Formula Guide

Weight of Evidence Module

The purpose of the Weight of Evidence (WoE) module is to provide flexible tools to recode the values in continuous and categorical predictor variables into discrete categories automatically, and to assign to each category a unique WoE value. This recoding is conducted in a manner that will produce the largest differences between the recoded groups with respect to the WoE values. In addition, other constraints are observed while the program determines solutions for the optimal "binning" of predictors.

Optimal Coding of Predictors

Specifically, the goal of the algorithms implemented in the automated WoE module is to identify the best groupings for predictor variables that will result in the greatest differences in WoE between groups. For continuous variables the automated WoE module identifies the best recoding to weight-of-evidence values. For categorical predictors or interactions between coded predictors, users can combine groups with similar observed WoE to create new coded predictors with continuous weight-of-evidence value.

Continuous Variables

For continuous predictors, first a default coding is derived using the Classification and Regression Trees (C&RT) algorithm. For default categories with fewer than 20 groups *STATISTICA* will explicitly search through all possible combinations of default groups to achieve the least numbers of groups with the greatest Information Value (IV). When the number of groups is greater than 20, *STATISTICA* uses the CHAID approach. The CHAID approach is a modification to the CHAID algorithm where instead of the customary X^2 criterion, the change in WoE is used as the criterion.

Three types of constrained WoE recoding solutions are provided subject to their existence:

- Monotone solutions, where the WoE values of all adjacent recoded groups (intervals) will either increase (positive monotone relationship of predictor intervals to WoE), or the WoE values of all adjacent recoded groups will always decrease (negative monotone relationship of predictor intervals to WoE).
- Quadratic solutions, where the relationship between the coded value ranges (intervals) to WoE can have a single reversal so that the resulting function is either U-shaped or inverse-U-shaped.
- Cubic solutions, where the relationship between the coded value ranges (intervals) to WoE values can have two reversals so that the resulting function is S-shaped.

Two types of unconstrained WoE recoding solutions are provided:

- Custom coding is based on the default binning scheme with either C&RT or 10 equal groups of approximately equal size.
- The no restrictions coding is based on the custom solution after the running either the exhaustive search or the CHAID algorithm.

Note that the initial bins maybe adjusted prior to the algorithm in order to make sure that each bin satisfies the minimum N and minimum Bad N user specified parameters.

Categorical Variables

For categorical (discrete) predictors, the default (original) grouping is further refined using the modified CHAID approach.

Two types of unconstrained WoE recoding solutions are provided:

- Custom coding is based on the default binning of the group.
- The no restrictions coding is based on the default categorization provided by the modified CHAID algorithm.

Note that the initial bins maybe adjusted prior to the algorithm in order to make sure that each bin satisfies the minimum N and minimum Bad N user specified parameters.

Interactions

For pairs of coded predictors the modified CHAID approach is implemented using interaction coding of the two-way interaction table or user-defined coding.

Statistics

Chi-square

$$X^2 = \sum_{i=1}^{K} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

- This statistic is distributed according to a chi-square distribution with degrees of freedom equal to the difference between the number of parameters under the alternative hypothesis and the number of parameters under the null hypothesis.

Cramer's V

$$V = \sqrt{\frac{X^2/N}{min_{(i-1)(j-1)}}}$$

Notation:

N = Total number of observations

$min_{(i-1)(j-1)}$ = Minimum of row dimension minus 1 and column dimension minus 1

F-test

$$F = \frac{\sum_i n_i \left(\bar{Y}_{i.} - \bar{Y}\right)^2 / K - 1}{\sum_{ij} \left(Y_{ij} - \bar{Y}_{i.}\right)^2 / N - K}$$

Notation:

$\bar{Y}_{i.}$ = sample mean of the i[th] group

n_i= number of observations in the i[th] group

$\bar{Y}$ = overall mean of the data

K = denotes the number of groups

Y_{ij} = j[th] observation in the i[th] out of K groups

N= overall sample size

Gini

$$g = 2\left(\frac{Number\ of\ Bads}{N}\right)\left(\frac{Number\ of\ Goods}{N}\right)$$

Notation:

N = Total number of observations

Information Value (IV)

$$IV = \sum_{i-1}^{K}\left[\left(Relative\ Frequency\ of\ Goods_i - Relative\ Frequency\ of\ Bads_i\right) * ln\left(\frac{Relative\ Frequency\ of\ Goods}{Relative\ Frequency\ of\ Bads}\right)\right]$$

- The IV of a predictor is related to the sum of the (absolute) values for WoE over all groups. Thus, it expresses the amount of diagnostic information of a predictor variable for separating the Goods from the Bads.

Kolmogorov-Smirnov (KS) test

- For all Good observations, predicted probability of Bad is computed, that is the relative frequency of bad cases in the bin a Good observation is placed. This process is repeated for all Bad observations. The KS test is then completed with the Good/Bad indicator as the group variable and the predicted probability of Bad as the response.

$$Z = max_j{}^{|D_j|}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}j$$

Significance level (p) approximation is based on the formula:

$$p = 2\sum_{i=1}^{\infty}(-1)^{i-1}\, e^{-2i^2\left(KS\sqrt{\frac{n_1}{n_1+n_2}+0.12+0.11}\Big/\sqrt{\frac{n_1}{n_1+n_2}}\right)^2}$$

Logit Transformation (Logg Odds)

$$Logit = ln\left(\frac{\frac{Number\ of\ Goods}{N}}{\frac{Number\ of\ Bads}{N}},\right)$$

Mean

$$\bar{x} = \frac{\sum x}{n}$$

Somer's D

If ties are present:

$$d = \frac{(n_c - n_d)}{t}$$

If ties are not present:

$d = 2c - 1$ where $c = \left(n_c + 0.5(t - n_c - n_d)\right)/t$

Notation:

(Note: Sorting of cases for calculation of Somer's d is based on the relative frequency of bad, that is, estimated probably of bad.)

t = total number of pairs with different responses of good/bad

n_c = number of pairs of cases where the case with the lower ordered response value has a lower predicted mean score than the case with the higher ordered response value.

n_d = number of pairs of cases where the case with the lower ordered response value has a higher predicted mean score than the case with the higher ordered response value.

Weight of Evidence (WoE)

$$WoE = \left[ln\left(\frac{Relative\ Frequency\ of\ Goods}{Relative\ Frequency\ of\ Bads}\right)\right] * 100$$

- The value of WoE will be 0 if the odds of Relative Frequency of Goods / Relative Frequency Bads is equal to 1. If the Relative Frequency of Bads in a group is greater than the Relative Frequency of Goods, the odds ratio will be less than 1 and the WoE will be a negative number; if the Relative Frequency of Goods is greater than the Relative Frequency of Bads in a group, the WoE value will be a positive number.

Notes

The WoE recoding of predictors is particularly well suited for subsequent modeling using Logistic Regression. Specifically, logistic regression will fit a linear regression equation of predictors (or WoE-coded continuous predictors) to predict the logit-transformed binary Goods/Bads dependent or Y variable. Therefore, by using WoE-coded predictors in logistic regression, the predictors are all prepared and coded to the same WoE scale, and the parameters in the linear logistic regression equation can be directly compared, for example, when using the new modeling tools for Marginal Stepwise Logistic Regression.