Understanding linear regressions

A linear regression is a simple and easy predictive model that is used to provide the "next value" from a dependant variable of some predictors, or explanatory variables. This model is built using a linear function and using the values of the predictors.

It has so many practical applications and uses. Those could be divided in two types:

- Forecasting a dependant variable
- Quantify the level of relationship between the dependent variable and the predictors.

The algorithm used to calculate the coefficients of the predictors and the intercept of the linear function is ordinary least squares, one of the most known and analyzed fitting methods. This algorithm attempts to minimize the mean squared error in the dependant variable.

To start this analysis, BIRT Analytics requires a Domain (filter or a full table) where the Dependent variable belongs (the one we want to predict with linear function) and the Independents variables, also known as predictors or explanatory variables.

The distinct sets of data provided must be continuous to allow all the calculations required for the least squares algorithm.

Understanding results and goodness of fit

The results of the calculation are the coefficients that accompany each predictor and the intercept. Interceptor is the value of the dependent variable (the one to be predicted) when all the predictors are zero.

Each coefficient had associated some additional parameters like:

- *Standard error*
- *t-stats test*. It's associated with the Student-t distribution, and when it has larger values implies that coefficient is not zero.
- *p-value test* . The p-value shows the results of the hypothesis test as a significance level. In that case smaller values than 0.5 are taken as evidence that the coefficient is nonzero.
- *Upper and Lower Confidence Level* (95%). Assuming that the error in the prediction of the dependant variable is normally distributed, BA 4.4 is able to calculate the confidence interval of the linear regression. The results are two linear functions with coefficients and intercepts for the 95% upper confidence level and the 95% lower confidence level, also known as the 95% confidence bands.

For the global results of the linear regression, the statistics that measure the goodness of fit are:

- *R Squared*. Also known as coefficient of determination indicates how well data provided fits with the linear regression model that has been calculated. This coefficient

ranges from 0 to 1. Values close to 1 indicates a good fitting, and values close to 0 points to no linear relationship of the data.

- *Adjusted R Squared*: Sometimes, R Squared suffers a increase in its value due to the addition of extra predictors, although the fit is not better. This parameter doesn't grow if a new independent variable is added to the equation and its addition doesn't affect to the goodness of fit in a positive way. Always Adjusted R Squared would be less than or equal to R Squared.

Also, the tool shows the total records used from the total selected in the Domain. The invalid records come from those that have null values in some of the variables.

How to create a linear regression

1. In Analytics-Advanced, choose Linear Regression.
2. Drag and drop the segment to analyze in the Domain.
3. Drag and drop the column to be predicted in the Dependent Variable.
4. In the left panel of Domain columns, expand the database and the appropriate tables.
5. Drag the appropriate columns from the left panel and drop them in the right panel. The columns specify the continuous independent variables which will be the predictors of the dependent variable.
6. Choose Train. In the results tab is shown a linear equation that giving any value to the predictors, it's possible to predict the dependent variable. Also appears a stars ratio to measure the goodness of fit of the model. In the main panel, there is a visualization of the function and a sample of the values used in the calculations.
7. For more advanced results, in the statistics tab is possible to analyze each calculated coefficient of the equation and their goodness of fit and relevance in the model.
8. Once saved this analysis, it's possible to apply the model to predict values of the dependant variable using the equation calculated.