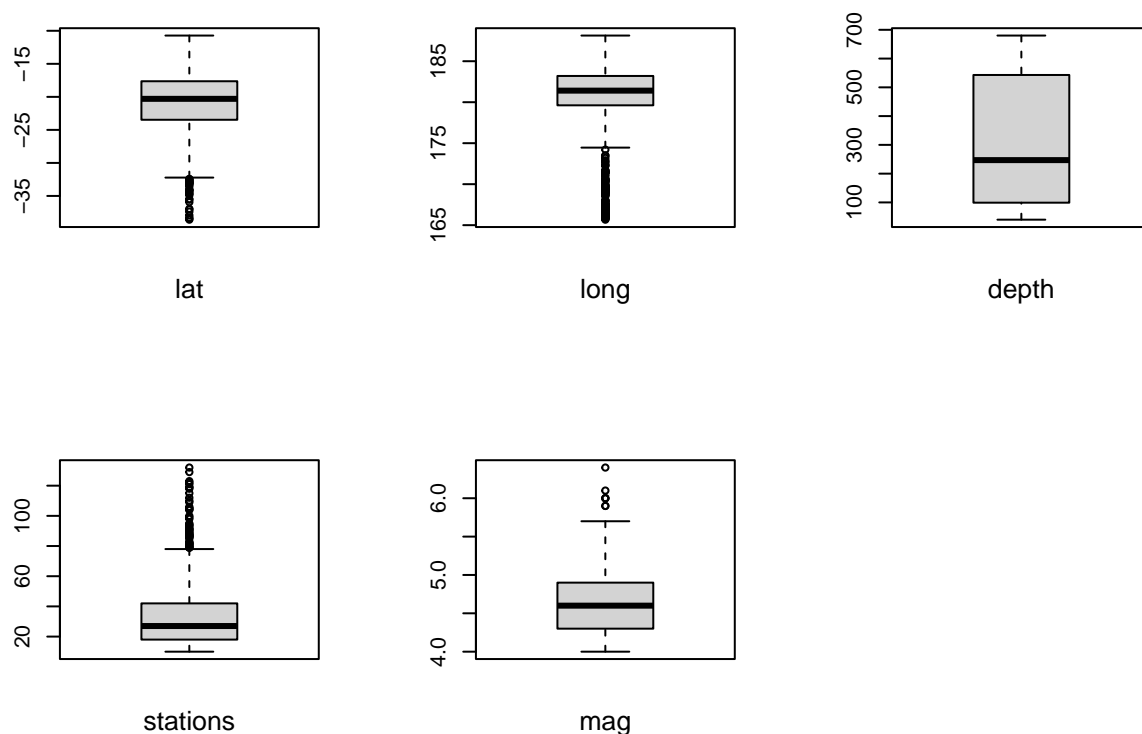


From the above distribution we can say that variable are asymmetric.

Next I will produce boxplot for each of the variables to see the outliers

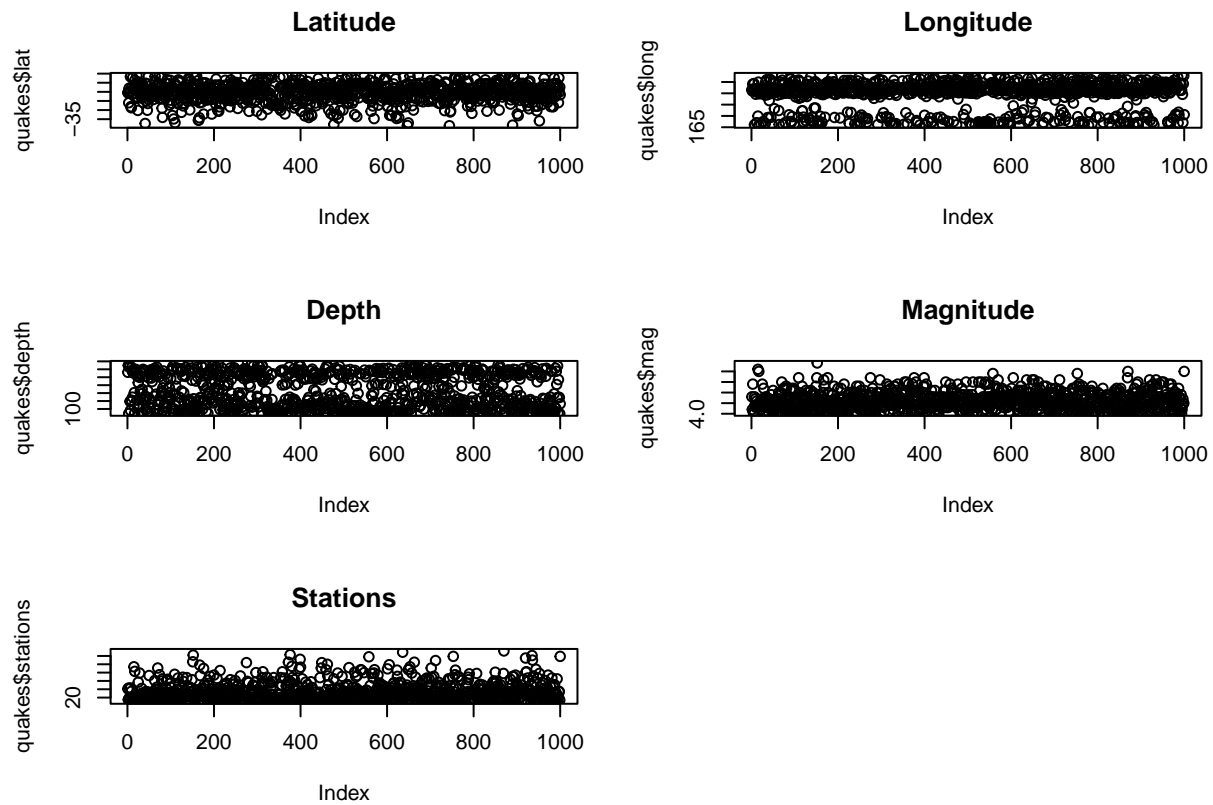
For each variables, we consider observations that lie outside $1.5 * \text{IQR}$ as outliers.

```
par(mfrow = c(2,3))
for ( i in 1:5) {
  boxplot(data_quakes[[i]])
  mtext(names(data_quakes)[i], cex = 0.8, side = 1, line = 2)
}
```



Create a histogram to get better information which boxplot can't provide

```
par(mfrow = c(3, 2))
plot(quakes$lat, main="Latitude")
plot(quakes$long, main="Longitude")
plot(quakes$depth, main="Depth")
plot(quakes$mag, main="Magnitude")
plot(quakes$stations, main="Stations")
```

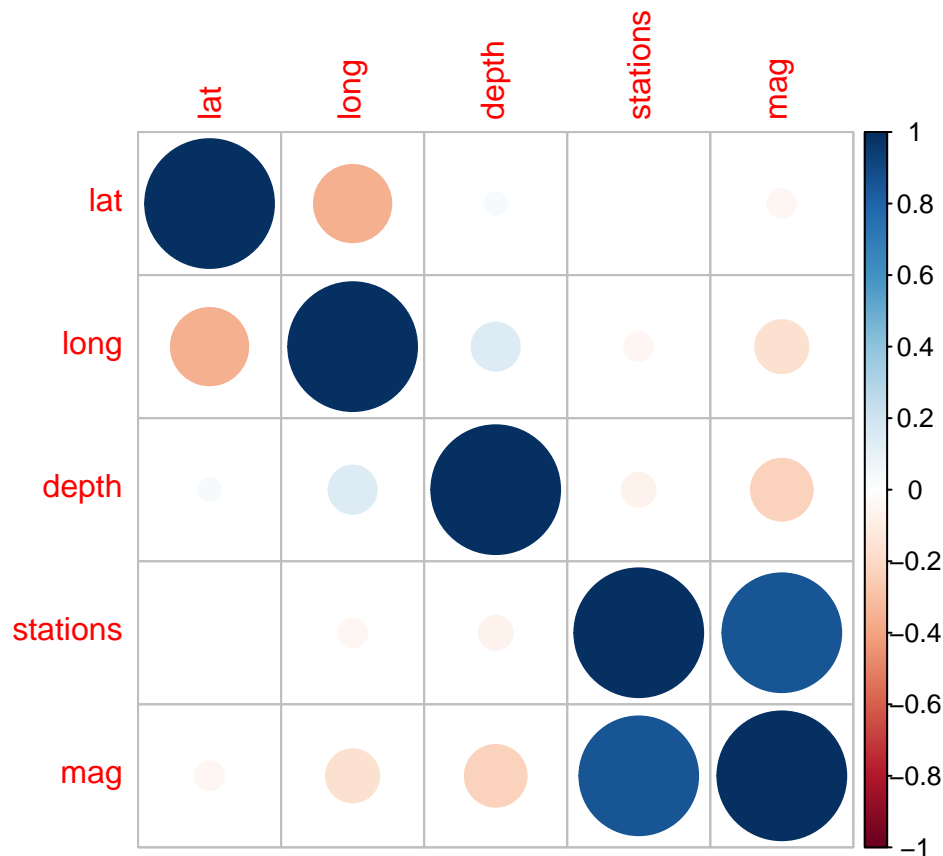


Correlation matrix

```
cormat<- round(cor(data_quakes),2)
cormat
```

```
##      lat  long depth stations  mag
## lat    1.00 -0.36  0.03    0.00 -0.05
## long   -0.36  1.00  0.14   -0.05 -0.17
## depth    0.03  0.14  1.00   -0.07 -0.23
## stations 0.00 -0.05 -0.07    1.00  0.85
## mag     -0.05 -0.17 -0.23    0.85  1.00
```

```
corrplot(cormat)
```



From the above correlation matrix we can say that mag has negative correlation with latitude and depth, it has positive correlation with longitude and no correlation with stations.

Outlier detection

```
outliers = c()
for ( i in 1:5 ) {
  stats = boxplot.stats(data_quakes[[i]])$stats
  bottom_outlier_rows = which(data_quakes[[i]] < stats[1])
  top_outlier_rows = which(data_quakes[[i]] > stats[5])
  outliers = c(outliers , top_outlier_rows[ !top_outlier_rows %in% outliers] )
  outliers = c(outliers , bottom_outlier_rows[ !bottom_outlier_rows %in% outliers] )
}
outliers
```

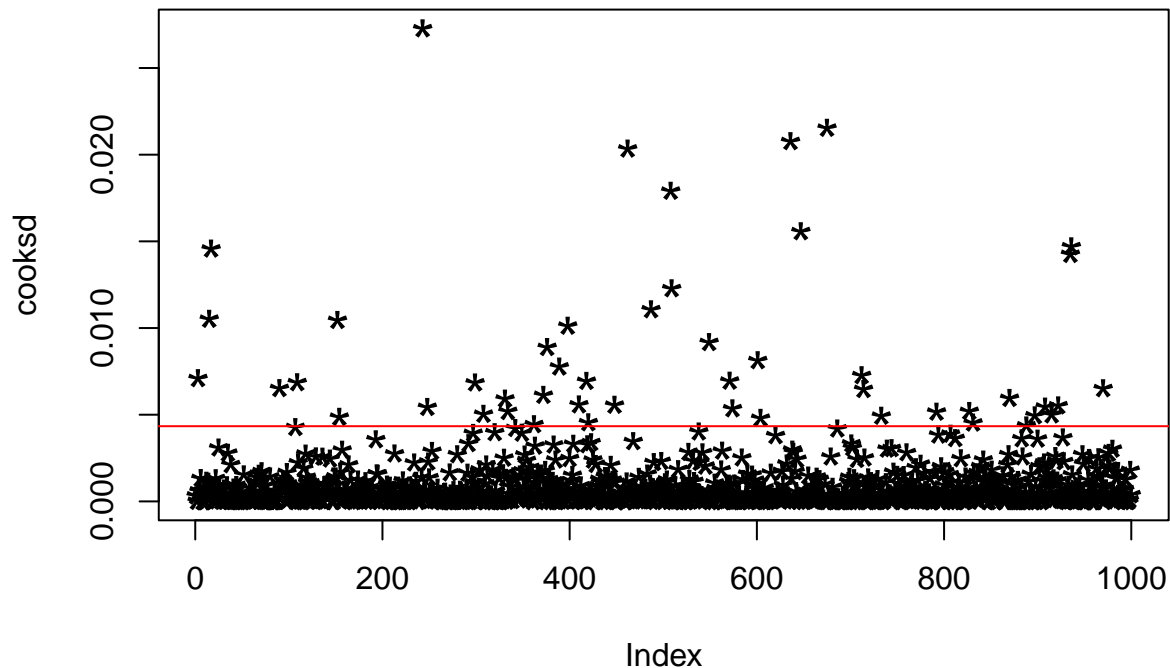
```
## [1] 41 81 104 107 110 164 165 166 310 410 418 419 425 426 476
## [16] 477 484 487 525 530 570 606 610 621 622 627 647 649 744 890
## [31] 903 952 7 12 15 17 22 27 32 37 40 45 48 53 63
## [46] 64 73 78 87 91 92 94 99 108 117 118 119 120 121 126
## [61] 133 136 141 143 145 148 152 154 155 157 159 160 163 170 192
## [76] 205 222 226 230 239 243 250 251 252 254 258 263 267 268 292
## [91] 300 301 305 311 312 318 320 321 325 328 330 334 352 357 360
## [106] 365 381 382 384 389 400 402 408 413 416 417 429 437 441 443
## [121] 453 456 467 474 490 492 496 504 507 508 509 517 524 527 528
## [136] 531 532 534 536 538 539 541 542 543 544 545 546 547 552 553
## [151] 560 571 581 583 587 593 594 596 597 612 613 618 620 625 629
```

```
## [166] 638 642 653 655 656 672 675 681 686 699 701 712 714 716 721
## [181] 725 726 735 754 756 759 765 766 769 779 781 782 787 797 804
## [196] 813 825 827 837 840 844 852 853 857 860 865 866 869 870 872
## [211] 873 883 884 887 888 891 893 908 909 912 915 916 921 927 930
## [226] 962 963 969 974 980 982 986 987 988 997 1000 28 70 151 167
## [241] 176 214 275 297 313 358 372 376 380 397 399 448 449 459 462
## [256] 486 512 558 601 605 623 636 651 657 663 702 753 758 850 920
## [271] 922 935 936 944
```

We use the Cook's distance to detect influential observations.

```
mod = lm(mag ~ ., data = data_quakes)
cooks = cooks.distance(mod)
plot(cooks, pch = "*", cex = 2, main = "Influential Obs by Cooks distance")
abline(h = 4*mean(cooks, na.rm = T), col = "red")
```

Influential Obs by Cooks distance



Clean outliers

```
clean_outliers = as.numeric(rownames(data_quakes[cooks > 4 * mean(cooks, na.rm=T), ]))
outliers = c(outliers, clean_outliers[!clean_outliers %in% outliers])

clean_Data = data_quakes[-outliers, ]
summary(clean_Data)
```

```
##      lat      long      depth      stations
## Min.   :-32.20 Min.   :177.8 Min.   : 40.0 Min.   :10.00
## 1st Qu.: -23.74 1st Qu.:181.0 1st Qu.:143.0 1st Qu.:17.00
## Median : -20.88 Median :182.0 Median :407.5 Median :23.50
```



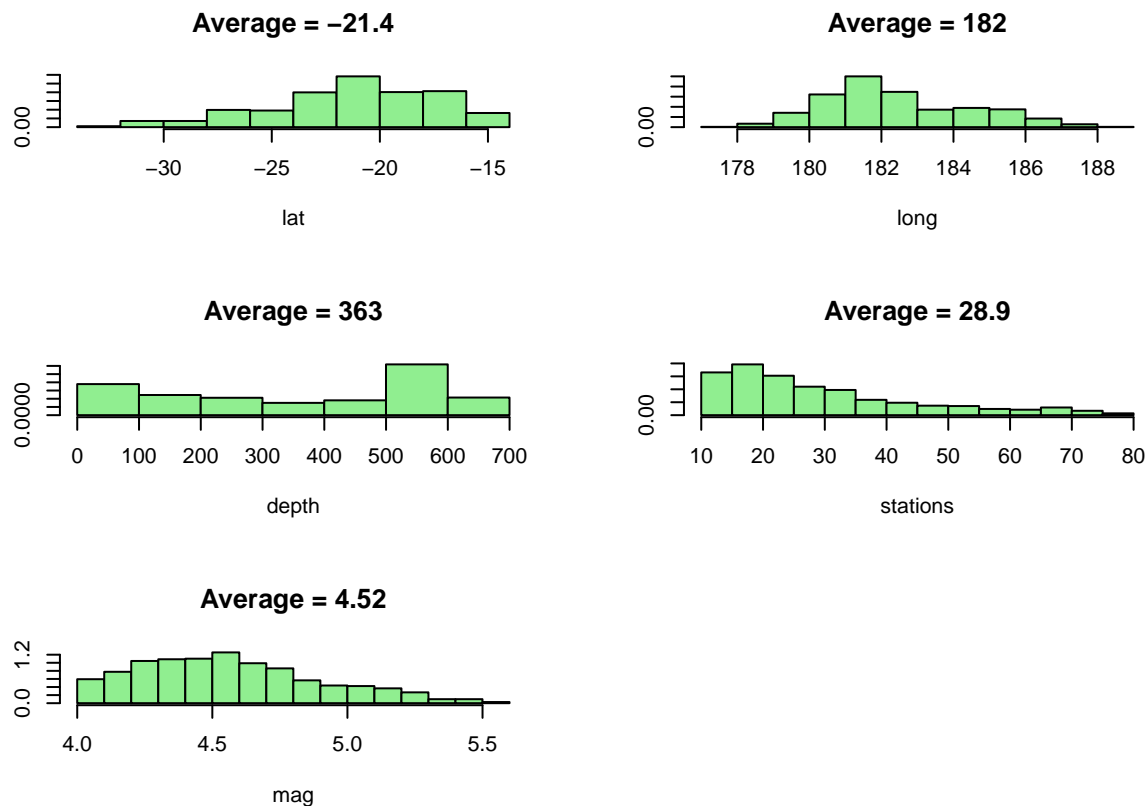
```
## Mean      :-21.36    Mean      :182.4    Mean      :362.7    Mean      :28.93
## 3rd Qu.   :-18.12    3rd Qu. :183.9    3rd Qu. :562.0    3rd Qu. :36.00
## Max.      :-14.85    Max.     :188.1    Max.     :680.0    Max.     :78.00
##          mag
## Min.      :4.000
## 1st Qu.   :4.300
## Median    :4.500
## Mean      :4.518
## 3rd Qu.   :4.700
## Max.      :5.500
```

```
str(clean_Data)
```

```
## 'data.frame': 708 obs. of 5 variables:
## $ lat      : num -20.4 -20.6 -18 -20.4 -19.7 ...
## $ long      : num 182 181 182 182 184 ...
## $ depth     : int 562 650 626 649 195 194 211 622 583 554 ...
## $ stations  : int 41 15 19 11 12 15 35 19 13 19 ...
## $ mag       : num 4.8 4.2 4.1 4 4 4.4 4.7 4.3 4.4 4.4 ...
```

Histogram plot after remove outliers

```
par(mfrow=c(3,2))
for ( i in 1:5 ) {
  truehist(clean_Data[[i]], xlab = names(clean_Data)[i], col = 'lightgreen', main = paste("Average =", ,
})
```



By removing the outliers, the dataset size reduced to 708 observations of 5 variables. Now, the variables are

approximately normally distributed. By comparing with the previous histogram that contains high influence outliers we can see that the skewness is reduced in the new histogram.

2.1 Model building

Now we fit a simple linear regression model to predict earthquake magnitude based on earthquake depth.

$$mag = \beta_0 + \beta_1.depth$$

```
lmmod<- lm(mag~depth, data = clean_Data)
summary(lmmod)

##
## Call:
## lm(formula = mag ~ depth, data = clean_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59315 -0.26308 -0.05362  0.19998  0.93990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.614e+00  2.474e-02 186.477  < 2e-16 ***
## depth       -2.645e-04  5.896e-05  -4.486  8.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3311 on 706 degrees of freedom
## Multiple R-squared:  0.02771,    Adjusted R-squared:  0.02634
## F-statistic: 20.12 on 1 and 706 DF,  p-value: 8.476e-06

anova(lmmod)

## Analysis of Variance Table
##
## Response: mag
##      Df Sum Sq Mean Sq F value    Pr(>F)
## depth    1  2.206  2.20597   20.123 8.476e-06 ***
## Residuals 706 77.395  0.10963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted $R^2 = 0.02634$, RSE= 0.3311 so the model is not good fit.

2.1 To check for non constant variance

ANOVA for reduced model

```
red<- resid(lmmod)
rs<- red^2
red.lm<- lm(rs~depth, data = clean_Data[1:708,])
summary(red.lm)

##
```

```
## Call:
## lm(formula = rs ~ depth, data = clean_Data[1:708, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11005 -0.09598 -0.05240  0.03625  0.77461
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.102e-01  1.071e-02  10.285  <2e-16 ***
## depth       -2.332e-06  2.552e-05  -0.091   0.927
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1433 on 706 degrees of freedom
## Multiple R-squared:  1.183e-05, Adjusted R-squared:  -0.001405
## F-statistic: 0.008351 on 1 and 706 DF, p-value: 0.9272
anova(red.lm)

## Analysis of Variance Table
##
## Response: rs
##           Df Sum Sq Mean Sq F value Pr(>F)
## depth      1  0.0002  0.0001715   0.0084 0.9272
## Residuals 706 14.5005  0.0205390
```

Breush-Pagan test for constancy of error variance

$$\chi_0 = \frac{n^2}{2} * \frac{SSR^*}{SSE^2}$$

From above anova table and summary we have $n = 708$, $SSR = 14.500$, $SSE^* = 77.395$

```
chi0 <-((708^2)/2) * (14.500/(77.395^2))
chi0
```

```
## [1] 606.7066
```

```
chi_crit <- qchisq(0.95,706)
chi_crit
```

```
## [1] 768.924
```

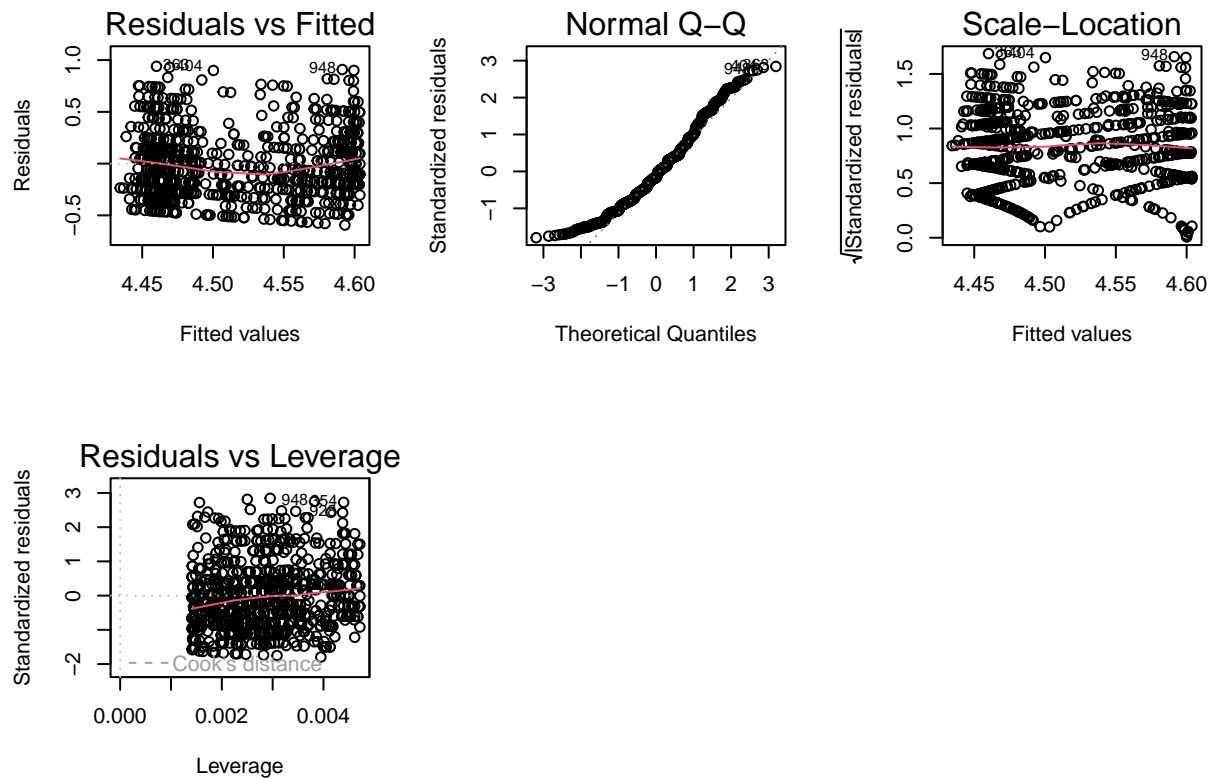
Hypothesis(null and alternative):

$H_0: \beta_1 = 0$ and $H_1: \beta_1 \neq 0$

Since the $\chi_0 = 606.7066 < \chi_{crit} = 768.924$ so we can reject null hypothesis.

Plot for full model without outliers in the dataset to predict magnitude based on depth

```
par(mfrow=c(2,3))
plot(lmmod)
```



Based on the above graphs, we observe the following -

- Residual vs fitted: There is curvature in the plot indicating that there is non linear relationship in the datasets.
- The normal Q-Q plot shows a fairly straight line, indicating the errors are more-or-less normally distributed.

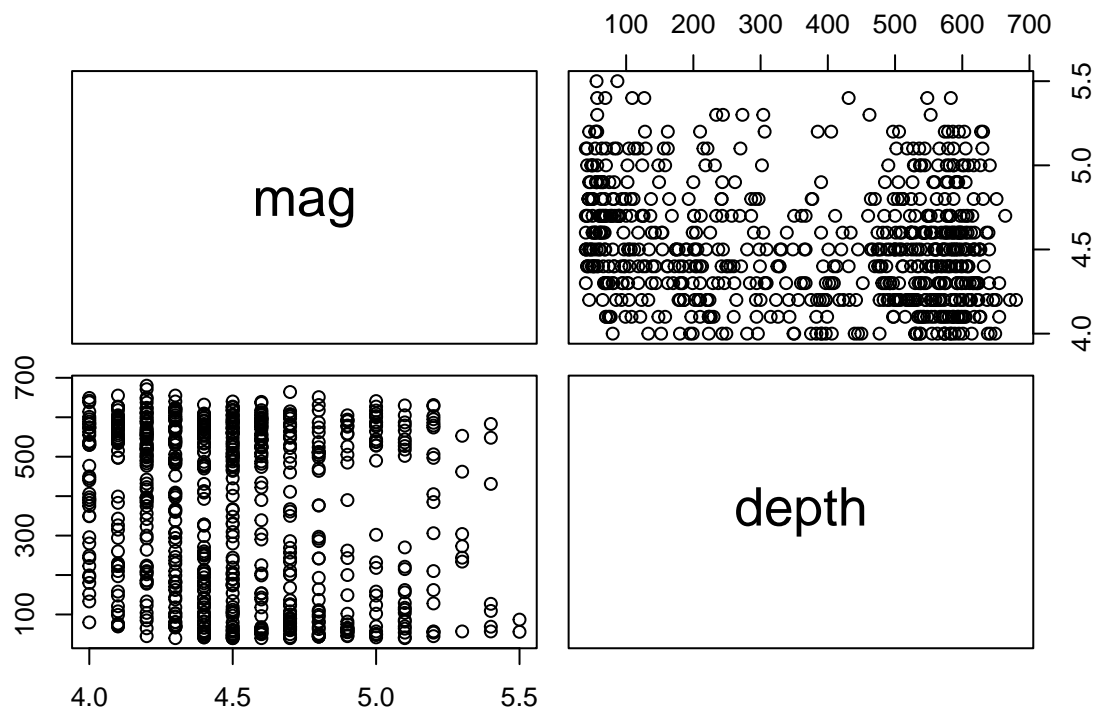
Based on the above summary of the fitted model we make the following observations:

- The multiple R- squared of the full and reduced model is 0.02771 and 1.183e-05. Adjusted R-square value is 0.02634 and -0.001405 respectively for full and reduced model.
- Since the errors seem to follow normal distribution based on Q-Q plot so taking level of significance 0.01.

Identify multicollinearity of with.

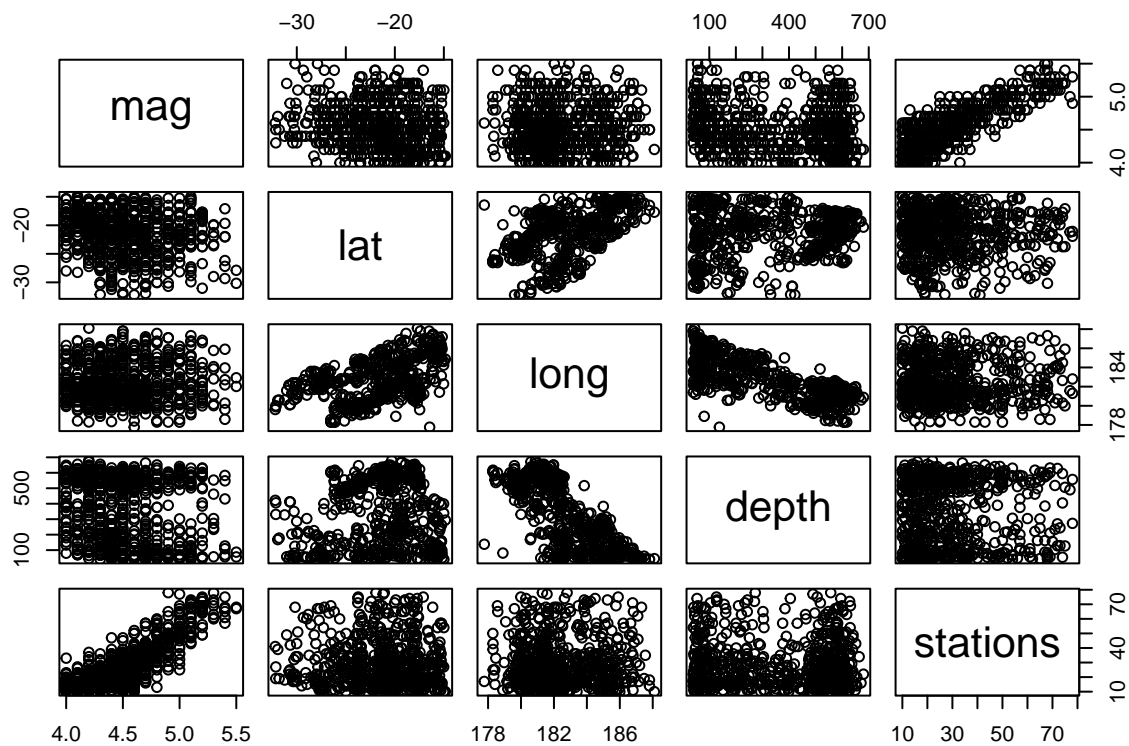
Now we look for the deeper analysis of the data

```
pairs(mag~depth, data = clean_Data[1:708,])
```



With all variables

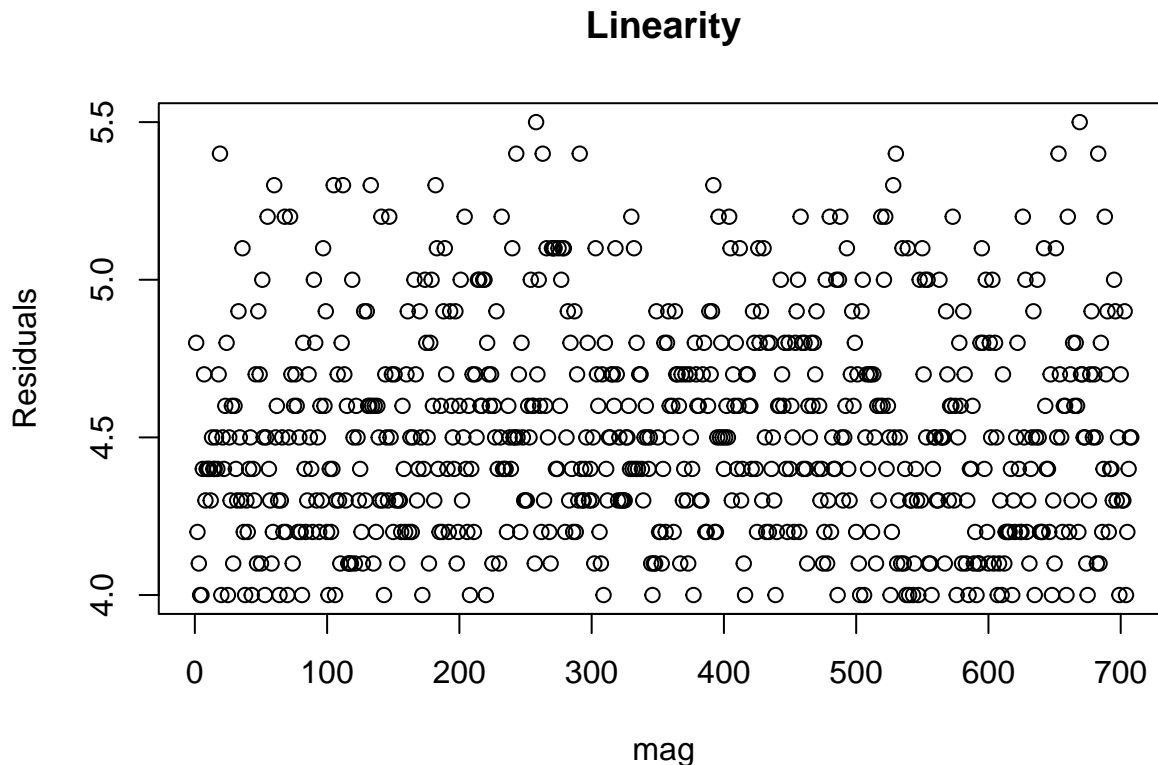
```
pairs(mag~., data = clean_Data[1:708,])
```



Multicollinearity occurs when the features (or independent variables) are highly correlated

Linearity

```
plot(clean_Data$mag,clean_Data$residuals,xlab="mag",ylab="Residuals",main="Linearity")
```



from the scatterplot we can see that the relationship between response and feature variables is linear.

Further analysis

Multiple Linear Regression

Now, we fit a multiple linear regression model with mag as the response and all other variables as regressors. We plot the basic summary plots based on the fitted model, lmmod1, say, to get more idea about the data

```
lmmod1<-lm(mag~.,data=clean_Data[1:708,])  
summary(lmmod1)
```

```
##  
## Call:  
## lm(formula = mag ~ ., data = clean_Data[1:708, ])  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -0.51312 -0.12390 -0.00318  0.12307  0.45420   
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.507e+00  1.660e+00   2.716 0.006772 **
## lat         -1.175e-02  3.117e-03  -3.769 0.000178 ***
## long        -3.602e-03  8.657e-03  -0.416 0.677453
## depth       -2.124e-04  7.662e-05  -2.772 0.005717 **
## stations     1.708e-02  4.044e-04  42.235 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1737 on 703 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.732
## F-statistic: 483.7 on 4 and 703 DF,  p-value: < 2.2e-16

anova(lmmod1)

## Analysis of Variance Table
##
## Response: mag
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## lat         1  2.971   2.971   98.4350 < 2.2e-16 ***
## long        1  1.538   1.538   50.9706 2.344e-12 ***
## depth       1  0.047   0.047    1.5494  0.2136
## stations    1 53.831  53.831 1783.8104 < 2.2e-16 ***
## Residuals 703 21.215   0.030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusted $R^2 = 0.732$, $RSE = 0.1737$

2.2 Check for non constant variance

ANOVA for reduced model

```
red<- resid(lmmod1)
rs<- red^2
red.lm<- lm(rs~ lat+long+depth+stations, data = clean_Data[1:708,])
summary(red.lm)

##
## Call:
## lm(formula = rs ~ lat + long + depth + stations, data = clean_Data[1:708,
##    ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03721 -0.02607 -0.01411  0.01220  0.23350
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.623e-01  3.664e-01   0.716   0.474
## lat         1.502e-04  6.881e-04   0.218   0.827
## long        -1.228e-03  1.911e-03  -0.643   0.521
## depth       -1.121e-05  1.692e-05  -0.663   0.508
## stations    -3.449e-05  8.928e-05  -0.386   0.699
```

```
##
## Residual standard error: 0.03835 on 703 degrees of freedom
## Multiple R-squared: 0.001375, Adjusted R-squared: -0.004307
## F-statistic: 0.242 on 4 and 703 DF, p-value: 0.9145
```

```
anova(red.lm)
```

```
## Analysis of Variance Table
##
## Response: rs
##           Df Sum Sq   Mean Sq F value Pr(>F)
## lat        1 0.00051 0.00051145  0.3477 0.5556
## long       1 0.00002 0.00001873  0.0127 0.9102
## depth      1 0.00067 0.00067415  0.4584 0.4986
## stations   1 0.00022 0.00021943  0.1492 0.6994
## Residuals 703 1.03394 0.00147075
```

Breush-Pagan test for constancy of error variance

$$\chi_0 = \frac{n^2}{2} * \frac{SSR^*}{SSE^2}$$

From above anova table and summary we have $n = 708$, $SSR = 1.03394$, $SSE^* = 21.215$

```
chi0 <-((708^2)/2) * (1.03394/(21.215^2))
chi0
```

```
## [1] 575.7657
```

```
chi_crit <- qchisq(0.95,703)
chi_crit
```

```
## [1] 765.7925
```

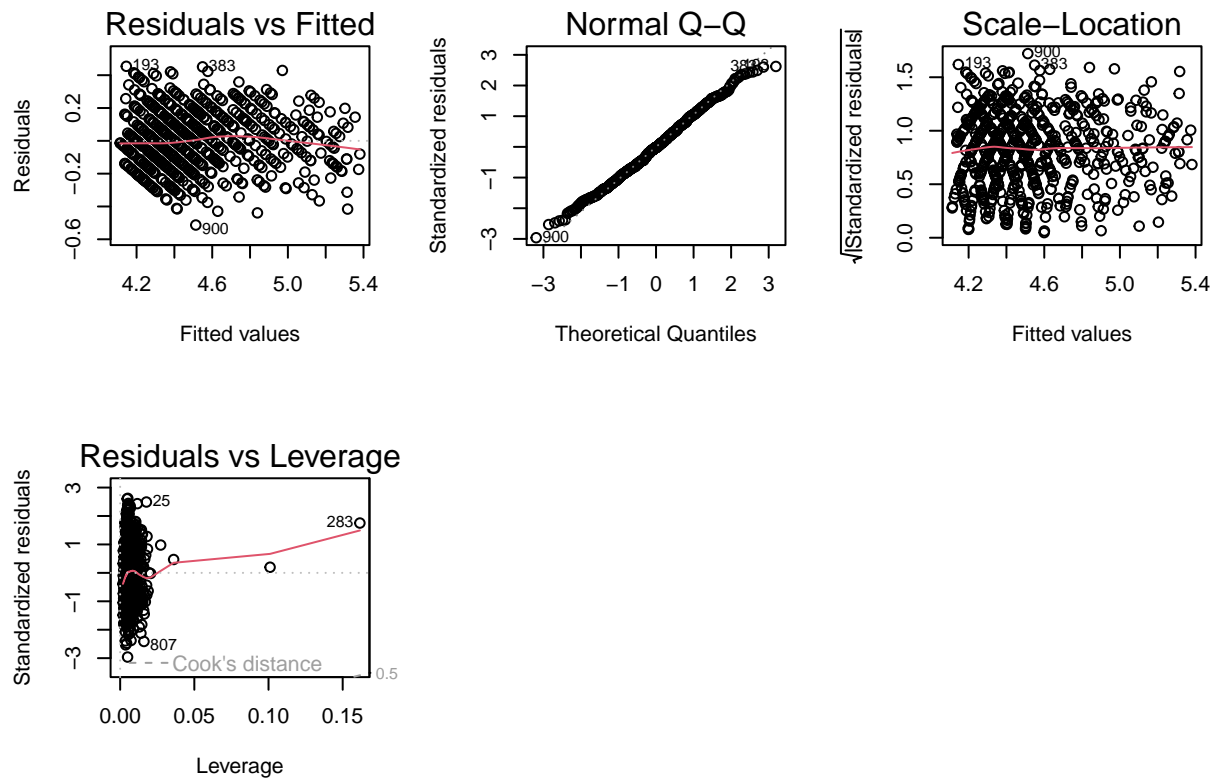
Hypothesis(null and alternative):

$H_0: \beta_1 = 0$ and $H_1: \beta_1 \neq 0$

Since the $\chi_0 = 575.7657 < \chi_{crit} = 765.7925$ so we can reject null hypothesis.

Plot for full model without outliers

```
par(mfrow=c(2,3))
plot(lmmod1)
```

Based on the above graphs, we observe the following -

- Residual vs fitted: There is curvature in the plot indicating that there is non linear relationship in the datasets.
- The normal Q-Q plot shows a fairly straight line, indicating the errors are more-or-less normally distributed.

Based on the above summary of the fitted model we make the following observations:

- The multiple R- squared of the full and reduced model is 0.7335 and 0.001375. Adjusted R-square value is 0.732 and -0.004307 respectively for full and reduced model.
- Since the errors seem to follow normal distribution based on Q-Q plot so taking level of significance 0.01.

3. Model Selection

In this section we will develop a best subset model for predicting the Earthquakes.

```
lm.1 <- lm(mag ~ ., clean_Data)
lm.1

##
## Call:
## lm(formula = mag ~ ., data = clean_Data)
##
```

```
## Coefficients:
## (Intercept)      lat      long      depth      stations
##  4.5073439   -0.0117481  -0.0036022  -0.0002124   0.0170808
```

```
plt<- ols_step_best_subset(lm.1)
plt
```

```
##          Best Subsets Regression
```

```
## -----
## Model Index    Predictors
## -----
##      1         stations
##      2         lat stations
##      3         lat depth stations
##      4         lat long depth stations
```

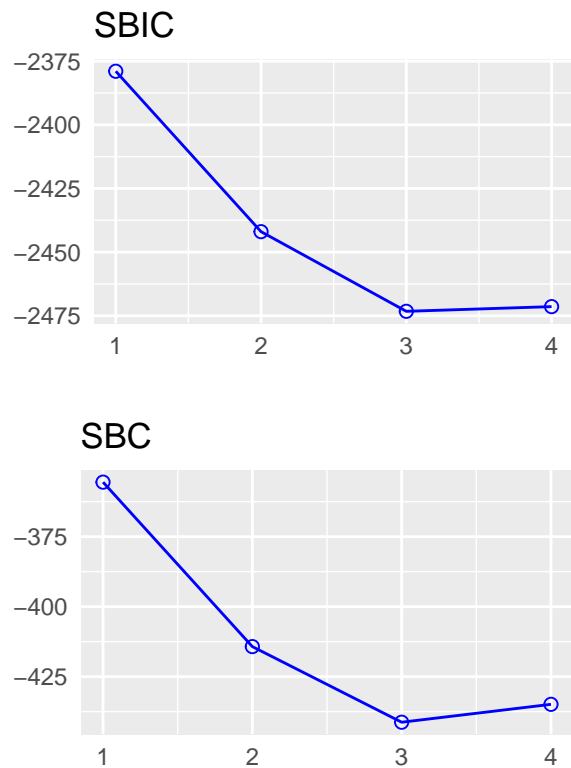
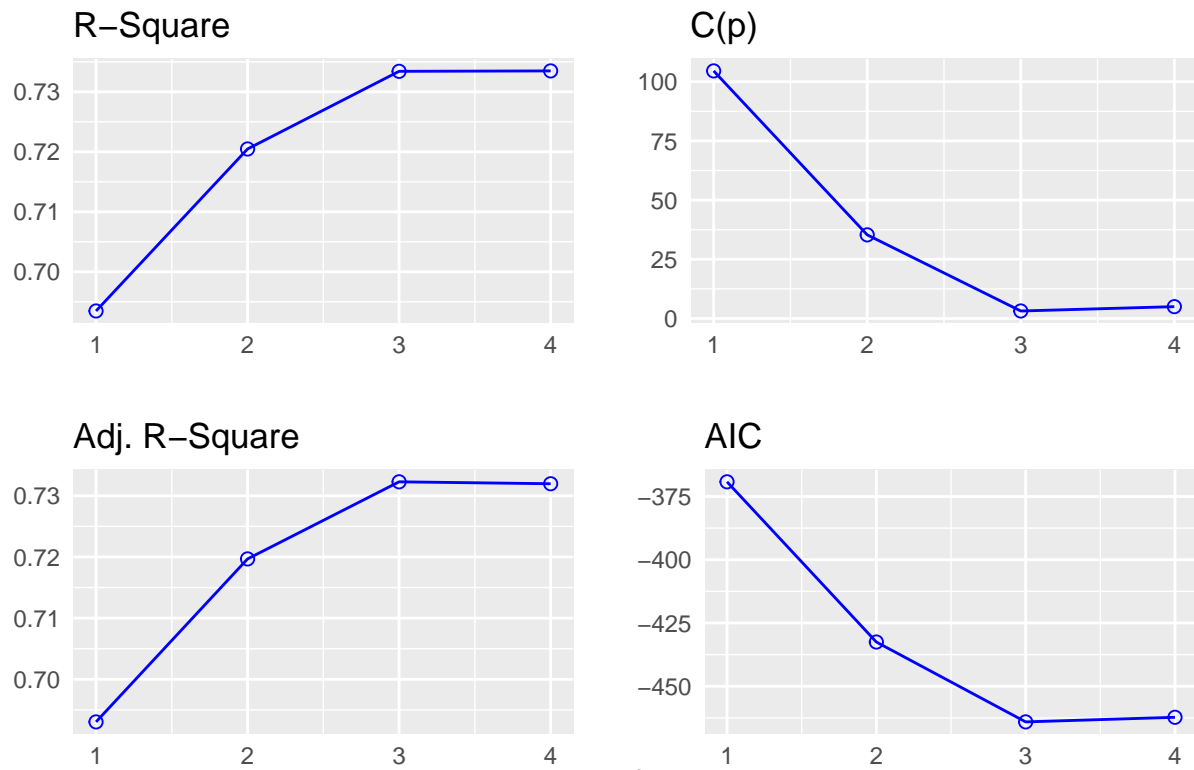
```
## -----
##
```

```
##                               Subsets Regression Summary
```

```
## -----
##                               Adj.      Pred
## Model    R-Square    R-Square    R-Square    C(p)      AIC      SBIC      SBC      MS
## -----
##      1      0.6935      0.6930      0.6918    104.5580   -369.2199  -2378.9653  -355.5326  24.
##      2      0.7205      0.7197      0.7181     35.3204   -432.5187  -2441.9774  -414.2689  22.
##      3      0.7334      0.7323      0.7304      3.1732   -464.0916  -2473.2537  -441.2793  21.
##      4      0.7335      0.7320      0.7295      5.0000   -462.2659  -2471.4119  -434.8913  21.
```

```
## -----
## AIC: Akaike Information Criteria
## SBIC: Sawa's Bayesian Information Criteria
## SBC: Schwarz Bayesian Criteria
## MSE: Estimated error of prediction, assuming multivariate normality
## FPE: Final Prediction Error
## HSP: Hocking's Sp
## APC: Amemiya Prediction Criteria
```

```
plot(plt)
```



We can see that model 4 with lat, long, depth, stations as predictor variables is selected based on R^2 adjusted criterion with highest R^2 adjusted value. $C(p)$ value leads to model 4 as for this model which is small. This 4 predictor variable model is also selected by the AIC. We can see SBC criterion which leads to model 4.

Based on the 4 criteria model 4 turned out to be the best model.

4. Summary

In this project I have predicted earthquake magnitude based on earthquake depth. For this I have collected data with 1000 observations of 5 variables and all the values in the data sets are numeric.

In the preliminary analysis part of the project I analyzed data with various method. A few approaches were taken to addressed the analysis of the data. Box plot shows the outliers of the data and then for better analysis I plotted histogram. It shows there more than 200 outliers on the datasets. I checked for influence of the outliers by cooks distance and clean the outliers.

After that I fitted a model with magnitude and depth which is a simple linear regression model. By the assumptions of linear regression i checked for multicollinearity, linearity of the variables. From the scatter plot of multicollinearity and linearity its visible that the data in linear and normally distributed.

For the further analysis I was fitted a model with earthquake magnitude based on other variables.

The diagnostic plots show an improvement over the base one. However, the performance of the model decreases as showcased by smaller R-squared and RMSE values for both methods.

In the last section the best models shows that with lat, long, depth and stations