**Linear Regression**

The linear regression analysis consists of a statistical analysis with the objective to verify the existence of a functional relationship between a dependent variable with one or more independent variables. In other words, this analysis consists of obtaining a linear equation that try to explain the variation from the dependent variable by variation from the independent variable level.

The statistical model for this case should be:

Where, *Y* represents the dependent variable, *a* represents the line slope of the linear model, and *b* represent the y-axis intercept. The regression line is a method to calculate the parameters *a* and *b*.

The determination coefficient, frequently called R2, or simply r2, for the linear regression case, provide auxiliary information to the analysis of the result of regression variance, as a way to verify if the proposed model is suitable or not to describe the phenomenon.

In this work, we verified how close is the relationship between the error probability (independent variable) and the other variables dependent: LSS (Largest Sorted Subside) and UEQ (Unordered Elements Quantity) to a linear equation.

In order to try to establish an equation that represents the phenomenon in studying it can make a graph, called scatter plot, to verify by visualization how is the variation between the variables, dependent and independent. Figures 1 and 2 are the scatter plots for LSS and UEQ for each algorithm, respectively. At these figures, we can see the line from the linear equation from the estimated model that used the minimum mean square error as criteria. The red dotted line represents the bounds for 95% of the confidence interval.

Uma imagem contendo captura de tela, mapa

Descrição gerada automaticamenteUma imagem contendo captura de tela, mapa

Descrição gerada automaticamenteUma imagem contendo captura de tela

Descrição gerada automaticamenteUma imagem contendo captura de tela, mapa

Descrição gerada automaticamente

Figure 1 – Scatter plot and the linear regression for LSS versus Error Probability

Uma imagem contendo mapa

Descrição gerada automaticamenteUma imagem contendo mapa, texto

Descrição gerada automaticamente

Uma imagem contendo mapa

Descrição gerada automaticamenteUma imagem contendo captura de tela

Descrição gerada automaticamente

Figure 2 – Scatter plot and the linear regression for UEQ versus Error Probability

However, it can be verified that the points at scatter plots, don´t have perfect adjust to the proposed math model line. There are in major part, a great distance between the points at the plot and the model line. This happens due to the fact that in each error probability we have a lot of values of dependent variable and with large dispersion. This explains why the determination coefficient r2 has a low value in most of the cases. The table

Table 1 – Parameters from linear regression for the LSS variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Bubble** | **Quick** | **Merge** | **Insertion** |
| Intercept (b) | 36.576 | 14.86 | 20.651 | 8.5707 |
| Slope (a) | -512.32 | -176.4 | -333.64 | -90.711 |
| r2 | 0.141 | 0.0499 | 0.101 | 0.0399 |
| Root Mean Squared Error (RMSE) | 21.5 | 13.1 | 17 | 7.59 |

Table 2 – Parameters from linear regression for the UEQ variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Bubble** | **Quick** | **Merge** | **Insertion** |
| Intercept (b) | 1.4429 | 4.3258 | 4.1165 | 10.426 |
| Slope (a) | -214.56 | 268.59 | 408.4 | 295.39 |
| r2 | 0.899 | 0.823 | 0.809 | 0.826 |
| Root Mean Squared Error (RMSE) | 1.23 | 2.12 | 3.38 | 2.31 |