Any data collected will have some noise in it. To make the data noise free we make use of error correction techniques. These techniques detect the delays/errors in transmission and facilitate error free data. Statistical regression techniques give us a relationship between two variables, one dependent and the other independent. In other words, this regression analysis gives us an idea how a dependent variable varies when an independent variable is held fixed.

The data we collected has to be pruned and corrected. For this we have explored 3 regression techniques namely, Weighted Least Mean Squares(LMSQ), Linear Regression and K Nearest Neighbor.

**<<LMSQ>>**

**Least squares** is a standard method in [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). It is used to the approximate solution of [overdetermined systems](https://en.wikipedia.org/wiki/Overdetermined_system). They are sets of equations in which there are more equations than unknowns. Least squares indicates that the overall solution reduces the sum of the squares of the errors made in the results of every single equation. Linear least squares is a way fitting a [mathematical](https://en.wikipedia.org/wiki/Mathematical_model) or [statistical model](https://en.wikipedia.org/wiki/Statistical_model) to [data](https://en.wikipedia.org/wiki/Data). When we have idealized value provided by the model for any data point, it can be expressed linearly in terms of the unknown [parameters](https://en.wikipedia.org/wiki/Parameter) . The resulting fitted model can be used to [compile](https://en.wikipedia.org/wiki/Descriptive_statistics) the data, to [find](https://en.wikipedia.org/wiki/Prediction) the unobserved values from the same system and to understand the structure that may determine the system. This method is optimum when the sum, *S*, of squared residuals is a minimum.

S=\sum_{i=1}^{n}{r_i}^2

The most important application of least squares is [data fitting](https://en.wikipedia.org/wiki/Curve_fitting). It best fits in minimizing the sum of squared [residuals](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics), a residual which is the difference between an experimented value and the fitted value provided by a model.When the problem has substantial doubts in the [independent variable](https://en.wikipedia.org/wiki/Independent_variable), then simple regression and least squares methods have problems. In such cases, the methodology essential for fitting [errors-in-variables models](https://en.wikipedia.org/wiki/Errors-in-variables_models) may be measured.

A special case of this [generalized least squares](https://en.wikipedia.org/wiki/Generalized_least_squares) is called weighted least squares. It occurs when all the off-diagonal entries of *Ω* (the correlation matrix of the residuals) are null and the [variances](https://en.wikipedia.org/wiki/Variance) of the observations (along the covariance matrix diagonal) are unequal.

Hence, the expressions given above are based on the implicit hypothesis that the errors are not related with each other. Also we observe it with the independent variables and have equal variance. The [Gauss–Markov theorem](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem) shows that, when this is so, \hat{\boldsymbol{\beta}}is a [best linear unbiased estimator](https://en.wikipedia.org/wiki/Best_linear_unbiased_estimator).

However, the measurements are uncorrelated but have different indecisions. For this a modified approach might be adopted. [Aitken](https://en.wikipedia.org/wiki/Alexander_Aitken) showed that when a weighted sum of squared residuals is minimized, \hat{\boldsymbol{\beta}}is the [BLUE](https://en.wikipedia.org/wiki/Best_linear_unbiased_estimator) if each weight is equal to the reciprocal of the variance of the measurement.

 S = \sum_{i=1}^{n} W_{ii}{r_i}^2,\qquad W_{ii}=\frac{1}{{\sigma_i}^2} 

The gradient equations for this sum of squares are

-2\sum_i W_{ii}\frac{\partial f(x_i,\boldsymbol {\beta})}{\partial \beta_j} r_i=0,\qquad j=1,\ldots,n

In a linear least squares system give the modified normal equations,

\sum_{i=1}^{n}\sum_{k=1}^{m} X_{ij}W_{ii}X_{ik}\hat{ \beta}_k=\sum_{i=1}^{n} X_{ij}W_{ii}y_i, \qquad j=1,\ldots,m\,.

It is usually supposed that the response data has equal quality and has constant variance. If this is violated, your fit is unduly influenced by data of poor class. To improve this fit, you can use weighted least-squares regression where an additional scale factor like weight is included in the fitting process. Weighted least-squares regression minimizes the error estimate as,

i(yi-yi^)2

where *wi* are the weights. The weight determines how much each response value influences the final parameter estimates. High-quality data points influence the fit more than a low-quality data points. Therefore, weighting your data is commended if the weights are known, or if there is validation that they follow a particular format.

In weighted least mean squares, we calculate the error function at a specific location L for a given source, S, as follows:

For example we are given rtt(S; H), the top K(= 3 by default) peaks of the probability distribution of the distance for a small rtt range around rtt(S; H). All our delay measurements are Round Trip Time measurements and not one way delay. For each peak P among the K different peaks, let us compute the error to be err(P; S; L)=(d(P ) d(S; L))2=prob(P ), where d(P ) and prob(P )are referred to the distance value of the peak P and its probability density in the Probability Distribution Function. Among the K different error values, we compute the minimum value of error as the error measure of the location with respect to single source S. Therefore error(S; L) = minP (err(P; S; L)). The error measure with respect to all sources for a specific location L is the sum total of all the errors from different sources to destinations. This can be shown as, error(L) = S error(S; L). The LMSQ algorithm accounts the location L with the least value of error(L). The reason for opting the top K peaks is to take into consideration the occurrence of multiple disjoint peaks in the probability distribution function. The prob(P ) term was used to evaluate the error inversely to the probability distribution around the peak values.

Weighted least squares regression is not related with a particular type of function used to define the connection between the process variables. On the contrary, weighted least squares reflects the behavior of the casual errors in the model. Also it can be used with functions that have either [linear](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm#def) or [nonlinear](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm#def) parameters. It is done by integrating additional nonnegative constants or weights linked with each data point, into the fitting criterion. The size of the weight specifies the accuracy of the information present in the accompanying observation. Enhancing the weighted fitting criterion to find the parameter estimations permits the weights to determine the contribution of each observation to the final parameter estimations. It is significant to note that the weight for each observation is given comparative to the weights of the other observations; so different sets of absolute weights can have identical effects.

|  |  |
| --- | --- |
|  | Weighted least squares is an efficient method that makes use of small data sets. It shares the ability to provide different types of easily interpretable statistical intervals for assessment, expectation, standardization and optimization. Moreover, the main advantage that weighted least squares has over the other methods is the ability to handle regression conditions in which the data points are of fluctuating quality. If the standard deviation of the random errors in the data is not persistent across all levels of the explanatory variables, using weighted least squares with weights that are inversely proportional to the variance at each level of the explanatory variables yields the most precise parameter estimates possible. |
|  | The major disadvantage of weighted least squares, is probably the fact that the theory behind this method is depended on the postulation that the weights are known accurately. This is almost never the case in real applications, so estimated weights must be used in its place. The effect of using estimated weights is difficult to evaluate, but experience specifies that small variations in the weights due to approximation do not often disturb a regression analysis or its explanation. Nevertheless, when the weights are assessed from small numbers of replicated remarks, the results of an analysis can be very badly and randomly affected. This is likely to be the case when the weights for extreme values of the predictor are estimated using only a few observations. It is important to be aware of this potential problem and to only use weighted least squares when the weights can be estimated precisely relative to one another . | |
|  |  | |

**K Nearest Neighbor** is a method that is used for data classification and regression. Here data is classified as training set and test set.

In KNN, basic regression algorithm is that, once a new point is determined, its distance to all its neighbors is calculated, sorted and nearest k-neighbors are determined. For a new data point, mean of its k nearest neighbors will be assigned.

Below is the command in R language that implements the regression for given set of data.

**knn(train, test, k = 1, use.all = TRUE)**

Arguments:

train- a set of data classified based on a condition

test- second set of the same data, that does not classify the condition

k- number of neighbors to be considered

use.all- handles ties (in case of votes)

**Linear Regression** is one of the most popularly used statistical regression technique in practical applications. Models which depend linearly on their unknown parameters have their outputs easier to determine than the ones that are non linearly dependent. It can be fitted in many ways, but commonly, linear regression fits in using the "least squares" approach.

Let Y be the dependent variable ; let x1,x2,... xk be the independent variables which should be predicted. The equation for predicting the value can be written as:

image002.png

where **bi** is the change in the predicted value of Y per unit of change in Xi, other things being equal.  The additional constant **b0**, is the intercept, is the prediction that the model would make if all the X’s were zero.

Least squares is used to estimate coefficients and intercepts that is, we set them equal to distinct values to minimize sum of squared errors within the data in which model would be fitted. And the model's prediction errors are typically assumed to be independently and identically normally distributed.

Applying linear regression to our data,

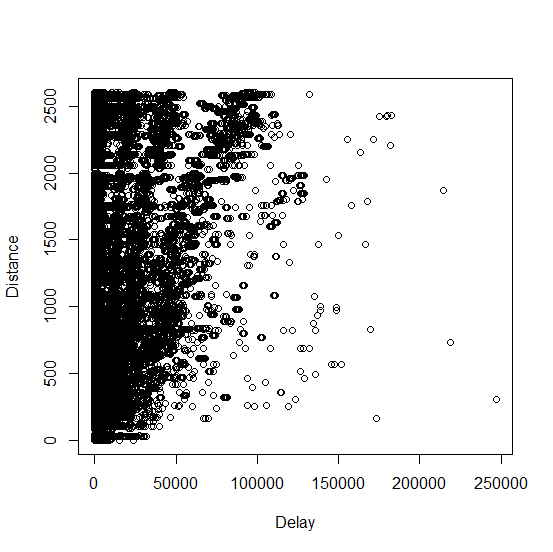
We have selected the data set to be the logset collected on 29th Oct for file size 1MB .

Step 1: Import the log file into R using the command:

f = read.csv("download1.csv")

Step 2: Plot a scattered graph with this un corrected data using the dataset "f"

plot(f)



We see that the data is scattered. The scantily spread data is the noise that has to be corrected.

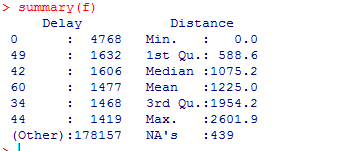
Step 3: Check for the structure of f

str(f)

Capture.PNG

Step 4: Check for the Delay and distances summary estimations

summary(f)



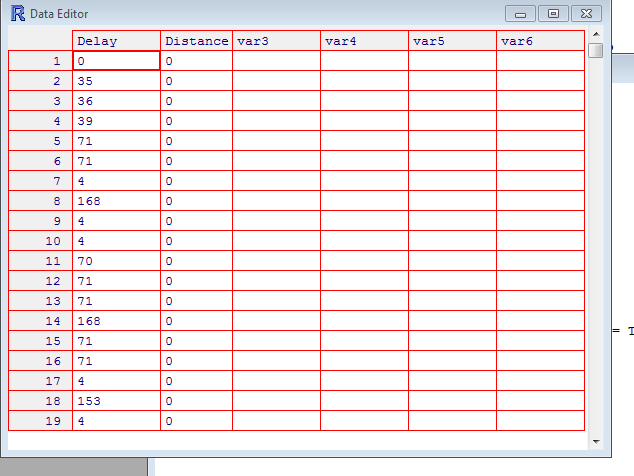
Step 5: Load required libraries for plotting linear regression

load=function() {

+ library(MASS) }

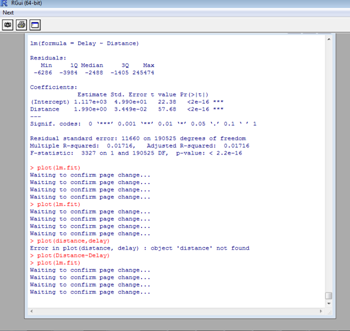
Step 7: Generate a table from which graph has to be generated for given delay to diatance

fix(f)

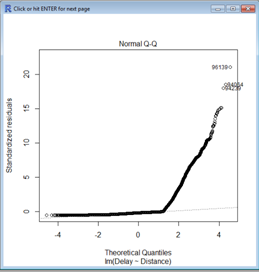


Step 6: Apply linear regression to the dataset

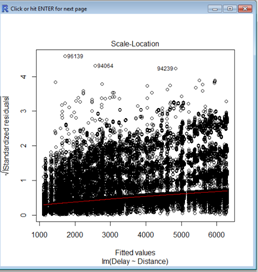
lm.fit=lm(Delay~Distance)



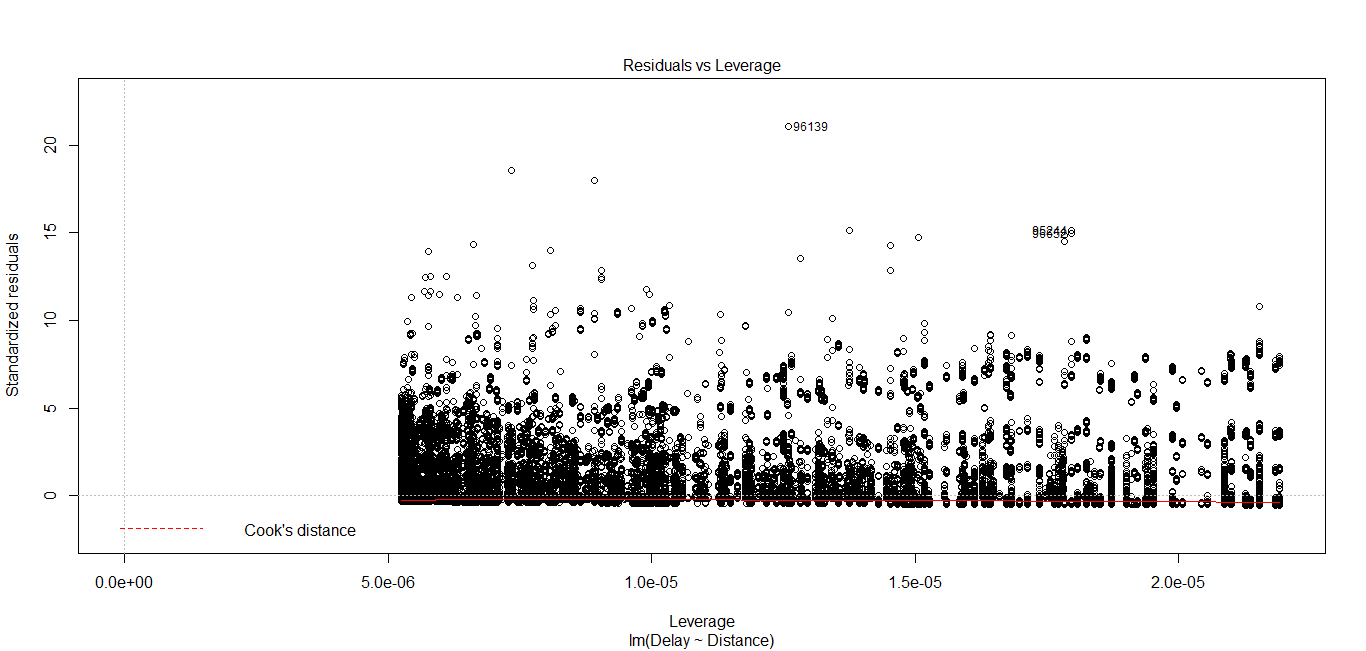
We get the following graphs upon regression:



The Q-Q plot gives us the graph based on actual data to estimated values. It gives us a straight line along which all the data is aligned.



This plot is used to know the non linearity in data. In other words, it measures the error variances. We can see that it corrects the data to be concentrated around the "red" line.



The leverage vs residuals plot gives us the cooks's distance, which estimates how much error would be reduced, it a detected point would be deleted.

**Conclusion:**

From the analysis of collected data, we can conclude that we have achieved good amount of error correction.

<<can add more>>