Logistic Regression Basics

Joseph J Guido, MS, Paul C Winters, MS, Adam B Rains, MSc
University of Rochester Medical Center, Rochester, NY

ABSTRACT

What is regression? What's the difference between linear and logistic regression? When and how should I use them? While these are common questions when students first encounter modeling procedures, there are very few sources which succinctly summarize the process for the SAS® system. After several years of teaching courses in the use of SAS/STAT® for public health data analysis, we developed a primer to quickly impart a working knowledge of logistic regression to our students. While logistic regression analyses may be performed using a variety of SAS® procedures (CATMOD, GENMOD, PROBIT, LOGISTIC and PHREG), this paper focuses on the LOGISTIC procedure as it is particularly well-suited to the needs of our students. Casting regression as a part of a systematic approach to data analysis, we use examples to demonstrate the LOGISTIC procedure's basic syntax (MODEL, CLASS, OUTPUT statements), model construction and selection options (FORWARD, BACKWARD, STEPWISE, HIERARCHY), and output interpretation. Where relevant the authors have highlighted parallels between LOGISTIC and the, more familiar, REG procedure used for linear regression. The authors hope this paper will serve as a concise reference for those seeking a rapid introduction to logistic regression in SAS®.

INTRODUCTION

Many students, when encountering regression in SAS for the first time, are somewhat alarmed by the seemingly endless options and voluminous output. Though the LOGISTIC procedure does indeed have its complexities, most problems and much confusion can be avoided by taking a systematic approach to your analyses. The most common problem we as SAS instructors encounter with regard to regression is misuse. Sadly, many students have a tendency to employ regression when simpler analytic methods would do or, more egregiously, rush straight to regression without completing the necessary preliminary analytics. Though the focus of this paper is primarily upon the LOGISTIC procedure, we want to emphasize that regression should be used as part of a deliberate and well-though-out analytic plan. Remember, no software or series of equations can do you thinking for you.

We generally recommend that, prior to commencing a logistic regression analysis, students develop an *a priori* conceptual model of how the variables in their dataset might interact and affect one another. Once this has been done descriptive frequencies for each variable (using the UNIVARIATE, MEANS, and/or FREQ procedure) should be examined to determine distribution, coding, and the presence of any potentially influential outliers. What you do with this information will depend largely upon the results you receive. Even in the event that you make no modification to the dataset prior to proceeding with your analyses, it is important that you have a good understanding of each variable and how it behaves. Only once you have gathered and interpreted basic descriptive information for your variables should you proceed to higher level analyses.

WHAT IS REGRESSION?

In simplest terms, regression is a statistical procedure which attempts to predict the values of a given variable, (termed the dependent, outcome, or response variable) based on the values of one or more other variables (called independent variables, predictors, or covariates). The result of a regression is usually an equation (or model) which summarizes the relationship between the dependent and independent variable(s). Typically, the model is accompanied by summary statistics describing how well the model fits the data, the amount of variation in the outcome accounted for by the model, and a basis for comparing the existing model to other similar models. By comparing these statistics across multiple models the user is able to determine a combination and order of independent variables that most satisfactorily predict the values of the outcome.

Numerous forms of regression have been developed to predict the values of a wide variety of outcome measures. Since the focus of regression modeling is on the response variable, the type of regression you use will be dictated by the type of response variable you are analyzing and by your eventual analytic goal. Given the limited space, we will limit our discussion to the kinds of regression most commonly used by our students: linear regression and logistic regression.

LINEAR REGRESSION

Since the syntax for the SAS LOGISTIC and REG procedures are somewhat similar, a quick review of linear regression would be valuable at this point. Linear regression is used to predict the values of a continuous outcome (dependent) variable based on the values of one or more independent predictor variables. The REG procedure in SAS optimizes this linear model based on the method of least squares. The result of this type of regression can be expressed as follows:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ . \ . \ . \ b_kx_k + e$$

Where **Y** represents the continuous dependent variable whose values are being modeled, b_0 is the y-intercept, and x_1 to x_k represent the k independent variables included in the model. Since no model fits the data perfectly an error term, **e**, is included to account for differences between the values predicted by the model and those observed in the dataset. The terms b_1 to b_k are coefficients indicating the degree of association between each independent variable and the outcome. In other words, each coefficient represents the amount of change we would expect in the outcome variable if there were a one-unit change in the independent variable.

The syntax for the REG procedure is characteristic of most regression code in the SAS system. Apart from options common to all SAS procedures (e.g., DATA=), you'll notice a few unique commands. Since REG, like most SAS modeling procedures, is interactive (allowing you to submit multiple models within one PROC step) and hence requires a QUIT statement to explicitly terminate the modeling process.

```
PROC REG DATA=Dataset Options;
  MODEL Dependent = Independent(s) / Options;
RUN;
QUIT;
```

Another commonality among SAS modeling procedures is the MODEL statement. This statement, as we shall see later in reference to the LOGISTIC procedure, establishes the speculated relationship between the outcome and the various independent predictors.

LOGISTIC REGRESSION

The primary distinction between the REG and LOGISTIC procedures is that the latter is intended for the modeling of dichotomous categorical outcomes (*e.g.,* dead vs. alive, cancer vs. none,…). Logistic and linear regression are both based on many of the same assumptions and theory. While convenient in many ways this presents a minor problem with regard to the outcome. Since the outcome is dichotomous, predicting unit change has little or no meaning. As an alternative to modeling the value of the outcome, logistic regression focuses instead upon the relative probability (odds) of obtaining a given result category. As it turns out the natural logarithm of the odds is linear across most of its range, allowing us to continue using many of the methods developed for linear models. The result of this type of regression can be expressed as follows:

$$Ln\ [\ p\ /\ (1\text{-}p)\ ] = \ b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ . \ . \ . \ b_kx_k + e$$

Where **p** represents the probability of an event (*e.g.,* death), b_0 is the y-intercept, and x_1 to x_k represent the independent variables included in the model. As with the linear model, each independent variable's association with the outcome (log odds) is indicated by the coefficients b_1 to b_k . Again, an error term is included to account for differences between the observed outcome values and those predicted by the model. In effect, we are trying to model the probability that an event is a result of a linear combination of variables as indicated in the equation above. The general syntax resembles the REG procedure and is as follows:

```
PROC LOGISTIC DATA=Dataset Options;
  CLASS classification variables;
  MODEL Dependent = Independent(s) / Options;
RUN;
QUIT;
```

Note the addition of the CLASS statement. Unlike many other SAS procedures, where this statement is used to define analytic subgroups, CLASS in the LOGISTIC procedure instructs SAS to create dummy variables for each categorical variable you specify. By adding options to the CLASS statement, it is possible to control the dummy-creation process in a variety of ways (e.g., specification of a reference group, parameterization method,…).

Reading the output from PROC LOGISTIC can, at first, be somewhat daunting. Confusion can be minimized by taking your time to focus on each element of the output separately. If we look at the -2 log likelihood test (labeled "-2 LOG L" in the output), which is analogous to the Global F test in PROC REG, we have a starting point for determining whether the overall model is significant or not *(see David Hosmer and Stanley Lemeshow, Applied Logistic Regression, 2nd ed (2000) for details on computing this statistic)*. The Hosmer-Lemeshow Goodness-of-Fit test tells us whether we have constructed a valid overall model or not. To request this option, we use the keyword LACKFIT on the PROC LOGISTIC statement. If the model is a good fit to the data then the Hosmer-Lemeshow Goodness-of-Fit test should have an associated p-value greater than 0.05. While this information is included last in the output, it should be reviewed early to determine a valid model has been selected.

Similar to Linear Regression each independent variable is also tested for statistical significance. As part of the LOGISTIC output SAS generates a table showing each predictor variable with its calculated parameter estimate, variability, Wald Chi-Square test, and associated p-value. Since, under the logit transformation, the odds ratio for any given independent variable is simply Napier's constant (e) raised to the power of the parameter estimate, SAS is able to easily compute the associated odds ratios for significant independent variables.

Odds Ratio = e^b = 2.71828^b

The odds ratio can be interpreted as having a harmful or protective effect upon the subject depending on how far it deviates from 1 (i.e., no effect). Odds ratios whose confidence limits exclude 1 are statistically significant.

We will now consider a real life example to demonstrate PROC LOGISTIC. This example is taken from a Prostate Cancer Study from Hosmer and Lemeshow (2000). The goal of the analysis is to determine if variables measured at baseline can predict whether a tumor has penetrated the prostatic capsule. We have written the following PROC LOGISTIC code to analyze the data from this study.

```
PROC LOGISTIC DATA=Prostate DESCENDING;
  CLASS Dpros (PARAM=Ref REF=First);
  MODEL Capsule= Age|Race|Dpros|Dcaps|PSA|Vol|Gleason @2/
            SELECTION=Stepwise HIERARCHY=Multiple LACKFIT;
RUN;
QUIT;
```

Variables from the Dataset Prostate (Hosmer and Lemeshow, 2000):

Variable	Label	Values
ID	Patient ID	1 – 380
Capsule	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration
Age	Age in Years	Number
Race	Race of Patient	1 = White, 2 = Black
Dpros	Results of the Digital Rectal Exam	1 = No Nodule, 2 = Left Lobe, 3 = Right Lobe, 4 = Both Lobes
Dcaps	Detection of Capsular Involvement	1 = No, 2 = Yes
PSA	Prostatic Specific Antigen Value	mg / ml
Vol	Tumor Volume Obtained from US	cm3
Gleason	Total Gleason Score	2 - 10

SELECTION=STEPWISE HIERARCHY=MULTIPLE

The LOGISTIC Procedure
Model Information

Data Set	WORK.PROSTATE
Response Variable	CAPSULE Tumor Penetration of Prostatic Capsule
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	380
Number of Observations Used	374

Response Profile

Ordered Value	CAPSULE	Total Frequency
1	1:Penetration	151
2	0:No Penetration	223

Probability modeled is CAPSULE='1:Penetration'. ❶

Stepwise Selection Procedure

Class Level Information ❷

Class	Value	Design Variables		
DPROS	1:No Nodule	0	0	0
	2:Unilobar Nodule (Left)	1	0	0
	3:Unilobar Nodule (Right)	0	1	0
	4:Bilobar Nodule	0	0	1

Step 0. Intercept entered:

Model Convergence Status ❸

Convergence criterion (GCONV=1E-8) satisfied.
-2 Log L = 504.526

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
134.1456	42	<.0001

[STEP 1, STEP 2, and STEP 3 OMITTED BECAUSE OF SPACE LIMITATIONS]

[SUMMARY OF STEPWISE LOGISTIC REGRESSION PROCEDURE]

Analysis of Maximum Likelihood Estimates ❹

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.4723	1.1401	55.2256	<.0001
DPROS	2:Unilobar Nodule(Lt)	1	1.3211	0.4753	7.7269	0.0054
DPROS	3:Unilobar Nodule(Rt)	1	1.4868	0.5057	8.6426	0.0033
DPROS	4:Bilobar Nodule	1	2.2866	0.6722	11.5729	0.0007
PSA		1	0.0275	0.00960	8.2231	0.0041
VOL		1	0.00451	0.0150	0.0900	0.7641
VOL*DPROS	2:Unilobar Nodule(Lt)	1	-0.0396	0.0202	3.8404	0.0500
VOL*DPROS	3:Unilobar Nodule(Rt)	1	0.00473	0.0205	0.0531	0.8178
VOL*DPROS	4:Bilobar Nodule	1	-0.0624	0.0322	3.7455	0.0530
GLEASON		1	1.0338	0.1681	37.8386	<.0001

Odds Ratio Estimates ❺

Effect	Point Estimate	95% Wald Confidence Limits	
PSA	1.028	1.009	1.047
GLEASON	2.812	2.023	3.909

Hosmer and Lemeshow Goodness-of-Fit Test ❻

Chi-Square	DF	Pr > ChiSq
7.5638	8	0.4772

❶ This statement indicates the category of the outcome variable (Capsule) modeled in the regression output.

❷ This table indicates which variables were included in the CLASS statement and the manner in which temporary dummy variables were created and their coding.

❸ Model converges and -2 Log L (analogous to Global F-test in PROC REG) shows that overall model is significant (see Residual Chi-square test p< .0001).

❹ The “Analysis of Maximum Likelihood Estimates” table summarizes information regarding the independent variables including parameter estimates, variability, and significance.

❺ The “Odds Ratio Estimates” table summarizes the significant independent variables and indicates their associated odds ratios and confidence limits.

❻ The “Hosmer and Lemeshow Goodness of Fit Test” indicates the quality of model fit. If the associated p-value is significant (p<0.05) this would be an indication that we need to rethink our analytic strategy.

DISCUSSION

Now that we have run the PROC LOGISTIC and annotated the partial output let's discuss some of the options we have used by examining the SAS code.

```
PROC LOGISTIC DATA=Prostate DESCENDING;
```

This statement above tells SAS to begin a logistic regression analysis using the SAS dataset "Prostate". Since the DESCENDING option has been specified, SAS knows that we are attempting to model the probability of the higher of our two outcome values (*i.e.,* 1 or capsule penetration). In the absence of the DESCENDING option SAS would have attempted to model the probability of the lowest numbered outcome category (in this case, 0 or no capsule penetration).

```
CLASS Dpros (PARAM=Ref REF=First);
```

The statement above tells SAS we have a classification variable that we want to use in our analysis. In our case since the variable Dpros takes on 4 values or levels the CLASS statement will temporarily create 3 dummy variables for the analysis. Since we want the result to be reported relative to the "no nodule" category we use the PARAM= and REF= options to specify this category. The option "PARAM=Ref" tells SAS we want to use reference group parameterization, in which each category's effect is expressed relative to a single category. The "REF=First" option that follows indicates which category is to be used as the referent group (the category against which the other categories are compared). Since the desired referent group is the "no nodule" category, and this category is coded "1", we can identify it explicitly by specifying "REF=1" or, more generically, "REF=First". In the absence of this statement SAS would have automatically selected the highest numbered category to conduct the parameterization.

```
MODEL Capsule= Age|Race|Dpros|Dcaps|PSA|Vol|Gleason @2/
      SELECTION=Stepwise HIERARCHY=Multiple LACKFIT;
```

This statement above tells SAS that our dependent variable is "Capsule" and that are independent variables are "Age", "Race", "Dpros", "Dcaps", "PSA", "Vol" and "Gleason". By using a bar between each variable name, and following the model statement with the "@2" option indicates to SAS that we want to include second order interaction terms in the model. Our selection option, STEPWISE, tells SAS that the independent variables should be entered into the model, evaluated, and then retained or discarded based on their assessed contribution to the overall model. Another popular method is FORWARD which is similar to STEPWISE except that once a term is entered into the model it stays in regardless of whether it remains significant. There is also the BACKWARD selection method which begins with all independent variables in the model (a saturated model) and progressively eliminates those that fail to meet the specified criterion (usually a parameter estimate p-value <0.05).

We have also included the HIERARCHY option which tells SAS that the main effects (the variables alone) and their associated interaction terms be moved in or out of the model in the manner specified (i.e. SINGLE, MULTIPLE, etc). In this example the "HIERARCHY=Multiple" option informs SAS to move the independent variables and their associated interaction terms as a unit.

The OUTPUT statement, not included here, is used to preserve the output from a given procedure by placing it in a separate dataset file. This output file includes not only the variables from the analysis dataset but numerous other diagnostic measures. Similarly, the OUTROC option (when placed on the MODEL statement) will create a dataset containing measures of the model's sensitivity, specificity, predictive values, and error rates. Datasets created by the OUTPUT and OUTROC options can be plotted and used to further determine the quality of the model.

CONCLUSIONS

While theory and methods underlying logistic regression are far too involved to be addressed entirely in this paper, our hope is that this short work will provide a sound foundation for students and other seeking to explore logistic regression in SAS.

REFERENCES

SAS Institute Inc., SAS/STAT User's Guide, Version 9.1. Cary, NC, SAS Institute Inc, 2002-2003

Hosmer, David and Lemeshow, Stanley. *Applied Logistic Regression*, *2nd Edition,* John Wiley and Sons, 2000.

Karp, Andrew. *Getting Started with PROC LOGISTIC*, NESUG 13 Proceedings, pp 709-13, Philadelphia 2000

Allison, Paul. *Logistic Regression Modeling Using the SAS System:Theory and Applications*, SAS Institute, 1998.

ACKNOWLEDGMENTS

SAS is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors in care of:

Joseph J Guido, MS
University of Rochester Medical Center
Department of Community and Preventive Medicine
300 East River Road, Suite G-100
Rochester, NY 14623
Work Phone: (585)273–2671
Fax: (585)276–2054
Email: Joseph_Guido@urmc.rochester.edu
Web: http://www.urmc.rochester.edu/cpm/directory/jguido.html