Understanding logistic regressions

A logistic regression is a probabilistic and predictive model that provides a probability of occurrence of a binary response (dependent variable) based in the values of one or more predictor variables (independent variables). This model is built using a logistic function whose coefficients are calculated from the variety of occurrences of the binary response with the predictor variables.

It has so many practical applications and uses. For example, the Trauma and Injury Severity Score (TRISS) determines the probability of survival of a patient that enters into hospital emergency wards, and it uses a logistic regression for its calculation.

The algorithm used to calculate the coefficients of the predictors and the intercept of the logistic function is maximum likelihood estimation. This method could not be calculated directly and needs an iterative process based in Newton's method. It starts with an initial solution and in each iteration improves the result. The process ends when improvement is minute, when it had converged.

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 logistic function) and the Independents variables, also known as predictors or explanatory variables.

The dependant variable must be a binary response (only had 0 or 1 values). The predictors provided could be continuous or categorical (with a binary response). These requirements are mandatory in order to allow all the calculations done in the maximum likelihood estimation algorithm (through Newton's method).

Understanding results and goodness of fit

The results of the calculation are the coefficients that accompany each predictor and the intercept.

Each calculated coefficient had associated some additional parameters in order to measure the contribution of each predictor to the model. Those tests are:

- *Standard error*
- *Odds ratio.* This ratio quantifies how strongly the presence or absence of certain property is associated with the presence or absence of another property in a given domain. As bigger is the ratio, better is the relationship between dependent variable and the independent related to the coefficient.
- *Odds Upper and Lower Confidence Level* (95%). It has the same calculation as confidence level for a domain mean, but it's calculated on the natural log scale. It gives two functions to define the confidence interval or band.
- *Log likelihood p value.* 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.

- *Significance Level.* Based in the distinct range of significance values of p-value it's possible classify the level of significance of this coefficient. It's a range from 0 to 5, where 0 is none significance level and 5 is a highly relevant significance level.

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

- *Chi Squared test*. Also known as the likelihood ratio test, it's an asymptotically distributed Chi Squared test with certain degrees of freedom. As bigger is its value, better is the goodness of fit of the model.
- *Chi Squared p*-value. Is the statistical significance testing from the Chi Squared test. The p-value is the probability of obtaining the observed sample results (or a more extreme result) when the null hypothesis is actually true. So, when p-value is very small (less than a certain threshold), it tells that the modeled data is inconsistent with the assumption that the null hypotheses is true. In other words, this hypothesis could be rejected, so the modeled data could be accepted as true.
- *Log likelihood.* It's the logarithm of the likelihood ratio. This will always be negative, with higher values (closer to zero) indicating a better fitting model.

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 logistic 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 logistic equation that giving any value to the predictors, it's possible to predict the dependent variable (given as a probability of occurrence). 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.
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.