Regression

Sujay Vadlakonda

2023 Feb 18

1. Linear Regression

A linear regression attempts to find a line that minimizes the distance between each of the data points graphed with predictors on the x axis and target column on the y axis. A linear regression seeks to predict a quantitative target. Linear regression is a high bias algorithm, which means it is prone to underfitting. A linear regression wants to see a linear relationship between the target and its predictors and cannot observe other relationships.

2. Load Data

I am using a dataset about earthquakes I found here.

```
df <- read.csv("earthquakes.csv", header=TRUE)
```

2a. Create Test and Train Data

```
set.seed(1234)
i <- sample(1:nrow(df), 0.8*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

2b. Training Data Exploration

```
str(train)
```

```
## 'data.frame':    29864 obs. of  23 variables:
##  $ Time           : chr  "1992-06-02T11:48:58.690Z" "1933-10-01T14:34:59.270Z" "1925-03-22T14:05:38.
##  $ Place          : chr  "east of the Kuril Islands" "198 km E of Hasaki, Japan" "252 km W of Abepura
##  $ Latitude       : num  47.69 35.78 -2.19 -21.5 -60.77 ...
##  $ Longitude      : num  155 143 138 170 154 ...
##  $ Depth          : num  44.6 15 15 33 33 71 10 10 36.9 15 ...
##  $ Mag            : num  5.5 5.55 6.01 5.6 5.7 5.5 5.6 5.7 6 6.07 ...
##  $ MagType        : chr  "mw" "mw" "mw" "ms" ...
##  $ nst            : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ gap            : num  NA NA NA NA NA NA 60 30 NA NA ...
##  $ dmin           : num  NA NA NA NA NA ...
##  $ rms            : num  1 NA NA 1.2 1.3 1.1 1.08 0.82 1.3 NA ...
##  $ net            : chr  "us" "iscgem" "iscgem" "us" ...
##  $ ID             : chr  "usp00058jt" "iscgem905846" "iscgem910424" "usp00038sj" ...
##  $ Updated        : chr  "2016-11-09T22:40:04.308Z" "2022-04-25T21:59:32.398Z" "2022-04-25T23:24:56.8
##  $ X              : logi  NA NA NA NA NA NA ...
##  $ Type           : chr  "earthquake" "earthquake" "earthquake" "earthquake" ...
##  $ horizontalError: num  NA NA NA NA NA NA 5.2 7.1 NA NA ...
```

```
##  $ depthError     : num  NA 7.8 14.5 NA NA NA 1.7 1.8 11.9 35.9 ...
##  $ magError       : num  NA 0.38 0.2 NA NA NA 0.073 NA NA 0.22 ...
##  $ magNst         : int  NA NA NA 4 NA 30 18 NA NA NA ...
##  $ status         : chr  "reviewed" "reviewed" "reviewed" "reviewed" ...
##  $ locationSource : chr  "us" "iscgem" "iscgem" "us" ...
##  $ magSource      : chr  "hrv" "iscgem" "iscgem" "us" ...
```

```
names(train)
```

```
##  [1] "Time"            "Place"           "Latitude"        "Longitude"
##  [5] "Depth"           "Mag"             "MagType"         "nst"
##  [9] "gap"             "dmin"            "rms"             "net"
## [13] "ID"              "Updated"         "X"               "Type"
## [17] "horizontalError" "depthError"      "magError"        "magNst"
## [21] "status"          "locationSource"  "magSource"
```

```
dim(train)
```

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

```
head(train)
```

```
##                          Time                             Place Latitude
## 15241 1992-06-02T11:48:58.690Z         east of the Kuril Islands   47.694
## 33702 1933-10-01T14:34:59.270Z         198 km E of Hasaki, Japan   35.783
## 35716 1925-03-22T14:05:38.070Z    252 km W of Abepura, Indonesia   -2.194
## 17487 1987-09-27T21:22:35.470Z 196 km E of Tadine, New Caledonia  -21.503
## 15220 1992-06-15T14:16:50.060Z          west of Macquarie Island  -60.774
## 19838 1982-10-26T03:24:30.970Z     27 km NNW of La Serena, Chile  -29.683
##       Longitude Depth  Mag MagType nst gap dmin rms    net           ID
## 15241   155.377  44.6 5.50      mw  NA  NA   NA 1.0     us   usp00058jt
## 33702   143.025  15.0 5.55      mw  NA  NA   NA  NA iscgem iscgem905846
## 35716   138.398  15.0 6.01      mw  NA  NA   NA  NA iscgem iscgem910424
## 17487   169.778  33.0 5.60      ms  NA  NA   NA 1.2     us   usp00038sj
## 15220   154.040  33.0 5.70      mw  NA  NA   NA 1.3     us   usp000595e
## 19838   -71.367  71.0 5.50      mb  NA  NA   NA 1.1     us   usp0001qha
##                        Updated  X       Type horizontalError depthError
## 15241 2016-11-09T22:40:04.308Z NA earthquake              NA         NA
## 33702 2022-04-25T21:59:32.398Z NA earthquake              NA        7.8
## 35716 2022-04-25T23:24:56.825Z NA earthquake              NA       14.5
## 17487 2022-04-27T21:36:14.419Z NA earthquake              NA         NA
## 15220 2022-04-28T18:44:11.379Z NA earthquake              NA         NA
## 19838 2022-04-28T00:24:13.472Z NA earthquake              NA         NA
##       magError magNst   status locationSource magSource
## 15241       NA     NA reviewed             us       hrv
## 33702     0.38     NA reviewed         iscgem    iscgem
## 35716     0.20     NA reviewed         iscgem    iscgem
## 17487       NA      4 reviewed             us        us
## 15220       NA     NA reviewed             us       hrv
## 19838       NA     30 reviewed             us        us
```

```
summary(train)
```

```
##      Time              Place              Latitude         Longitude
##  Length:29864       Length:29864       Min.   :-77.080   Min.   :-180.00
##  Class :character   Class :character   1st Qu.:-16.524   1st Qu.: -76.07
##  Mode  :character   Mode  :character   Median :  1.192   Median :  98.16
```

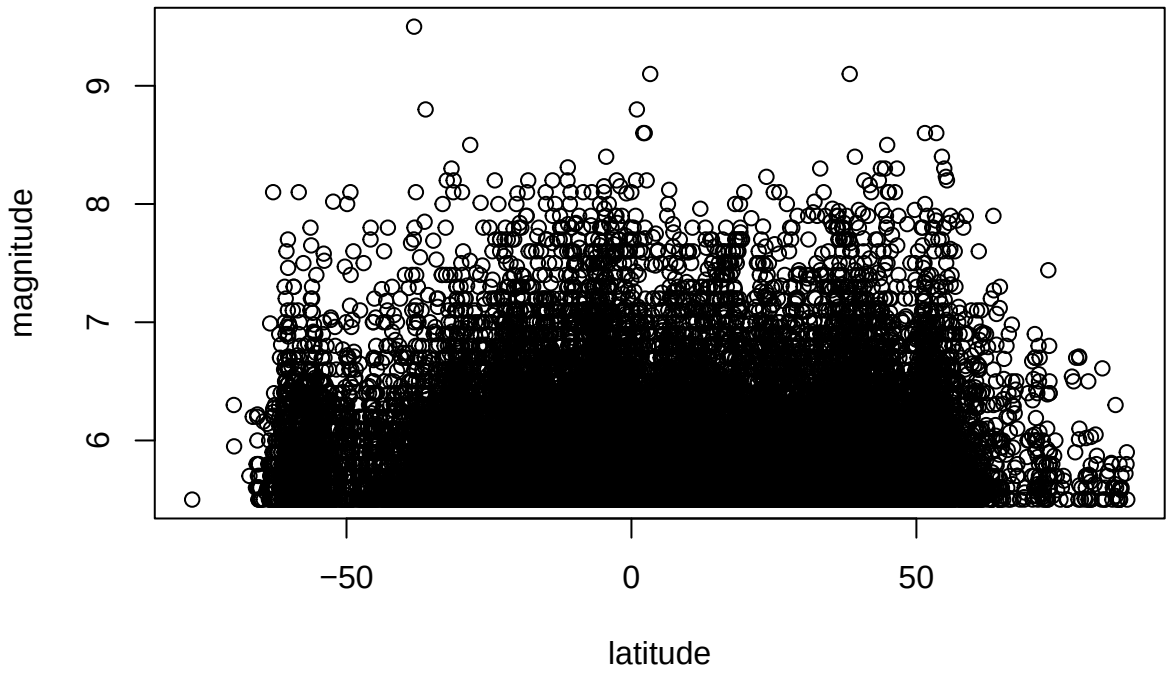
```
##                                       Mean   :  5.528   Mean   :  38.63
##                                       3rd Qu.: 33.934   3rd Qu.: 143.38
##                                       Max.   : 87.007   Max.   : 180.00
##
##      Depth             Mag           MagType               nst
##  Min.   : -4.00   Min.   :5.500   Length:29864       Min.   :  0.0
##  1st Qu.: 15.00   1st Qu.:5.600   Class :character   1st Qu.:135.0
##  Median : 28.22   Median :5.800   Mode  :character   Median :242.5
##  Mean   : 59.08   Mean   :5.948                      Mean   :265.3
##  3rd Qu.: 41.43   3rd Qu.:6.140                      3rd Qu.:373.0
##  Max.   :700.00   Max.   :9.500                      Max.   :929.0
##  NA's   :117                                         NA's   :23908
##       gap              dmin             rms             net
##  Min.   :  8.00   Min.   : 0.005   Min.   : 0.03   Length:29864
##  1st Qu.: 24.10   1st Qu.: 1.158   1st Qu.: 0.89   Class :character
##  Median : 36.00   Median : 2.528   Median : 1.00   Mode  :character
##  Mean   : 44.94   Mean   : 4.323   Mean   : 1.00
##  3rd Qu.: 54.60   3rd Qu.: 5.141   3rd Qu.: 1.11
##  Max.   :360.00   Max.   :39.730   Max.   :42.41
##  NA's   :21839    NA's   :26342    NA's   :13725
##       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.: 5.715   1st Qu.:  3.60   1st Qu.:0.200   1st Qu.: 17.00
##  Median : 7.100   Median :  6.10   Median :0.200   Median : 31.00
##  Mean   : 7.336   Mean   : 10.68   Mean   :0.262   Mean   : 47.38
##  3rd Qu.: 8.500   3rd Qu.: 16.20   3rd Qu.:0.330   3rd Qu.: 55.00
##  Max.   :99.000   Max.   :569.20   Max.   :1.840   Max.   :941.00
##  NA's   :26681    NA's   :13170    NA's   :16595   NA's   :25550
##     status          locationSource      magSource
##  Length:29864       Length:29864       Length:29864
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
mean(train$Mag)
```

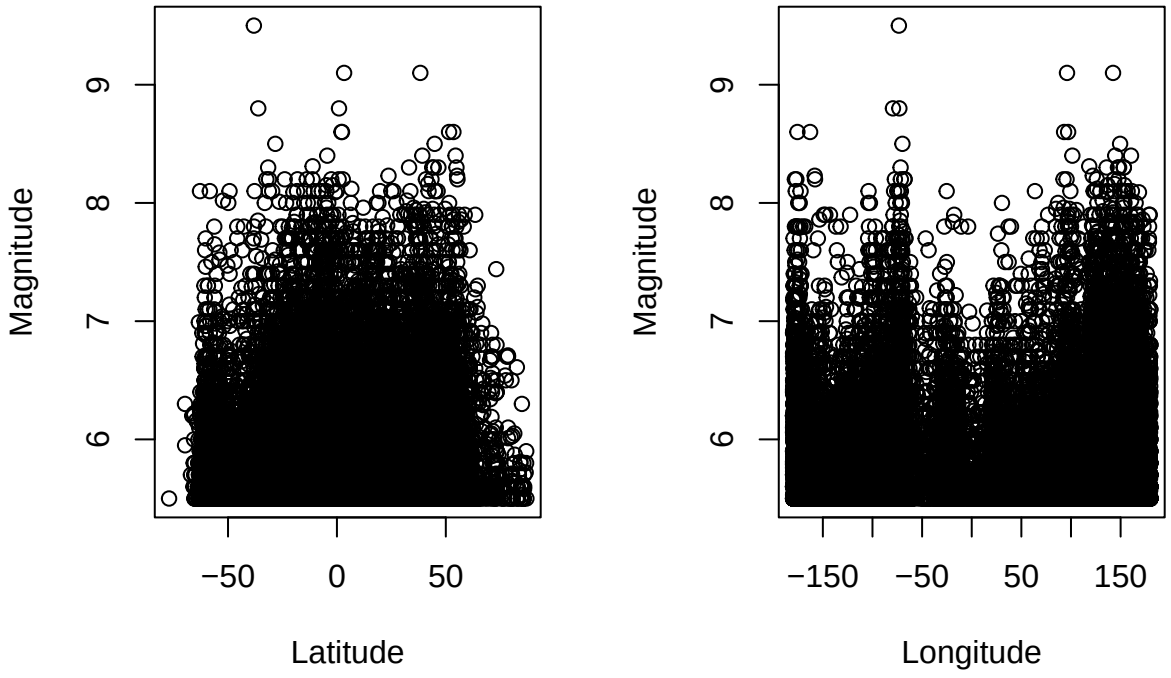
```
## [1] 5.947551
```

2c. Data Graphing

```
plot(train$Mag~train$Latitude, xlab="latitude", ylab="magnitude")
```




```
par(mfrow=c(1,2))
plot(train$Latitude, train$Mag, xlab="Latitude", ylab="Magnitude")
plot(train$Longitude, train$Mag, xlab="Longitude", ylab="Magnitude")
```



2d. Simple Linear Regression

```
simple_linear_regression_model <- lm(Mag~Latitude, data=train)
simple_linear_regression_model
```

```
##
## Call:
```

```
## lm(formula = Mag ~ Latitude, data = train)
##
## Coefficients:
## (Intercept)     Latitude
##   5.9446518    0.0005244
```

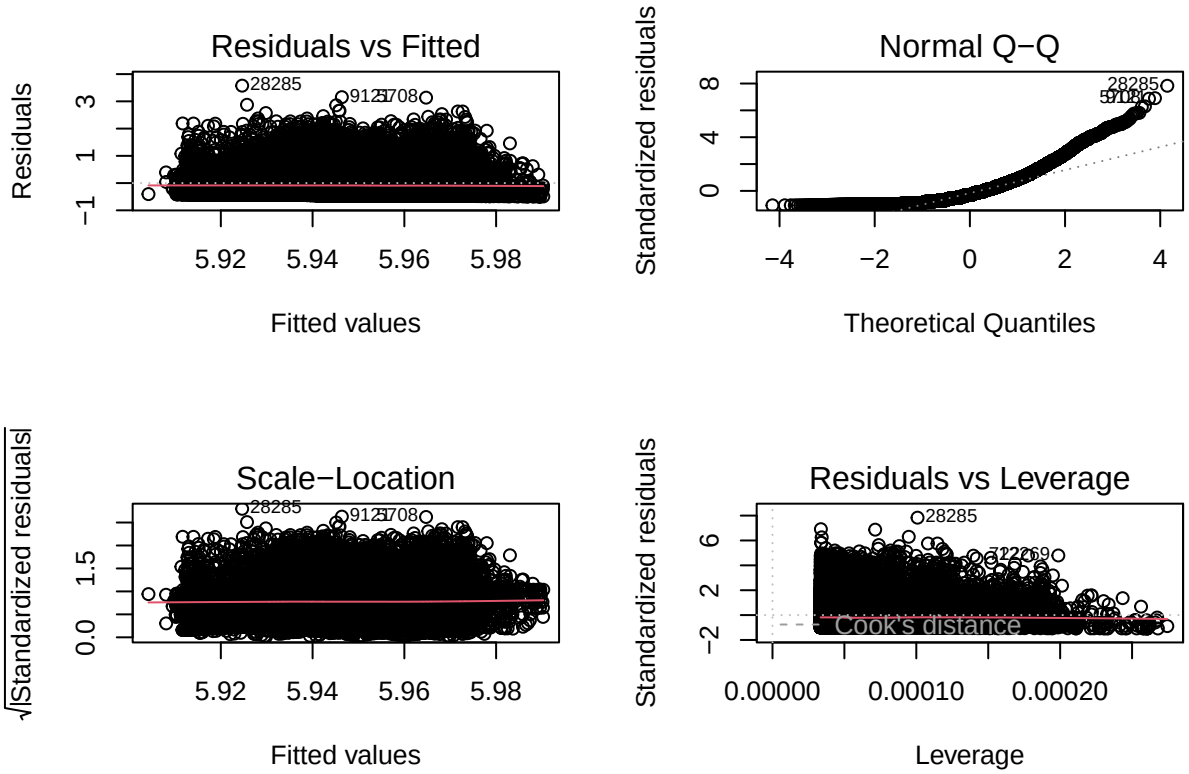
```
summary(simple_linear_regression_model)
```

```
##
## Call:
## lm(formula = Mag ~ Latitude, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4903 -0.3409 -0.1365  0.1884  3.5754
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 5.945e+00  2.684e-03 2214.926  < 2e-16 ***
## Latitude    5.244e-04  8.579e-05    6.113 9.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4565 on 29862 degrees of freedom
## Multiple R-squared:  0.00125,    Adjusted R-squared:  0.001216
## F-statistic: 37.37 on 1 and 29862 DF,  p-value: 9.913e-10
```

The line generated by the simple linear regression model is magnitude = 0*latitude + 5.945. The R-squared value is 0.00125. This is extremely low and suggests that the data is not a good fit for linear regression. The p-value of latitude is very small, which indicates that latitude is statistically significant to magnitude.

2e. Residual Plot

```
par(mfrow=c(2,2))
plot(simple_linear_regression_model)
```



The residuals vs. fitted graph does not show any non-linear patterns, which indicates that linear regression might be a good fit. The Q-Q Plot is not linear, which indicates that the residuals are not normally distributed. The scale-location plot has a horizontal line, which indicates that the residuals are spread out evenly. The Residuals vs. Leverage graph shows that the majority of the data points are not bound by Cook's distance lines, which suggests that a lot of points have leverage and are outliers.

2f. Multiple Linear Regression

```
multiple_linear_regression_model <- lm(Mag~Latitude+Depth, data=train)
multiple_linear_regression_model
```

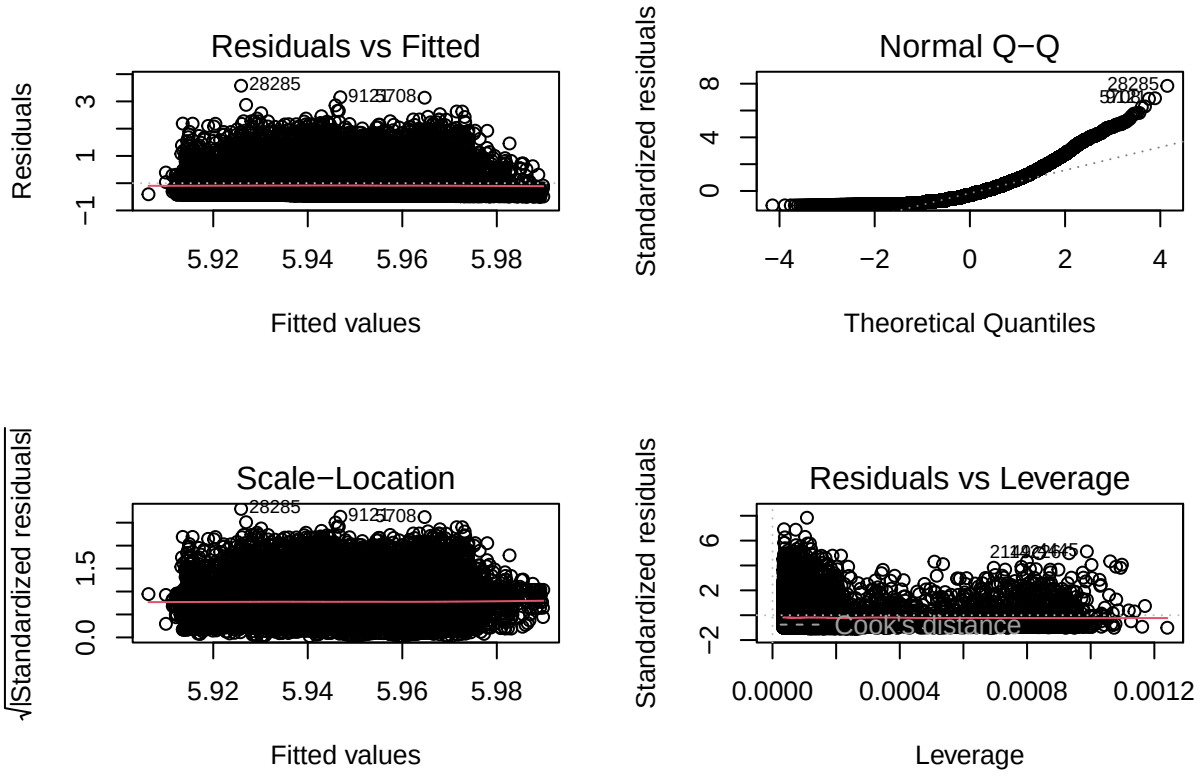
```
##
## Call:
## lm(formula = Mag ~ Latitude + Depth, data = train)
##
## Coefficients:
## (Intercept)     Latitude        Depth
##   5.9457218    0.0005089   -0.0000168
```

```
summary(multiple_linear_regression_model)
```

```
##
## Call:
## lm(formula = Mag ~ Latitude + Depth, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4898 -0.3405 -0.1364  0.1881  3.5741
##
## Coefficients:
```

```
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  5.946e+00  3.061e-03 1942.538  < 2e-16 ***
## Latitude     5.089e-04  8.650e-05    5.883 4.06e-09 ***
## Depth       -1.680e-05  2.403e-05   -0.699    0.485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.456 on 29744 degrees of freedom
##   (117 observations deleted due to missingness)
## Multiple R-squared:  0.001224,   Adjusted R-squared:  0.001157
## F-statistic: 18.23 on 2 and 29744 DF,  p-value: 1.224e-08
```

```
par(mfrow=c(2,2))
plot(multiple_linear_regression_model)
```



2g. Improved Linear Regression

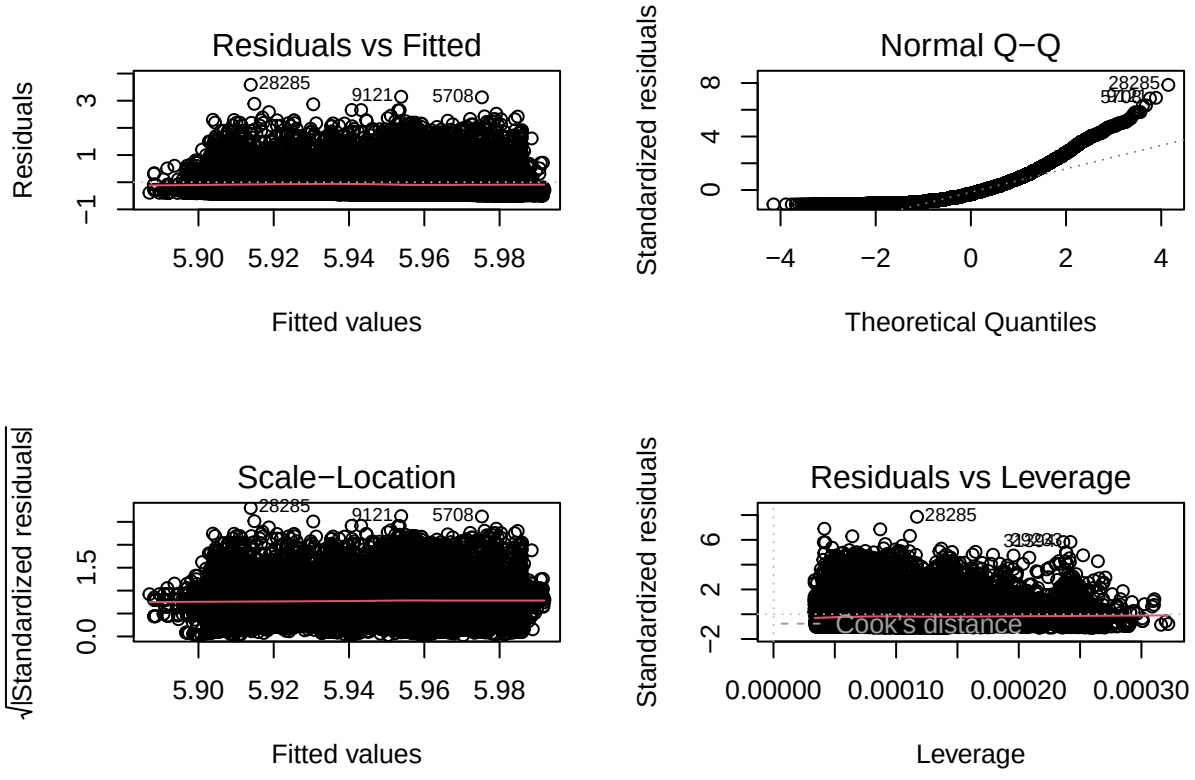
```
improved_linear_regression_model <- lm(Mag~Latitude+Longitude, data=train)
improved_linear_regression_model
```

```
##
## Call:
## lm(formula = Mag ~ Latitude + Longitude, data = train)
##
## Coefficients:
## (Intercept)     Latitude    Longitude
##   5.9401785    0.0004443    0.0001273
```

```
summary(improved_linear_regression_model)
```

```
##
## Call:
## lm(formula = Mag ~ Latitude + Longitude, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4915 -0.3437 -0.1348  0.1948  3.5861
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 5.940e+00  2.789e-03 2129.904  < 2e-16 ***
## Latitude    4.443e-04  8.683e-05    5.117 3.13e-07 ***
## Longitude   1.273e-04  2.172e-05    5.859 4.70e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4563 on 29861 degrees of freedom
## Multiple R-squared:  0.002397,   Adjusted R-squared:  0.00233
## F-statistic: 35.87 on 2 and 29861 DF,  p-value: 2.761e-16
```

```
par(mfrow=c(2,2))
plot(improved_linear_regression_model)
```



2h. Model Comparison

I think Multiple < Simple < Improved. The Multiple Linear Regression used depth which had a high p-value, which suggests it does not correlate with magnitude at all. The Improved Linear Regression had an r-squared value twice as large as the Simple Linear Regression, which suggests that the Improved Linear Regression has a greater linear relationship.

2i. Test Models

```
simple_prediction <- predict(simple_linear_regression_model, newdata=test)
simple_correlation <- cor(simple_prediction, test$Mag)
simple_mse <- mean((simple_prediction-test$Mag)^2)
simple_rmse <- sqrt(simple_mse)

print(paste('correlation:', simple_correlation))
```

Simple Linear Model

```
## [1] "correlation: 0.0478554384992787"
```

```
print(paste('mse:', simple_mse))
```

```
## [1] "mse: 0.200803546284805"
```

```
print(paste('rmse:', simple_rmse))
```

```
## [1] "rmse: 0.448111086991613"
```

```
multiple_prediction <- predict(multiple_linear_regression_model, newdata=test)
multiple_correlation <- cor(multiple_prediction, test$Mag)
multiple_mse <- mean((multiple_prediction-test$Mag)^2)
multiple_rmse <- sqrt(multiple_mse)

print(paste('correlation:', multiple_correlation))
```

Multiple Linear Model

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

```
print(paste('mse:', multiple_mse))
```

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

```
print(paste('rmse:', multiple_rmse))
```

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

```
improved_prediction <- predict(improved_linear_regression_model, newdata=test)
improved_correlation <- cor(improved_prediction, test$Mag)
improved_mse <- mean((improved_prediction-test$Mag)^2)
improved_rmse <- sqrt(improved_mse)

print(paste('correlation:', improved_correlation))
```

Improved Linear Model

```
## [1] "correlation: 0.0524918795310418"
```

```
print(paste('mse:', improved_mse))
```

```
## [1] "mse: 0.200680708906132"
```

```
print(paste('rmse:', improved_rmse))
```

```
## [1] "rmse: 0.447974004721404"
```

There is no correlation between the testing data and the multiple linear regression model. There is a 0.04 correlation between the testing data and the simple linear regression model. There is a 0.05 correlation between the testing data and the improved linear regression model. Both the improved and simple linear regression models have a mean squared error of 0.2. I think the multiple linear regression model did not have any correlation because depth had a very large p-value. I think the improved linear regression model is better than the simple linear regression model because it includes longitude as a predictor and longitude has a low p-value.