Formula Guide

Stepwise Model Builder

The Stepwise Model Builder will compute the marginal predictor statistics given a current model. Specifically, the variables listed in the Marginal results table will be entered one at a time into a logistic regression model containing the predictors listed in the Model results table. This enables the analyst to evaluate the unique contribution of each predictor candidate not in the equation. The model will be estimated after recoding all Bad code values to 1 and Good code values to 0.

Notation

The following notation is used throughout this model section of this document:

n	Number of observed cases
p	Number of parameters
y	*n* x 1 vector with y_i being the observed value of the *i*th case of the chosen dichotomous good/bad dependent variable
X	*n x p* matrix with x_{ij} being the observed value of the *i*th case of the *j*th parameter
β	*p x 1* vector with β_j being the coefficient for the *j*th parameter
w	*n* x 1 vector with w_i being the weight for the *i*th case.
l	Likelihood function
L	Log likelihood function
I	Information matrix

Model

The Stepwise Model Builder uses the linear logistic model. This model has a dichotomous dependent variable, and in this module the outcome is labeled as either having a Good or Bad outcome. Included in the Stepwise Model Builder dialog are options for Dependent Variable (Y), Good code and Bad code. For the model, the dependent variable is assumed to have a probability π, where for the *i*th case:

$$\pi_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

or

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \boldsymbol{X_i}'\boldsymbol{\beta}$$

For *n* observations, y_1 through y_n, with probabilities π_i through π_n and case weights w_1 through w_n, the likelihood function is:

$$l = \prod_{i=1}^{n} \pi_i^{w_i y_i} (1-\pi_i)^{w_i(1-y_i)}$$

The logarithm of *l* is:

$$L = \ln(l) = \sum_{i=1}^{n} (w_i y_i ln(\pi_i) + w_i(1-y_i)\ln(1-\pi_i))$$

The derivative of *L* with respect to β_j is:

$$L^*_{X_j} = \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \boldsymbol{w}_i (y_i - \pi_i) x_{ij}$$

Estimation

Maximum Likelihood Estimation is achieved through the Fisher Scoring algorithm.

$$\beta_{k+1} = \beta_k + I^{-1}S$$

$$Ipxp = -E_\beta \left[\frac{\partial^2 l(\beta)}{\partial \beta^2} \right]$$

$$S = \frac{\partial l(\beta)}{\partial \beta}$$

Statistics

Estimated Variance Covariance Matrix

The estimated covariance matrix is the inverse of the information matrix (negative of the expected Hessian) evaluated at the MLE values of the parameters.

$$I(\theta_{MLE})^{-1}$$

$$I(\theta_{MLE}) = -E_\theta[H(\theta)]$$

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2}$$

Estimated Correlation Matrix

The estimated correlation matrix is the standardized version of the covariance matrix, that is, all entries are divided by the product of the standard deviations.

Gini Coefficient

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

Notation:

N = Total number of observations

Hosmer-Lemeshow (HL) Goodness of Fit Statistic

$$H = \sum_{g=1}^{n} \frac{\left(O_g - E_g\right)^2}{N_g \pi_g \left(1 - \pi_g\right)}$$

Notation: O_g = Observed event *g*

E_g = Expected event *g*

N_g = Observations of event *g*

π_g = Predicted risk for the g^{th} risk decile group

n = number of groups

Note: The Hosmer-Lemeshow statistic is asymptotically distributed and follows a X^2 distribution with *n*-2 degrees of freedom.

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:

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

Lift Value

$$Lift = \frac{Result\ Predicted\ by\ Model}{Result\ Predicted\ with\ No\ Model}$$

ROC – Area Under Curve (AUC)

$$AUC = \frac{G+1}{2}$$

Notation: *G* = Gini coefficient

ROC – Sensitivity

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

Note: If a test has two outcomes, positive and negative:

- True Positive Both the observed and predicted response is positive.
- False Positive Predicted response is positive but the observed response is negative.
- True Negative Both the observed and predicted response is negative.
- False Negative Predicted response is negative but the observed response is positive.

ROC – Specificity

$$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$$

Somers' 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 Somers' 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.

Wald Statistic

For continuous variables:

$$W_i = \left(\frac{\beta_i}{SE_{\beta_i}}\right)^2$$

For categorical variables:

If β_i is a vector of MLEs associated with m-1 dummy coded variables, and **C** is the asymptotic covariance matrix for β_i, the Wald statistic is calculated as:

$$W_i = \beta'_i \boldsymbol{C^{-1}} \beta_i$$

Note: Asymptotically distributed as a X^2 distribution with degrees of freedom equal to the number of parameters estimated and is analogous to the t-test in linear regression.

Wald Statistic - Standard Error

The standard error (SE) is the square root of the i[th] diagonal entry of the inverse information matrix.

Wald Statistic Confidence Interval

$$100\ (1 - \alpha)\%\ CI\ for\ \beta_i = \hat{\beta}_i \pm z_{1-\alpha/2} SE_{\beta_i}$$

Notation: $z_{1-\alpha/2}$= $100\ (1 - \alpha/2)th$ percentile of the standard normal distribution

$\hat{\beta}_i$ = Estimate of parameter β_i

SE_{β_i} = Standard error estimate of $\hat{\beta}_i$