Paper 087-2012

Black Box PROCs: PROC LOGISTIC Discovered

Isabel H. Perry, Kaiser Permanente, Oakland, CA

ABSTRACT

Black box procedures are incredibly useful for the modern statistician, but they might not be aligned to what we think is happening behind the scenes. Students might find themselves consulting many resources to ensure that the code that they have just copied and pasted from an Internet source is performing the theory that they have just learned from their textbook. This paper combines statistical theory and the SAS® code needed to implement a binary logistic regression analysis using health care data.

INTRODUCTION

As a recent graduate student I've been catapulted into the "real world" of data analysis which results in many nights cozying up to my favorite statistics textbooks and SAS white papers. In an effort to consolidate these resources I began to write a comprehensive guide with statistical theory and the SAS programming needed to execute it with a student's perspective in mind. This paper reveals the theory behind the black box SAS procedure, PROC LOGISTIC. This procedure is not the only way to analyze data via a binary logistic regression but doing so enables students or novice SAS users to have a focused understanding of the coding syntax and thus use it confidently. I choose to begin with the logistic procedure because it has been ubiquitous in my biostatistics classes in graduate school and my recent healthcare data analyst position with Kaiser Permanente so it has proven valuable to learn.

BACKGROUND

Many categorical variables have only two categories, for example, those who have a disease and those who do not, smokers or nonsmokers, male or female, etc. Logistic regression models are a special case of generalized linear models (GLMs) with a categorical response variable as binary (e.g. yes or no, success or failure) or ordinal (e.g. high, medium, low). The explanatory variables can be either qualitative (categorical), quantitative or a combination of both. In this paper the binary logistic regression model is highlighted using a generated data set.

All GLMs have three components: a random component which assumes a probability distribution for the response variable, the systematic component which identifies the explanatory variables (denotes the expected value of the response), and the third is the link describing the function relationship between the random and systematic components. For a binary logistic regression the random component has a binomial distribution, the link function is the logit transformation of the response of linear regression.

So for a binary response we have P(Y=1) = π of success (e.g. probability disease is present) and P(Y=0) = $1 - \pi$ for a failure (e.g. probability disease is not present) where Y is the response variable.

Think back to linear regression. One approach to modeling the effect of X (the explanatory variable) is when the expected value of Y is a linear function of X.

$$\pi = \alpha + \beta x \quad \text{(EQ2)}$$

For binary logistic regression, or logit models, the probability on the left-hand side of the linear regression model is transformed such that values are appropriately bounded between 0 and 1.

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \quad \text{(EQ2)}$$

This logistic function above is used to estimate the probability an event (e.g. a patient has a disease) will occur as a function of the explanatory variable(s). In simple terms, this equation predicts the probability a

person will have the event of interest (e.g. disease is present) as a function of the explanatory variables at hand.

An alternative formula of EQ2 can be derived with algebra to the following form.

$$\pi = (e^{\alpha+\beta})/\,(1 + e^{\alpha+\beta}) \quad \text{(EQ3)}$$

BASIC SKELETON FOR PROC LOGSTIC

Knowing the basic form of PROC LOGISTIC is useful to know what must be included as a bare minimum. Not included these components would result in an error.

```
PROC LOGISTIC DATA=<Data Set Name> <OPTIONS>;
      MODEL <Dependent Variable> = <Independent Variable(s)> / <OPTIONS>;
RUN;
```

This logistic regression has continuous explanatory variable(s) and the categorical response variable (either binary or ordinal). Notice we are able to specify options that will change the analysis or enable more output.

STRUCTURING THE DATA SET PRIOR TO ANALYSIS

It is important to be mindful of how the data set is structured in order to run PROC LOGISTIC correctly and also ensure you are analyzing what you intended. For example, there is more than one way to flag the response variable as having the event of interest (e.g. disease) and a corresponding way to tell SAS which one you want to model (e.g. either disease or no disease)

Intuitively, we flag our data as zero for not having the event and one for the event of interest, however, PROC LOGISTIC will model on whatever value is the lowest. This means when the above basic syntax is ran, SAS will model the probability a patient does not have the disease. This is not usually what the researcher wishes to answer. We are most interested in the probability we have the disease and how the explanatory variables relate to it. So a simple solution is to either change the way you are flagging the event in the data set (e.g. “1” for disease and “2” for no disease) or include the DESCENDING option.

```
PROC LOGISTIC DATA = <Data Set Name> DESCENDING;
      MODEL <Dependent Variable> = <Independent Variable(s)> / <OPTIONS>;
RUN;
```

The DESCENDING option models in the highest value rather than the default of the lowest. This option would not be necessary if our data was flagged as “Disease” and “No Disease” because the lowest value here is the event of interest, “Disease”.

If we have grouped the data based on similar attributes then the DESCENDING option is not necessary.

```
DATA MYDATA;
      INPUT VAR1 VAR2 EVENTS N @@;
DATALINES;
7 1.0 0 10  14 1.0 0 31  27 1.0 1 56  51 1.0 3 13
7 1.7 0 17  14 1.7 0 43  27 1.7 4 44  51 1.7 0  1
7 2.2 0  7  14 2.2 2 33  27 2.2 0 21  51 2.2 0  1
7 2.8 0 12  14 2.8 0 31  27 2.8 1 22  51 4.0 0  1
7 4.0 0  9  14 4.0 0 19  27 4.0 1 16
;
RUN;
PROC LOGISTIC DATA = MYDATA;
      MODEL EVENTS/N = VAR1 VAR2;
RUN;
```

The above data set, MYDATA, has two exploratory variables, VAR1 and VAR2, and has counted the number of events, EVENTS, out of a total of N observations for that combination level of VAR1 and VAR2.

Another way the data may be structured is a similar to the above example except now we have an event variable, OPINION.

```
DATA PIE;
      INPUT TEMP APPLES OPINION COUNT @@;
DATALINES;
350 5 1 250 350 5 0 389 375 4 1 177 375 4 0 283
;
RUN;
PROC LOGISTIC DATA = PIE DESCENDING;
      FREQ COUNT;
      MODEL OPINION = APPLES TEMP;
RUN;
```

The OPINION is flagged as “1” for the apple pie being liked and “0” for the pie being disliked. The exploratory variables are the temperature cooking level (TEMP) and the number of apples included (APPLES) the pie. The COUNT variable is very useful because it is a much more concise way to express how many people rated the pie given the temperature and apples. Otherwise, for example, we would have to repeat the data line “350 5 1” 250 times.

Another way that I believe is the most direct way of coding successes for the binary variable is to include an EVENT option in the model statement.

```
PROC LOGISTIC DATA = PIE;
      FREQ COUNT;
      MODEL OPINION(EVENT=1) = APPLES TEMP;
RUN;
```

GENERATING THE HEALTHCARE DATA

The dummy data set is generated with a DATA step and the below table lists their metadata.

SAS name	Type	Role	Values
HEART	Binary	Response	1=Had a heart attack, 2=No heart attack
AGEGRP	Multinomial	Input	Age groups '20 to 35', '36 to 50', and '51 to 65'
GENDER	Binary	Input	1=Male, 2=Female
SMOKER	Binary	Input	1=Smoker, 0=Nonsmoker
CALORIE	Binary	Input	1=Over consumes, 0=Balanced Diet

The code below is used to produce the data set.

```
*Generate random data;
DATA ATTACK;
      DO I=1 TO 300;
            DO GENDER=1 TO 2; *1=Male, 2=Female;
                  H=RANUNI(1);
                  S=RANUNI(2);
                  C=RANUNI(3);
                  AGE=FLOOR(20+(65-20)*RANUNI(3)); *Set ages 20 to 65;
                  IF AGE LE 35 THEN AGEGRP='20 to 35';
                  ELSE IF AGE GT 35 AND AGE LE 50 THEN AGEGRP='36 to 50';
                  ELSE IF AGE GT 50 THEN AGEGRP='51 to 65';
```

```
                    IF H LE .40 THEN HEART=1; *1=Had a heart attack, 0=No heart
                    attack;
                    ELSE HEART=0;
                    IF S LE .30 THEN SMOKER=1; *1=Smoker, 0=Nonsmoker;
                    ELSE SMOKER=0;
                    IF C LE .70 THEN CALORIE=1; *1=Over consumes, 0=Balanced
                    Diet;
                    ELSE CALORIE=0;
                    *Manipulating responses;
                    IF H LE .60 AND HEART=0 AND (AGEGRP='36 to 50' OR SMOKER=1
                    OR CALORIE=1) THEN HEART=1;
                    OUTPUT;
            END;
      END;
      DROP I H S C;
RUN;
```

ANALYSIS

In order to analyze the data with a binary logistic regression, we need to include a CLASS statement in the LOGISTIC procedure that tells SAS which variables are categorical. In this example all the variables are categorical. You will also see variable options in the CLASS statement that denote the reference level. This is not required but helpful to make useful comparisons within the groups. Another option to note in the CLASS statement is the PARAM=REF option. This is necessary if we want the coefficients from the model to be consistent with the odds ratios. Without this option, SAS will use coding effect coding (e.g. -1=male, 1=female) and not dummy variable coding (e.g. 0=male, 1=female). Dummy variable coding produces exponentiated coefficients consistent with the odds ratios.

```
PROC LOGISTIC DATA=ATTACK;
      CLASS AGEGRP(REF='20 to 35') GENDER(REF='1') SMOKER(REF='0')
      CALORIE(REF='0') / PARAM=REF;
      MODEL HEART(EVENT='1')= AGEGRP GENDER SMOKER CALORIE;
RUN;
```

This MODEL statement above produces a main effects model with four factors but could easily be altered to a full model with interactions.

```
MODEL HEART(EVENT='1')= AGEGRP|GENDER|SMOKER|CALORIE;
```

Model Selection

You may have noticed from the output that many of the variables are not significant when testing the local null hypotheses for an individual explanatory variable. A great utility of PROC LOGISTIC is using its MODEL SELECTION option which automatically chooses the optimal subset of explanatory variables via forward, backward, or stepwise selection.

```
MODEL HEART(EVENT='1')= AGEGRP GENDER SMOKER CALORIE / SELECTION=STEPWISE;
```

After doing this automatic model selection the only variable left in the model is AGEGRP. For demonstration purposes we will keep all the variables in the model.

Model Equation

The logistic regression model (EQ2) for the LOGISTIC procedure with main effects only is the following.

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where π = P(heart attack) and

$x_1 = 1$ for the age group '36 to 50', 0 otherwise
$x_2 = 1$ for the age group '51 to 65, 0 otherwise
$x_1 = x_2 = 0$ for the age group '20 to 35'
$x_3 = 1$ for female, 0 otherwise for male
$x_4 = 1$ for smokers, 0 otherwise for nonsmokers
$x_5 = 1$ for those who over consume their daily calorie allowance, 0 otherwise for balanced eaters

The fitted logistic regression model and parameter estimates from the PROC LOGISTIC output.

$$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -0.24 + 0.51agegrp_{36\ to\ 50} - 0.17agegrp_{51\ to\ 65} + 0.14gender + 0.07smoker + 0.23calorie$$

Interpretation

The LOGISTIC output already supplies us with the odds ratios but to see how they are consistent with the parameter estimates we exponentiate the beta coefficients in the model. Below is part of the LOGISTIC procedure output giving the odds ratios and their 95% Wald confidence intervals.

The LOGISTIC Procedure

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
AGEGRP	36 to 50 vs 20 to 35	1.669	1.127	2.471
AGEGRP	51 to 65 vs 20 to 35	0.847	0.567	1.263
GENDER	2 vs 1	1.148	0.830	1.589
SMOKER	1 vs 0	1.069	0.749	1.525
CALORIE	1 vs 0	1.257	0.881	1.794

The estimated odds for smokers having a heart attack are 1.069 times the odds for nonsmokers when all other variables remain constant.

We could also calculate the probability of a heart attack,$\hat{\pi}$, as a function of the variables by the following fitted equation (EQ3).

$$\hat{\pi} = \frac{\exp(-0.24+0.51agegrp_{36\ to\ 50}-0.17agegrp_{51\ to\ 65}+0.14gender+0.07smoker+0.23calorie}{1+\exp(-0.24+0.51agegrp_{36\ to\ 50}-0.17agegrp_{51\ to\ 65}+0.14gender+0.07smoker+0.23calorie}$$

So if we had a 52 year old female smoker who consumed over their daily calorie allowance, the estimated probability that she would have a heart attack is .73 but keep in mind this model has not been validated for goodness of fit and all variables have been kept in the model for demonstration purposes.

Global Null Hypothesis Test

This test is investigating whether any of the variables in the model are related to changes in probability of the event of interest. In other words it tests whether our model is at all useful. The null hypothesis is $Ho: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus $Ha: At\ least\ one\ \beta_i\ is\ not\ equal\ to\ zero$. Here our test statistic is the Likelihood Ratio (L-R) Chi-Square test, $\chi_{L-R}(df)$. This is often confused with goodness of fit tests. The global null hypothesis test does not assess the quality of the model nor how well the model fits the data, but whether any of the variables are meaningful.

The LOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	830.834	826.576
SC	835.231	852.958
-2 Log L	828.834	814.576

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.2583	5	0.0140
Score	14.1158	5	0.0149
Wald	13.8972	5	0.0163

The L-R test statistic is referred to as the deviance, $G^2 = -2(L_0 - L_1)$, where L_0 the log-likelihood for the null model is and L_1 is the log-likelihood for the alternative model. From the output, $-2L_0 = 828.834$ and $-2L_1 = 814.576$ so $\chi_{L-R}(5) = 828.834 - 814.576 = 14.2583$ with p-value less than 0.05 (assuming a significance level of 0.05) the conclusion is to reject the null hypothesis and we can claim that at least one variable in our model is useful.

MODEL FIT

Goodness of Fit Test

There are several tests to assess the fit of the model one of which will be demonstrated in this paper. A popular choice for goodness of fit testing is the Hosmer-Lemeshow test which tells us if the model we have selected is appropriate for the data. We are looking for a p-value greater than our significance level of 0.05. The LACKFIT option in the model statement requests the test be done.

```
PROC LOGISTIC DATA=ATTACK;
      CLASS AGEGRP(REF='20 to 35') GENDER(REF='1') SMOKER(REF='0')
      CALORIE(REF='0') / PARAM=REF;
      MODEL HEART(EVENT='1')= AGEGRP GENDER SMOKER CALORIE / LACKFIT;
RUN;
```

From the output, we have a p-value of 0.0992 which, although is greater than 0.05, it provides weak evidence for a valid model.

This test is useful for attaining a general idea of model fit, but does not give insight about the nature of the fit. Another way to go about model validation is to compare your final model against more complex models, such as the saturated model. The saturated model is the most complex model of your variables at hand. The deviance could be attained from both your model and the saturated model of which the difference is the statistic for comparing the two models. The statistic, the difference between the models' deviances, is large when the chosen model fits poorly compared to the saturated model.

CONCLUSION

This paper touches lightly on calculations and hypotheses behind the LOGISTIC procedure, as well as general guidelines to a binary logistic regression using a dummy healthcare data set. Learnings from this paper can be translated to other models with a binary response variable and more complex models including interactions. This paper does not include a comprehensive explanation of all the options or functionalities in the LOGISTIC procedure but may provide a student or novice statistician starting ground for understanding what is being done in the black box.

REFERENCES

Agresti, Alan (1996). *An Introduction to Categorical Data Analysis*. New York, NY: John Wiley & Sons, Inc.

Agresti, Alan (2002). *Categorical Data Analysis*, Second Edition. New York, NY: John Wiley & Sons, Inc.

SAS Institute Inc. SAS OnlineDoc® 9.2
http://support.sas.com/cdlsearch?ct=80000

CONTACT INFORMATION

You can contact Isabel via
e-mail: perryih@gmail.com
website: www.isabelcurve.com.