Model Analyzer Tool

This tool can compute both confidence intervals and prediction intervals for both simple linear regression and multiple linear regression. These intervals can then be applied to the predicted values, allowing you to establish confidence bounds and prediction bounds.

1- Simple Linear Regression

Simple Linear Regression is a statistical method used to model the relationship between a single independent variable (predictor) and a dependent variable (response). It aims to find a linear relationship between the two variables, which can be expressed as a straight line on a scatterplot.

In simple linear regression, the goal is to fit a linear equation of the form:

$$Y = b \times X + a$$

Where:

Y is the dependent variable (response).

X is the independent variable (predictor).

a is the intercept (the value of Y when X is zero).

b is the slope of the line (indicating the change in Y for a one-unit change in X).

a. Confidence Intervals

Confidence Intervals (CIs) are a statistical concept and a common tool used in inferential statistics to quantify the uncertainty or variability

associated with a sample statistic, such as a mean, proportion, regression coefficient, or any other parameter of interest. CIs provide a range of values within which you can reasonably expect the true population parameter to fall, with a specified level of confidence.

$$Y_\alpha = Y \pm t_\alpha \times SE \times \sqrt{\frac{1}{n} + \frac{(X - X_{mean})^2}{SS_{xx}}}$$

b. Prediction Intervals

Prediction Intervals (PIs) are a statistical concept used to quantify the uncertainty associated with individual predictions or future observations in the context of regression analysis or other predictive modeling techniques. Unlike Confidence Intervals (CIs), which focus on estimating population parameters, PIs are concerned with estimating the range of values within which a single future data point or observation is likely to fall.

$$Y_\alpha = Y \pm t_\alpha \times SE \times \sqrt{1.0 + \frac{1}{n} + \frac{(X - X_{mean})^2}{SS_{xx}}}$$

Y_α represents the upper and lower bounds of prediction for a given confidence level, denoted as α.

t_α is critical t-value for a confidence level of 1-α with (n-2) degrees of freedom, where 'n' represents the total number of observations in our dataset.

SE stands for the square root of MSE, with MSE representing Mean Square Error or residuals.

SS_{xx} represents the sum of squared differences from the mean in independent variable.

2- Multiple Linear Regression

Multiple Linear Regression is a statistical method used to model the relationship between a dependent variable (Y) and two or more independent variables (X1, X2, X3, etc.). It extends the principles of simple linear regression, where a single independent variable is used to predict the dependent variable, to situations in which multiple predictors are involved.

The relationship between the dependent variable and the independent variables is expressed using a linear regression equation. It can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Y is the dependent variable.

β_0 is the intercept (the value of Y when all independent variables are zero).

$\beta_1, \beta_2 \ldots$ are the coefficients (regression coefficients) representing the change in Y for a one-unit change in each respective independent variable.

a. Confidence Intervals

In multiple linear regression, Confidence Intervals (CIs) refer to the range of values within which the population parameters associated with the regression model are expected to fall with a specified level of confidence. Multiple linear regression involves predicting a dependent variable (Y) based on multiple independent variables (X1, X2, X3, etc.), and it aims to estimate the coefficients (parameters) associated with these independent variables.

$$Y_{bounded} = Y_{pred} \pm t_{(1-\alpha/2),(n-p-1)} \times SE_{mean}$$

$$SE_{mean} = \sqrt{MSE \times (1 + x_h^T \times (X^T \times X)^{-1} \times x_h)}$$

b. Prediction Intervals

In multiple linear regression, Prediction Intervals (PIs) are used to estimate the range of values within which a future observation or response variable is likely to fall. These intervals take into account the uncertainty associated with making predictions when you have multiple independent variables (predictors).

$$Y_{bounded} = Y_{pred} \pm t_{(1-\alpha/2),(n-p-1)} \times SE_{mean}$$

$$SE_{mean} = \sqrt{\frac{MSE}{n} \times (1 + x_h^T \times (X^T \times X)^{-1} \times x_h)}$$

SE_{mean} is the standard error of the mean for the expected value of the dependent variable (y).

MSE is the Mean Squared Error, which is often calculated as the average squared difference between the observed and predicted values during the training of your regression model.

n is the number of observations in your dataset.

x_h is a vector of the values of the independent variables for the new data point for which you want to make a prediction.

X represents the design matrix of the independent variables for your entire dataset.

$T_{(1-\alpha/2),(n-p-1)}$ is the critical value from the t-distribution corresponding to the desired confidence level ($\alpha/2$) and the degrees of freedom n-p-1 where n is the number of observations and p is the number of independent variables.

3- QQ Plot

To calculate the data for a QQ (Quantile-Quantile) plot, you need to compare the quantiles of your observed data to the quantiles of a theoretical distribution. The purpose of a QQ plot is to assess how well the distribution of your observed data matches the distribution of the theoretical distribution. Here's the mathematical process:

A. Empirical Quantiles (Observed Quantiles):

For each data point in your sorted data, calculate the empirical quantile. The empirical quantile is the rank of the data point divided by the total number of data points. This gives you the cumulative distribution function (CDF) of your observed data.

The formula for the empirical quantile (observed quantile) is:

$$Q_i = \frac{i - 0.5}{n}$$

where i is the rank of the data point, and n is the total number of data points.

B. Theoretical Quantiles:

Choose a theoretical distribution (e.g., the standard normal distribution) that you want to compare your data to. This distribution should have the same mean and standard deviation as your data if you're assessing normality. If you select the 'Use Standard' checkbox, the tool will not fit a custom distribution to your data; instead, it will apply a standard distribution format. For instance, in the case of a Normal distribution, the tool will utilize a standard Normal distribution with a mean of 0 and a standard deviation of 1.

In summary, the QQ plot compares the quantiles of your observed data (empirical quantiles) to the quantiles of a chosen theoretical distribution (theoretical quantiles). The plot helps you visually assess whether your data follows the theoretical distribution or if there are deviations from the expected pattern. Departures from a straight line suggest deviations from the assumed distribution.

4- Model Efficiency

1- Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yp_i|$$

2- Mean Bias Error (MBE): The Mean Bias Error (MBE) is a statistical measure of the average difference between predicted and observed values. It is calculated as the sum of the individual differences divided by the number of observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(yp_i - y_i)$$

3- Root Mean Squared Error (RMSE): The square root of the MSE, which provides an error metric in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yp_i)^2}$$

4- Normalized Root Mean Squared Error (NRMSE): The square root of the MSE, which provides an error metric in the same units as the target variable.

$$NRMSE = \frac{RMSE}{\bar{y}}$$

5- Mean Absolute Percentage Error (MAPE): Measures the percentage difference between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - yp_i}{y_i}\right|$$

6- R-squared (R2): A measure of the proportion of variance explained by the model.

$$R = \frac{n\sum y \times yp - \sum y \sum yp}{\sqrt{(n\sum y^2 - (\sum y)^2) \times (n\sum yp^2 - (\sum yp)^2)}}$$

$$\mathbf{Rsquared} = R^2$$

7- **Adjusted R-squared:** An adjusted version of R2 that accounts for the number of predictors.

$$AdjR^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

8- Nash–Sutcliffe Efficiency (NSE):

The Nash–Sutcliffe model efficiency coefficient (NSE) is a normalized statistic that determines the relative magnitude of the residual variance (model error) compared to the measured data variance.

$$NSE = 1 - \frac{\sum(y_i - yp_i)^2}{\sum(y_i - \bar{y})^2}$$

9- **Residual Sum of Squares (RSS):** The sum of the squared residuals.

$$RSS = \sum_{i=1}^{n}(y_i - yp_i)^2$$

10- **Explained Sum of Squares (ESS):** The sum of the squared differences between the predicted values and the mean of the actual values.

$$ESS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

5- Autocorrelation

Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series. Autocorrelation, as a statistical concept, is also known as serial correlation. Autocorrelation of model residuals is commonly used to detect and diagnose any remaining patterns or dependencies in the residuals that the model has

not captured. If autocorrelation is present in the residuals, it suggests that the model may not adequately account for temporal dependencies in time series data. This information can guide you in improving your model.

$$r_k = \frac{\sum_{t=k+1}^{T}[(y_t - \bar{y}) \times (y_{t-k} - \bar{y})]}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

$$Bounds = \pm \frac{Z}{\sqrt{T-k}}$$

Where T represents the number of samples, k is the lag number, and Z is the inverse cumulative distribution (CD) of the standard normal distribution, determined by the Confidence Level.

6- Model Residual Analysis

a. Standardization (Z-score normalization)

Standardization, also known as Z-score normalization, involves transforming data in such a way that it has a mean of 0 and a standard deviation of 1.

$$Z = \frac{X - \mu}{\sigma}$$

X is a data point.

μ is the mean of the data.

σ is the standard deviation of the data.

Standardization is useful when the data has different units or scales, and you want to make them comparable. It centers the data around zero and scales it based on its variability.

Reference:

Statistical Methods for Psychology in page 215 pdf and page 192 book

https://labs.la.utexas.edu/gilden/files/2016/05/Statistics-Text.pdf

https://real-statistics.com/regression/confidence-and-prediction-intervals/

https://amsi.org.au/ESA_Senior_Years/SeniorTopic4/4h/4h_2content_11.html#:~:text=Calculating%20a%20C%25%20confidence%20interval%20with%20the%20Normal%20approximation&text=%CB%89x%C2%B1zs,%2C%20we%20use%20z%3D1.64.

https://www.youtube.com/watch?v=o0UESA3UZss

https://www.stat.uchicago.edu/~yibi/teaching/stat224/L04.pdf

http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/3-7-UnivarPredict.html

https://doi.org/10.2307/2684843