GETTING STARTED WITH PROC LOGISTIC

Andrew H. Karp
Sierra Information Services, Inc.

Introduction

Logistic Regression is an increasingly popular analytic tool. Used to predict the probability that the 'event of interest' will occur as a linear function of one (or more) continuous and/or dichotomous independent variables, this technique is implemented in the SAS® System in PROC LOGISTIC. This paper gives an overview of how some common forms of logistic regression models can be implemented using PROC LOGISTIC as well as important changes and enhancements to the procedure in Releases 6.07 and above of the SAS® System.

Background

Logistic regression is commonly used to obtain predicted probabilities that a unit of the population under analysis will acquire the event of interest as a linear function of one or more:

- continuous-level variables
- dichotomous (binary) variables
- or, a combination of both continuous and binary independent variables.

Many concepts in logistic regression will be familiar to people who already have experience with simple and multiple regression models. In fact, much of the syntax in PROC LOGISTIC will be familiar to SAS System users already experienced with using PROCs REG and/or GLM.

In logistic regression, however, the dependent variable is *dichotomous* and is usually coded as:

- zero (event did not occur)
- one (event did occur)

for each particular subject in the data set upon which the analysis will be carried out.

The *logistic function* is used to estimate, as a function of unit changes in the independent variable, the probability that the event of interest will occur. This function is often called the *link function* in that it connects, or 'links' changes in values of the independent variables to increasing (or decreasing) probability of occurrence of the event being modeled by the dependent variable.

Implementation in the SAS System

Techniques for implementing logistic regression are found in PROC LOGISTIC in the STAT module of SAS System software, and is one of several procedures in this module which can be used for categorical data analysis. Other procedures for categorical data analysis in the STAT module include:

- FREQ
- GENMOD
- CATMOD
- PROBIT
- PHREG
- LIFETEST
- PROBIT

Data Preparation

As with other forms of data analysis, the results of a logistic regression analysis performed by PROC LOGISTIC can be seriously compromised if the analyst does not take care to prepare their data properly. Following important rules regarding construction and coding of the dependent variable are critical to the SAS System's generation of accurate results.

Dependent Variable

Your dependent variable should be:

- dichotomous
- coded zero for 'non-event'
- coded one for 'event'

Although *polytomous (or multinomial logistic) regression models* (those with three or more levels, or categories, of the dependent variable) can be

implemented in the SAS System, discussion of these types of models, and how they are implemented in the SAS System, is beyond the scope of this paper and will not be considered here. SAS System implementation of these types of models is discussed in the following SAS Institute publications:

- *Logistic Regression Examples Using the SAS System (1995)*
- *SAS Technical Report R-109: Conjoint Analysis Examples (1993)*

Effect of Coding Dependent Variable on how PROC LOGISTIC Works

The zero/one coding scheme is the most commonly employed method by which events/non-events are classified for the purposes of conducting a logistic regression analysis. By default, however, PROC LOGISTIC will attempt to model (that is, predict the probability of) the lower of the two values of the dependent variable. which is usually not the desired result.

For example, if a researcher were attempting to determine the probability that a patient will die, the variable representing "outcome" (i.e., "dead" or "alive") might be coded zero for patients who survived and one for patients who died. Since zero "sorts lower" than one, PROC LOGISTIC will attempt to 'model' (that is, predict) the probability that the patient will be coded zero (lived) rather than the probability that the patient will be coded one (died). This is most likely the *opposite* of what the researcher desired.

Overriding SAS System Defaults

Users can override the default attempt by PROC LOGISTIC to predict the probability of the non-event using one of three approaches:
- re-coding the dependent variable (e.g., 0 = 'died', 1 = 'lived')
- creating and applying a FORMAT to the dependent variable where the *formatted value* of the 'event' group 'sorts higher' than the 'non-event' group
  (i.e., the external representation of 0 = 'Alive' and 1 = 'Dead')

- use the DESCENDING option in the PROC LOGISTIC statement.

This DESCENDING option, new in Release 6.07, is probably the easiest and most straightforward method by which to override the SAS System default, as it avoids potentially unnecessary work in a DATA Step before applying PROC LOGISTIC.

Implementing a Logistic Regression Analysis

The structure and syntax of many features in PROC LOGISTIC are similar to those used in PROCs REG and GLM, which facilitates comparison of how to perform a logistic regression analysis with linear models such as regression and analysis of variance.

The important difference, for our purposes, between what is being estimated by a logistic regression model and that estimated by a linear model is:

- *linear regression* attempts to predict the *value* of the dependent variable as a linear function of one (or more) independent variables
- *logistic regression* attempts to predict the *probability* that a unit under analysis will acquire the event of interest as a function of one or more independent variables.

Put another way, the logistic regression equation predicts the probability that the unit under analysis will, as a function of one or more independent variables, obtain the condition of interest which is (usually) coded as 1 in a zero/one coding scheme.

The general form of PROC LOGISTIC is:

PROC LOGISTIC DATA=*dsn* [DESCENDING] ;
MODEL *depvar* = *indepvar(s)* /*options;*
RUN;

Interpretation of SAS System-Generated Results

Tests of the Global Null Hypothesis

The default output generated by PROC LOGISTIC looks very similar to that generated by PROCs REG and/or GLM. This output includes several tests of overall model adequacy which test the global null hypothesis that *none* of the independent variables in the model are related to changes in probability of event occurrence. Of these, the **-2 LOG L** test is perhaps the most easy to interpret and is analogous to the "Global F" test used in a linear regression analysis. The computation of and rationale for the **-2 LOG L** test, among others, is found in Hosmer and Lemeshow (1989). Other global tests, such as the **SCORE, Akaike Information Criterion,** and **Schwartz Bayesian Criterion** are also provided but are beyond the scope of this paper.

Tests of the Local Null Hypotheses
Tests of the 'statistical significance' of each independent variable are also provided. The **Wald Chi-Square** test (and its associated p-value) are printed along with the **parameter estimate** and **standardized parameter estimate.** As with linear regression analysis, the parameter estimate can be conceptualized as how much mathematical impact a unit changes in the value of the independent variable has on increasing or decreasing the probability that the dependent variable will achieve the value of one in the population from which the data are assumed to have been randomly sampled.

The Odds Ratio
Exponentiation of the parameter estimate(s) for the independent variable(s) in the model by the number *e* (about 2.17) yields the **odds ratio,** which is a more intuitive and easily understood way to capture the relationship between the independent and dependent variables. This quantity is automatically portrayed in PROC LOGISTIC starting in Release 6.07; users with earlier SAS System releases can easily compute this quantity by hand.

The odds ratio gives the increase or decrease in probability that a unit change in the independent variable has in the probability that the event of interest will occur. Two analytic scenarios will be presented here to further motivate this concept: a) categorical independent variable; and, b) continuous independent variable. Both examples are drawn from Hosmer and Lemeshow's (1989)

study of patient survival after admission to a hospital intensive care unit (ICU).

a) categorical independent variable
A logistic regression model was implemented using 'admission type' as an independent variable. This variable was coded one if the patient was admitted to the hospital via emergency room and zero if the patient was admitted via another hospital 'service', such as surgery, cardiology, etc..

The resulting odds ratio for this model was 8.89, which suggests that a patient admitted via the emergency room is about 9 times more likely to die than a patient admitted from another service.

b) continuous level independent variable
Consider another analytic situation where the event of interest to be predicted is patient's survival after admission to a hospital intensive care unit (ICU), and the independent variable is age of patient in years. Application of PROC LOGISTIC to the Hosmer and Lemeshow data set yielded a parameter estimated for the variable AGE as 0.0275; exponentiation of that estimate gives an odds ratio of 1.028. In this example, a one unit (that is, one year) increase in a patient's age increases by 2.8 percent the chance they will die (i.e., acquire the event of interest). [Of course, while this result may be 'statistically significant', the clinical relevance to a health care provider of patient age, without regard to other prognostic factors (such as disease severity) may limit the practical usefulness of the results.]

Customized Odds Ratios
As with the previous example, a unit change in the values of the independent variable(s) may not be substantively relevant or useful to the analyst. In the ICU survival study, a five, ten, or twenty year change in patient age may be of more clinical relevance than a change of just one year. Customized odds ratios can be obtained by:

- hand, using a calculator
- placing the parameter estimates generated by PROC LOGISTIC into an output SAS data set using the OUTEST option and then working in the data step

using the **UNITS** option, which is available in Release 6.10 and above.

For example:

UNITS AGE = 5 10 20 ;

Placed after the MODEL statement would generate customized odds ratios for five, ten and twenty year changes in patient age.

Confidence Intervals for Odds Ratios

As with regression analysis, the parameter estimates and associated odds ratios are *point estimates* of the true value of these quantities in the population from which the data under analysis are assumed to have been randomly sampled. Confidence intervals for the odds ratios can be obtained by:

- manual calculation
- use of the **RISKLIMITS** option, which was first available in Release 6.07

By default, the **RISKLIMITS** option produces 95% confidence intervals around the odds ratios for each independent variable in the model. Users can obtain customized confidence intervals by using the **ALPHA** option.

Multiple Logistic Regression Model

Researchers are frequently interested in examining either the joint effect of two more independent variables on the likelihood of event outcome. In other situations the effect of a single independent variable is analyzed controlling for (that is, holding constant the effect of) other independent variables. In these situations a *multiple logistic regression model* is required, and is implemented by placing the names of the independent variables of interest to the right of the equals sign in the MODEL statement.

Multiple logistic regression model results generated by PROC LOGISTIC are interpreted in much the same way as are results obtained from a multiple logistic regression model: the parameter estimates (and resulting odds ratios) are the unique effect (if any) on the probability of event occurrence as if each independent variable were entered in to the model last. This is analogous to "Type III" sum of squares analysis provided by PROCs REG and/or GLM.

Automated Selection of 'Optimal' Subsets of Independent Variables

PROC LOGISTIC implements three common methods to automate selection of a 'best subset' of independent variables:

- forward selection
- backward elimination
- stepwise selection

Implementation occurs when the user codes the **SELECTION=** option to the right of the slash sign in the model statement. The name of the desired selection method is placed following the equals sign.

Assessing Model Fit

PROC LOGISTIC provides several means of assessing how well the logistic regression model fits the data. These include:

- Hosmer and Lemeshow Chi-Square Goodness of Fit
- R-square 'like' statistics
- Classification Tables

Hosmer and Lemeshow Test

This approach provides a chi-square-based test which assesses how well the data under analysis perform under the null hypothesis that the model fits the data. This test, implemented by the SAS System in Release 6.07, is called by the **LACKFIT** option and is discussed at length by the authors in their text.

R-square 'Like' Statistics

These measures, implemented in SAS System Release 6.10, provide a generalization of the coefficient of determination to the logistic regression model. Their derivation is found both in the Hosmer and Lemeshow text and in *SAS/STAT Software: Changes and Enhancements, Release 6.10.* Two statistics are printed if the **RSQUARE** option is used: the 'adjusted R-square' statistic is appropriate for models containing one or more dichotomous independent variables.

Classification Tables

This approach provides a convenient way to assess the:

- sensitivity
- specificity
- false positive rate
- false negative rate
- proportion of cases correctly classified

by a particular logistic regression model. Classification tables are generated by use of the **CTABLE** option; additional use of the **PPROB** option avoids generation of unnecessary output.

Additional Functionalities in PROC LOGISTIC

PROC LOGISTIC provides a number of additional functionalities and tests not addressed in this paper. These include:

- detection of outliers and influential observations
- generation of values for a Receiver-Operator Characteristics (ROC) curve to an output data set for subsequent plotting by PROCs PLOT and/or GPLOT
- generation of false positive and false negative rates using Baye's Theorem.

Note: SAS is the trademark of SAS Institute, Cary, NC

References

Hosmer and Lemeshow ,*Applied Logistic Regression,* Wiley:, 1989

SAS Institute, Inc: *SAS/STAT Software, Volume 2: the LOGISTIC Procedure*

SAS Institute, Inc.: *SAS/STAT Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07*

SAS Institute, Inc.: *SAS/STAT Software: Changes and Enhancements Release 6.10*

SAS Institute, Inc.: *Logistic Regression Examples Using the SAS System* (1995)