Variable Selection

Like ordinary linear regression, logistic regression provides a coefficient **b** estimates, which measures

Contents

0.1 Variable Selection

Like ordinary regression, logistic regression provides a coefficient **b** estimates, which measures ¿¿¿¿¿¿ b9e1cf334ecff39f3286d124903162fb4867a913:LogisticRegression/LR08variableselection.tex each IV's partial contribution to variations in the response variables. The goal is to correctly predict the category of outcome for individual cases using the most parsimonious model.
To accomplish this goal, a model (i.e. an equation) is created that includes all predictor variables that are useful in predicting the response variable. Variables can, if necessary, be entered into the model in the order specified by the researcher in a stepwise fashion like regression.

There are two main uses of logistic regression:

- The first is the prediction of group membership. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an **odds ratio**.
- Logistic regression also provides knowledge of the relationships and strengths among the variables (e.g. playing golf with the boss puts you at a higher probability for job promotion than undertaking five hours unpaid overtime each week).

Stepwise Logistic Selection

- Stepwise logistic regression involves the stepwise (or one-by-one) selection of variables, providing a fast and effective method to screen a large number of variables, and to fit multiple logistic regression equations simultaneously.
- In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.
- Stepwise binary logistic regression is very similar to stepwise multiple regression in terms of its advantages and disadvantages. Stepwise logistic regression is designed to find the ***most parsimonious*** set of predictors that are most effective in predicting the dependent variable.

SPSS Implementation

SPSS provides a table of variables included in the analysis and a table of variables excluded from the analysis. It is possible that none of the variables will be included. It is possible that all of the variables will be included.

The order of entry of the variables can be used as a measure of relative importance.

Once a variable is included, its interpretation in stepwise logistic regression is the same as it would be using other methods for including variables.

Advantages and Disadvantages

- Stepwise logistic regression can be used when the goal is to produce a predictive model that is parsimonious and accurate because it excludes variables that do not contribute to explaining differences in the dependent variable.
- Stepwise logistic regression is less useful for testing hypotheses about statistical relationships. Its usage is recommended only for exploratory purposes, rather that as a formal procedure.
- Stepwise logistic regression can be useful in finding relationships that have not been tested before. Its findings invite one to speculate on why an unusual relationship makes sense.
- It is not legitimate to do a stepwise logistic regression and present the results as though one were testing a hypothesis that included the variables found to be significant in the stepwise logistic regression.
- Using statistical criteria to determine relationships is vulnerable to over-fitting the data set used to develop the model at the expense of generalisability.

  Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained."

Forward Selection

You can estimate models using block entry of variables or any of the following stepwise methods: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald.

Forward selection is the usual option for a stepwise regression, starting with the constant-only model and adding variables one at a time. The forward stepwise logistic regression method utilizes the likelihood ratio test which tests the change in 2LL between steps to determine automatically which variables to add or drop from the model.

Cross Validation of Stepwise Regression

When stepwise logistic regression is used, some form of validation analysis is a necessity. We will use 75/25% cross-validation.

To do cross validation, we randomly split the data set into a 75% training sample and a 25% validation sample. We will use the training sample to develop the model, and we test its effectiveness on the validation sample to test the applicability of the model to cases not used to develop it.

In order to be successful, the follow two questions must be answers affirmatively: Did the stepwise logistic regression of the training sample produce the same subset of predictors produced by the regression model of the full data set?

If yes, compare the classification accuracy rate for the 25% validation sample to the classification accuracy rate for the 75% training sample. If the **shrinkage** (accuracy for the 75% training sample - accuracy for the 25% validation sample) is 2% (0.02) or less, we conclude that validation was successful.

Note: shrinkage may be a negative value, indicating that the accuracy rate for the validation sample is larger than the accuracy rate for the training sample. Negative shrinkage (increase in accuracy) is evidence of a successful validation analysis.

If the validation is successful, we base our interpretation on the model that included all cases.

Stepwise Logistic Selection

Stepwise logistic regression involves the stepwise (or one-by-one) selection of variables, providing a fast and effective method to screen a large number of variables, and to fit multiple logistic regression equations simultaneously.

In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.

Stepwise binary logistic regression is very similar to stepwise multiple regression in terms of its advantages and disadvantages. Stepwise logistic regression is designed to find the ***most parsimonious*** set of predictors that are most effective in predicting the dependent variable.

0.2 Procedure for Stepwise Selection

- Variables are added to the logistic regression equation one at a time, using the statistical criterion of reducing the ***-2 Log Likelihood error*** for the included variables. (Recall: The lower the $-2LL$ value, the better the fit of the model).

- After each variable is entered, each of the included variables are tested to see if the model would be better off the variable were excluded. This does not happen often,but not impossible.

- The process of adding more variables stops when all of the available variables have been included or when it is not possible to make a statistically significant reduction in -2 Log Likelihood using any of the variables not yet included.

- Categorical variables are added to the logistic regression as a group. It is possible, and often likely, that not all of the individual dummy-coded variables will have a statistically significant individual relationship with the dependent variable.

0.3 SPSS Implementation

SPSS provides a table of variables included in the analysis and a table of variables excluded from the analysis. It is possible that none of the variables will be included. It is possible that all of the variables will be included. The order of entry of the variables can be used as a measure of relative importance. Once a variable is included, its interpretation in stepwise logistic regression is the same as it would be using other methods for including variables.

Model Selection Methods

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

1 **Enter**. A procedure for variable selection in which all variables in a block are entered in a single step.

2 **Forward Selection (Conditional)**. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates.

3 **Forward Selection (Likelihood Ratio)**. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates.
(LR stands for Likelihood Ratio and is considered the criterion least prone to error.)

4 **Forward Selection (Wald)**. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of the Wald statistic.

5 **Backward Elimination (Conditional)**. Backward stepwise selection.
Removal testing is based on the probability of the likelihood-ratio statistic based on conditional parameter estimates.

6 **Backward Elimination (Likelihood Ratio)**. Backward stepwise selection.
Removal testing is based on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates.

7 **Backward Elimination (Wald)**. Backward stepwise selection.
Removal testing is based on the probability of the Wald statistic.

Forward Selection

You can estimate models using block entry of variables or any of the following stepwise methods:

1. Forward conditional,
2. Forward LR,
3. Forward Wald,
4. Backward conditional,
5. Backward LR,
6. Backward Wald.