Paper TU09

The Logic and Logistics of Logistic Regression

Lawrence Rasouliyan and Dave P. Miller
ICON Clinical Research, San Francisco, California

ABSTRACT

Although logistic regression models are widely used in multivariable analyses with dichotomous outcomes, many of their features, which can be very helpful tools in better understanding the data, are often underutilized. In addition to the widely used and reported odds ratios and p-values, PROC LOGISTIC generates a plethora of statistics from which one can gain further insight, make stronger analytical inferences, and more easily identify errors in the model construction. Model options will be explored, and resulting outputs from real-world examples will be explained in detail. Topics covered include model fit statistics, maximum likelihood estimates, effect modification, Receiver Operator Characteristic (ROC) curves, and the Hosmer-Lemeshow Goodness of Fit Test.

INTRODUCTION

Logistic regression is extremely important in observational study designs, particularly in the analysis of health-related data. From estimating the relationship between smoking and lung cancer to determining which risk factors are associated with hypertension to observing whether a particular preventive therapy reduces the risk of experiencing an adverse event, logistic regression methods are applicable in any study where the outcome of interest is dichotomous; that is, either the outcome occurs, or it does not occur.

At the univariable level, when one wants to determine whether an exposure is associated with an outcome, a 2 x 2 frequency table is often constructed with the exposure status on one axis and the outcome status on the other. A chi-square test (or Fisher's exact test) is then performed to determine whether a statistically significant association exists. Alternatively, an odds ratio can be calculated by cross-multiplication (dividing the product of the concordant cells by the product of the discordant cells), and one can test whether the odds ratio is significantly different from unity.

Even if a significant association exists, however, one can only make limited conclusions at the univariable level unless the predictor variable was a randomized treatment assignment. In observational studies, comparisons in which the potential effects of other underlying variables (covariates) have not been taken into account are of limited use. Logistic regression is a method of multivariable data analysis in which associations between exposures and outcomes can be assessed while adjusting for variables that could potentially confound the relationship. When the final logistic regression model has been established, the odds of experiencing the outcome among subjects with any set of underlying characteristics can be compared to that among subjects with any other set of underlying characteristics, as specified by the model.

THE LOGIC

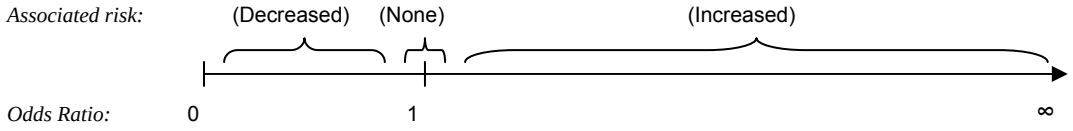
Most often when describing whether the outcome occurs, binary notation is used denote its presence (outcome=1) or absence (outcome=0). In logistic regression, however, the value of the outcome itself is not used as the dependent variable (in contrast to linear regression, where the outcome variable is continuous rather than dichotomous). The reason that the values of 1 or 0 are not regressed upon the dependent variable itself is because the resulting model will produce predicted values of the outcome that are impossible. That is, after the regression coefficients have been determined and values substituted for the independent variables, the predicted dependent variables could be less than 0, greater than 1, or in between 0 and 1. Since the outcome, by definition, either occurs (outcome=1) or it does not occur (outcome=0), how should these other predicted values be interpreted?

The ultimate goal of logistic regression is to establish a model in which one can predict the association between having a given set of factors and experiencing the outcome, while controlling for any extraneous variables. The measure of association derived from logistic regression is the odds ratio (OR), which is the odds of experiencing the outcome for subjects in one group divided by the odds of experiencing the outcome for subjects in another group. The OR ranges from 0 to infinity with a null value (no association) of 1. Therefore, when OR = 1, no association exists between exposure and outcome; when OR > 1, an increased

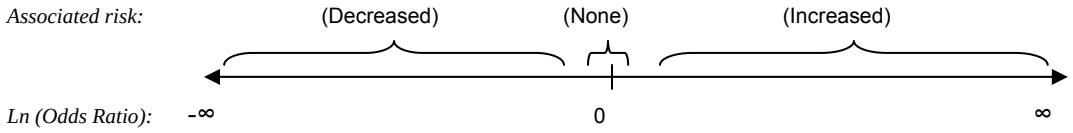
risk is associated with the exposure; and when OR < 1, a decreased risk is associated with exposure (Figure 1a).

Figure 1. *Odds ratio values and associated risk between exposure and outcome.*

(a)



(b)



One can deduce from Figure 1a that the distribution of odds ratios is skewed, with a smaller range of possible values occurring in the realm of decreased risk (from 0 to 1) than that of increased risk (from 1 to ∞). When applying parametric methods to odds ratios, it is preferable for this distribution to be symmetryic; therefore, the odds ratio is often logarithmically transformed. The possible distribution of Ln(OR) is symmetric and ranges from negative infinity to positive infinity, with 0 being the null value (Figure 1b).

These issues of continuity and symmetry described above, however, are overcome in logistic regression. Instead of modeling the value of the outcome (0 or 1) as the dependent variable, logistic regression models the odds of the outcome occurring (making it continuous) after being log transformed (making it symmetric). The resulting expression of the dependent variable is the log odds of the outcome occurring (also called the logit). Ultimately, the log odds is modeled as a linear function of the independent variables (Equation 1):

$$\text{Logit}\,[P(\text{outcome})] = Ln\left(\frac{P(\text{outcome})}{1 - P(\text{outcome})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \quad \text{(Equation 1)}$$

The odds of experiencing the outcome is then obtained by exponentiating both sides of the equation as follows (Equation 2):

$$\left(\frac{P(\text{outcome})}{1 - P(\text{outcome})}\right) = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots] = \exp[\beta_0]\exp[\beta_1 X_1]\exp[\beta_2 X_2]\exp[\beta_3 X_3]\ldots \quad \text{(Equation 2)}$$

The estimated β coefficients from this model, which correspond to the log odds ratios, have an approximately normal distribution. Therefore, one can predict the odds of the outcome occurring for subjects with any of the characteristics specified in the model by substituting in the applicable values of the independent variables into Equation 2. To calculate an odds ratio between two groups, simply calculate the odds of experiencing the outcome in each group and divide these values. By applying this methodology, one can calculate the odds ratio comparing subjects with any set of risk factors to subject with any other set of risk factors.

THE LOGISTICS

The most common procedure used in SAS® for performing logistic regression is PROC LOGISTIC, which has a multitude of options from which one can develop and acquire strong insight into the model. The general syntax of PROC LOGISTIC, as used in the context of this paper, is as follows (SAS Institute):

```
proc logistic < options >;
  where where-expression-1 < logical-operator where-expression-n>;
  class variable <(v-options)> <variable <(v-options)>... >
                    < / v-options >;
  model events/trials = < effects > < / options >;
  output < OUT=SAS-data-set > < keyword=name...keyword=name > /
                     < option >;
run;
```

In the CLASS statement, one can list non-ordinal categorical variables (e.g., gender, race) and specify the desired referent groups (Pasta); therefore, it is not necessary to reparameterize nominal variables into indicator variables. In the MODEL statement, the dichotomous outcome is specified on the left side of the equals sign, and the independent variables (both continuous and categorical) are specified on the right. In the OUTPUT statement, statistics generated from the model can be written to a dataset.

EXAMPLE

In order to illustrate some of the logistics of PROC LOGISTIC, data from a national asthma registry will be used as an example. The Epidemiology and Natural History of Asthma: Outcomes and Treatment Regimens (TENOR) is a prospective, observational, multi-center, large scale study of patients who were diagnosed as having severe or difficult-to-treat asthma. Details about the TENOR registry have been described elsewhere (Dolan). For this particular example using the TENOR data, we will be focusing on an analysis in which we want to determine which risk factors are associated with a clinical outcome that we have cleverly disguised and will refer to as poor lung function outcome (PLFO) to avoid any suggestion that the authors of this paper are making independent clinical inferences about the data. The variable PLFO will be assigned a value of 1 if the poor lung function outcome occurs and a value of 0 if it does not occur.

SPECIFYING THE MODEL

The first step in most analyses is to examine all potential predictors of the outcome at the univariable level and test for statistical significance with a t-test (for continuous variables) or a chi-square test (for categorical variables). Once variables significantly associated with the outcome have been identified, we can begin constructing the logistic regression model to see if these associations retain significance at the multivariable level. It is also possible for variables that are not significant at the univariable level to be significant at the multivariable level, but the focus of this analysis was restricted to significant univariable predictors. In this analysis, several clinical and demographic characteristics were tested for significant associations with PLFO. As a first step in constructing an appropriate model, we identified the significant characteristics and put them all into the model at once (variable names described by the comments):

```
* Logistic model – all significant univariable terms;

proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') hxar (ref='0') / param=ref;
  model PLFO (event='1') =
    /* Age per decade              */   agebl_dec
    /* Gender                      */   sexmf
    /* Black race                  */   raceblack
    /* Hispanic ethnicity          */   racehisp
    /* IgE >30, IgE >100           */   igegt30    igegt100
    /* Present smoker, Past smoker */   presentsmk pastsmk
    /* Skn test done, Skn test pos.*/   skntstdone skntstpos
    /* Hx of allergic rhinitis     */   hxar
    /* Diabetic                    */   idiabet
    /* Fam hx of allergy           */   ihxallergy_fam
    /* Fam hx of al. rhinitis      */   ihxar_fam
    /* Fam hx of atopic dermatitis */   ihxad_fam
    /* Moisture at home            */   imoist
```

```
  /* Mold at home                */   imold
  /* Has at least 1 pet          */   ipet
  /* Has a cat                   */   icat
  /* Has a dog                   */   idog
  /* Has multiple pets           */   imultpet
  /* Emotional stress triggers sx*/   sxemyn
  /* Sinus infection triggers sx */   sxsiyn
  /* Dust triggers sx            */   sxduyn
  /* Aspirin triggers sx         */   sxasyn
  /* Asthma duration per decade  */   amdurbl_dec
  /* College grad or higher      */   collgrad
  /* Unemployed                  */   unemployed
  /* Medicaid insurance          */   medicaid;
run;
```

In the WHERE statement, the standard analytic dataset was subset to include only those patients for whom PLFO status was known (defined as either 0 or 1). Although the standard analytic dataset is only comprised of patients who have non-missing values of PLFO, this step was taken as a precautionary measure. PROC LOGISTIC automatically excludes observations for which the outcome variable or independent variables in the model are missing.

The CLASS statement allows for parameterization of categorical variables via the `param=ref` option. One can specify the referent group (`ref=`) in parentheses after each variable. It should be noted that the value of the referent group must be placed in quotation marks regardless whether the variable is numeric or character. The variable for gender (sexmf) in the dataset was originally coded as 1 for females and 2 for males. Since the parentheses following sexmf in the CLASS statement specify that the referent group has a value of 1 (female), the logistic regression will model the risk for PLFO associated with being male relative to being female. The variable denoting history of allergic rhinitis (hxar) is also included in the class statement because it is a three-level variable, coded as 2 for yes, 1 for uncertain, and 0 for no. Since the value of 0 is specified as the referent group in the class statement, PROC LOGISTIC will generate two coefficients for this term: one coefficient for yes relative to no, and another for uncertain relative to no. For more details on parameterizing variables with the CLASS statement in PROC LOGISTIC, see “Parameterizing models to test the hypotheses you want: Coding indicator variables and modified continuous variables” (Pasta).

In the MODEL statement, the outcome variable (PLFO) is specified on the left side of the equals sign. By default, SAS® uses the lesser alphanumeric value of the outcome variable to denote an event occurring. In this analysis, we are interested in modeling when PLFO occurs (rather than when PLFO does not occur); however, based on the way the variable is coded, SAS® will model PLFO=0 as the event. In order to reverse the polarity of the outcome variable, we can specify in parentheses the value that SAS® should use as the event (`event='1'`). By using this syntax, PROC LOGISTIC now models the event as PLFO=1. This has no effect whatsoever on the p-values from the model, but it is critical for assessing the directionality of any identified effects. On the right side of the equals sign are the independent variables. All of the independent variables in this model are indicator variables (coded as 1 for yes and 0 for no) with the exception of the class variables previously mentioned and age per decade (agebl_dec) and asthma duration per decade (amdur_dec) which are continuous variables. The patient’s age and duration of asthma were converted into decades by dividing the number of years by 10. This conversion was done because it is more tangible to talk about a greater amount of risk per 10 years rather than a lesser amount of risk per year.

MODEL OUTPUT

After the preceding code is run, PROC LOGISTIC generates a plethora of statistics. One of the first elements of the output is the response profile:

```
          Response Profile

 Ordered                      Total
   Value        PLFO     Frequency

       1           0           393
       2           1           598

Probability modeled is PLFO=1.

NOTE: 26 observations were deleted due to missing values for the response or
explanatory variables.
```

The response profile describes how the outcome variable was parameterized. From the output, we can determine that 393 patients in our model did not have PLFO while 598 patients did. The output then confirms that PLFO=1 was modeled as the event (since we specified `event='1'` in the MODEL statement). It then goes on to note that 26 patients read from the dataset did not contribute to the model due to missing data. Since we specified the model to read only those patients who had non-missing values for PLFO in the WHERE statement, we know that the missing data must come from the independent variables.

The class level information is then presented which describes how the variables in the CLASS statement were parameterized:

```
Class Level Information

                      Design
Class     Value     Variables

sexmf     1          0
          2          1

hxar      0          0      0
          1          1      0
          2          0      1
```

The design variables are basically a set of indicator variables that SAS® creates internally to parameterize the class variables. Since we specified a value of 1 for sexmf as the referent group, SAS® assigns the value of the corresponding design variable to be 0 (which is the referent group with indicator variables). Since hxar is a three-level variable, SAS® creates an additional design variable to specify the additional level for allergic rhinitis status.

Further down on the output is the analysis of maximum likelihood estimates. This section is where the regression coefficients (also called parameters) are estimated. These values correspond to the βs from Equations 1 and 2. Below is the abbreviated output:

```
            Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter           DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept            1     -1.5256      0.6217        6.0213        0.0141
agebl_dec            1      0.3716      0.0657       32.0322        <.0001
sexmf          2     1      0.9011      0.1811       24.7661        <.0001
raceblack            1      0.6612      0.3026        4.7733        0.0289
racehisp             1     -1.0497      0.3767        7.7647        0.0053
igegt30              1      0.1306      0.2156        0.3671        0.5446
igegt100             1      0.1455      0.1913        0.5781        0.4471
presentsmk           1      1.3993      0.4224       10.9734        0.0009
pastsmk              1      0.4757      0.1767        7.2427        0.0071
skntstdone           1      0.2927      0.4264        0.4713        0.4924
skntstpos            1     -0.3947      0.3744        1.1112        0.2918
hxar           1     1     -0.5996      0.3921        2.3383        0.1262
hxar           2     1     -0.6863      0.3742        3.3626        0.0667
idiabet              1      0.1964      0.3336        0.3467        0.5560
...
```

The very first parameter listed is the intercept (β_0 in the equations), and it represents the predicted value of the logit when all other covariates in the model are equal to 0. Often times when continuous variables are in logistic regression models, the value of the intercept itself may have little practical meaning because it corresponds to the predicted value of a patient whose age is 0 years, weight is 0 pounds, height is 0 inches, etc. The other coefficients occur in the order they were specified in the MODEL statement. Between the columns marked “Parameter” and “DF” exists an untitled column where variables that appeared in the CLASS statement are indicated. The value in this column denotes that of the comparison group (the non-referent group). For sexmf, a value of 2 (male) is indicated because the referent group was specified as having a value of 1 (female) in the CLASS statement, and the estimate corresponds to that of males in

reference to females. The variable hxar appears twice in the output. Since it was a three-level variable and the referent group was specified as having a value of 0 (no allergic rhinitis), PROC LOGISTIC indicates a 1 and 2 to denote the estimates for the uncertain and yes values, respectively, relative to the no value. Further evidence of this parameterization is the indication of 1 degree of freedom for each level of the hxar estimates.

The values under the "Estimate" column represent the regression coefficients. Since these values are derived from modeling the log odds of the outcome as a linear function of the covariates, a value of 0 indicates no association, a positive value indicates increased risk, and a negative value indicates decreased risk (Figure 1b). The next column contains the standard errors of the coefficients. The Wald chi-square is equivalent to (Estimate / Std Error)^2. The final column indicates the p-values corresponding to the Wald chi-squares and indicates whether the coefficients are significantly different than 0.

In the next section of the output, the odds ratios are calculated from exponentiating the coefficients (Equation 2):

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
agebl_dec		1.450	1.275	1.649
sexmf	2 vs 1	2.462	1.727	3.511
raceblack		1.937	1.070	3.506
racehisp		0.350	0.167	0.732
igegt30		1.140	0.747	1.739
igegt100		1.157	0.795	1.683
presentsmk		4.052	1.771	9.274
pastsmk		1.609	1.138	2.275
skntstdone		1.340	0.581	3.091
skntstpos		0.674	0.323	1.404
hxar	1 vs 0	0.549	0.255	1.184
hxar	2 vs 0	0.503	0.242	1.048
idiabet		1.217	0.633	2.340
...				

Again for class variables, the output specifies which values are referent. For sexmf, the odds ratio corresponds to the odds of experiencing PLFO among males (sexmf=2) compared to that among females (sexmf=1). The odds ratio of 2.462 indicated that males are at increased risk of experiencing PLFO, and since the 95% confidence interval does not encompass the value of 1, the odds ratio is statistically significant. Similarly Hispanic ethnicity (racehisp) has an odds ratio of 0.350 (95% CI: 0.167-0.732). Therefore, the model suggests that being Hispanic is protective against PLFO. It should be noted that odds ratio confidence intervals are not symmetric around the point estimates because the standard error estimated from the model applies to the log odds (coefficient parameter estimates). Consequently, after exponentiating these values, the symmetry is lost.

SELECTION OF MAIN EFFECTS

Selecting which variables go into the final logistic regression model is as much of an art as it is a science. One must take into account issues such as statistical significance, clinical plausibility, collinearity, and effect modification. Several different statistically-driven algorithms exist for selecting the final model such as forward approaches (adding variables into the model one at a time and assessing the significance of each term), backward approaches (starting with all variables in the model and removing them one at a time), and stepwise approaches (allowing variables to be added or removed upon successive revisions). Furthermore, different significance levels for entry and exit criteria can be specified.

In assessing PLFO, we chose a stepwise selection method, with an entry and exit criteria of 0.05. In SAS® these different model selection methods are automated and can be performed by specifying the option at the end of the MODEL statement, where you can specify the method of selection (`selection`=), and the p-value criteria for entry into (`slentry`=) and exit from (`slstay`=) the model:

```
* Stepwise logistic regression;

proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') hxar (ref='0') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp igegt30 igegt100
                           presentsmk pastsmk skntstdone skntstpos hxar idiabet
                           ihxallergy_fam ihxar_fam ihxad_fam imoist imold
                           ipet icat idog imultpet sxemyn sxsiyn sxduyn
                           sxasyn amdurbl_dec collgrad unemployed medicaid
                           / selection=stepwise slentry=0.05 slstay=0.05 details;
run;
```

To perform a forward or backward selection method, one can simply substitute those words for "stepwise" in the selection option. Similarly, different p-values for entry and exit can be specified. The details option tells PROC LOGISTIC to print the details of model selection (i.e., the intermediate steps with respect to which variables entered and exited the model). At the very end of the output, the variable selection steps for the final model are summarized:

The LOGISTIC Procedure

Summary of Stepwise Selection

Step	Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	amdurbl_dec		1	1	92.9297		<.0001
2	agebl_dec		1	2	57.0683		<.0001
3	sexmf		1	3	37.8539		<.0001
4	raceblack		1	4	17.8072		<.0001
5	presentsmk		1	5	10.8102		0.0010
6	pastsmk		1	6	8.5910		0.0034
7	ihxad_fam		1	7	7.7448		0.0054
8	sxduyn		1	8	6.5230		0.0106
9	collgrad		1	9	7.0821		0.0078
10	ipet		1	10	5.4084		0.0200
11	racehisp		1	11	5.8737		0.0154

The output summarizes the order in which the variables were added and removed and the corresponding p-values. For our model, a total of 11 variables were added, and none were removed. All p-values are less than 0.05 as specified by the exit criteria.

EFFECT MODIFICATION

The logistic regression model with the resulting terms from the stepwise selection was presented to the investigators of the study. For this particular asthma outcome, however, it was previously suspected that the disease might manifest itself differently in males compared to females. In other words, the association between at least one of the risk factors and PLFO might be different depending on whether the patient is male or female. This phenomenon is known as effect modification, or interaction, because the effect of a risk factor may be modified depending on the group to which the patient belongs.

Effect modification was assessed by refitting the model multiple times, once for each of the main effects (generated from the stepwise selection process). In each model, a single interaction between gender and one of the other variables is added to the final main effects model. For instance, the model would be run with the main effects terms and the term for gender-college graduate interaction. Then the model would be run with the main effects terms and the gender-age interaction. All possible gender-main effects terms were assessed, and the only one that attained statistical significance was that between gender and duration of asthma. Interaction terms can be expressed in PROC LOGISTIC through multiplication syntax:

```
* Model with gender*duration of asthma interaction;

proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp presentsmk
                           pastsmk ihxad_fam ipet sxduyn amdurbl_dec collgrad
                           sexmf*amdurbl_dec;
run;
```

The resulting coefficients are as follows:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2035	0.4056	29.5126	<.0001
agebl_dec		1	0.3608	0.0603	35.7844	<.0001
sexmf	2	1	1.4646	0.3267	20.0966	<.0001
raceblack		1	0.8101	0.2821	8.2434	0.0041
racehisp		1	-0.8205	0.3555	5.3284	0.0210
presentsmk		1	1.3501	0.3957	11.6429	0.0006
pastsmk		1	0.4827	0.1710	7.9643	0.0048
ihxad_fam		1	-0.5862	0.1936	9.1690	0.0025
ipet		1	-0.3841	0.1530	6.3020	0.0121
sxduyn		1	-0.4603	0.1907	5.8285	0.0158
amdurbl_dec		1	0.4661	0.0622	56.1684	<.0001
collgrad		1	-0.3761	0.1568	5.7548	0.0164
amdurbl_dec*sexmf	2	1	-0.2194	0.1119	3.8461	0.0499

Therefore, we are observing a significant interaction between gender and duration of asthma. That is, the effect that duration of asthma has in predicting PLFO is different between males and females.

FURTHER REFINEMENT

After the main effects plus gender-asthma duration interaction model was presented to the investigators, there were a few additional variables they wanted to consider which were marginally insignificant in construction of the main effects model but had clinical relevance. These variables were history of allergic rhinitis (hxar), family history of allergic rhinitis (ihxar_fam), sinus infection triggering of asthma symptoms (sxsiyn), and aspirin sensitivity (sxasyn). Similar to how interaction was assessed, these marginal variables were individually added to and removed from the main effects plus gender-asthma duration model to assess whether any would attain statistical significance.

```
* Final logistic regression model;

proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp presentsmk
                           pastsmk ihxad_fam ipet sxduyn amdurbl_dec collgrad
                           sexmf*amdurbl_dec sxasyn;
run;
```

Aspirin sensitivity (sxasyn) was the only marginal variable that remained significant and was permanently added to the model to produce the final analytic model, which was comprised of the original main effects, the gender-asthma duration interaction, and aspirin sensitivity.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2898	0.4087	31.3897	<.0001
agebl_dec		1	0.3629	0.0604	36.1270	<.0001
sexmf	2	1	1.4947	0.3273	20.8615	<.0001
raceblack		1	0.7880	0.2819	7.8119	0.0052
racehisp		1	-0.8132	0.3572	5.1835	0.0228
presentsmk		1	1.3703	0.3969	11.9187	0.0006
pastsmk		1	0.4945	0.1713	8.3278	0.0039
ihxad_fam		1	-0.5746	0.1943	8.7437	0.0031
ipet		1	-0.3647	0.1538	5.6245	0.0177
sxduyn		1	-0.4691	0.1912	6.0196	0.0141
amdurbl_dec		1	0.4622	0.0621	55.4528	<.0001
sxasyn		1	0.4307	0.2184	3.8909	0.0485
collgrad		1	-0.3576	0.1575	5.1533	0.0232
amdurbl_dec*sexmf	2	1	-0.2208	0.1118	3.9026	0.0482

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
agebl_dec	1.437	1.277	1.618
raceblack	2.199	1.265	3.822
racehisp	0.443	0.220	0.893
presentsmk	3.936	1.808	8.569
pastsmk	1.640	1.172	2.294
ihxad_fam	0.563	0.385	0.824
ipet	0.694	0.514	0.939
sxduyn	0.626	0.430	0.910
sxasyn	1.538	1.003	2.360
collgrad	0.699	0.514	0.952

REPARAMETERIZING

The output for the final model includes coefficients for all terms specified in the MODEL statement; however, PROC LOGISTIC does not generate odds ratios for variables that participate in interactions. No odds ratios are reported for gender or asthma duration. The reason is because if effect modification exists, the odds ratios, by definition, will be different by group. That is, the odds ratio for asthma duration will be different for males and females. In order to obtain these odds ratios and confidence intervals, we can reparameterize the asthma duration variable by creating separate asthma duration variables for males and females:

```
data reparam01;
  set saslib.sad_PLFO;

  * Duration term for females;
  if sexmf eq 1 then amdurbl_dec_female = amdurbl_dec;
  else if sexmf eq 2 then amdurbl_dec_female = 0;
  else put "Warning: Female duration not categorized";

  * Duration term for males;
  if sexmf eq 2 then amdurbl_dec_male = amdurbl_dec;
  else if sexmf eq 1 then amdurbl_dec_male = 0;
  else put "Warning: Male duration not categorized";

run;
```

The duration variable for females (amdurbl_dec_female) is equal to the actual asthma duration if the patient is female and set to 0 if the patient is male. Similarly, the duration variable for males (amdurbl_dec_male) is set equal to the asthma duration if the patient is male and set to 0 if the patient is female. The model can now be regenerated using the gender-specific asthma duration terms in place of the original asthma duration term and the interaction term:

```
proc logistic data=saslib.sad_PLFO;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp presentsmk
                           pastsmk ihxad_fam ipet sxduyn sxasyn collgrad
                           sxasyn

                           amdurbl_dec_male amdurbl_dec_female;
run;
```

The resulting output now specifies the gender-specific duration odds ratios in addition to the gender odds ratio itself:

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
agebl_dec		1.437	1.277	1.618
sexmf	2 vs 1	4.458	2.347	8.467
raceblack		2.199	1.265	3.822
racehisp		0.443	0.220	0.893
presentsmk		3.936	1.808	8.569
pastsmk		1.640	1.172	2.294
ihxad_fam		0.563	0.385	0.824
ipet		0.694	0.514	0.939
sxduyn		0.626	0.430	0.910
sxasyn		1.538	1.003	2.360
collgrad		0.699	0.514	0.952
amdurbl_dec_male		1.273	1.057	1.533
amdurbl_dec_female		1.588	1.406	1.793

This model does not explicitly include the interaction term, so it is not helpful for testing for the presence of the interaction. Since we have already established the significance of the interaction, we can refit the model in a way that is easier to interpret.

INTERPRETING THE FINAL MODEL

All of the odds ratios in the final model were statistically significant as evidenced by the exclusion of the null value (OR=1) in all 95% confidence intervals. Therefore, assuming that all other covariates are equal, we can make the following conclusions about each predictor:

- *Age per decade:* A 40-year old patient is 1.4 times more likely to experience PLFO compared to a 30-year old patient.
- *Black race:* Black patients are 2.2 times more likely to experience PLFO than non-Black patients.
- *Hispanic ethnicity:* Being Hispanic appears to be protective against PLFO, as Hispanics are less than half as likely to experience PLFO compared to non-Hispanics.
- *Smoking:* Being either a present smoker or a past smoker is associated with increased odds of experiencing PLFO. Past smokers are 1.6 times more likely and present smokers are almost 4 times more likely than non-smokers. Furthermore, we may be observing some type of dose-response relationship, since the odds ratio for present smokers is over twice that of past smokers.

- *Family history of atopic dermatitis, having a pet, dust sensitivity:* All these predictors have protective effects on PLFO
- *College graduate or higher:* Patients with an advanced degree are less likely to experience PLFO than patients without an advanced degree.
- *Gender:* Males are 4.5 times more likely than females to experience PLFO.
  - All other variables can be held equal only at asthma duration of zero. Thus, this odds ratio cannot be interpreted in isolation.
- *Asthma duration:* The longer a patient has had asthma, the more likely the patient is to experience PLFO
  - *Among males:* Male patients who have had asthma for 20 years are 1.3 times more likely to experience PLFO than male patients who have had asthma for 10 years.
  - *Among females:* Female patients who have had asthma for 20 years are 1.6 times more likely to experience PLFO than female patients who have had asthma for 10 years.

OTHER ANALYTICAL TOOLS

Hosmer-Lemeshow Goodness-of-Fit

This statistic can be generated in PROC LOGISTIC by specifying the lackfit option at the end of the MODEL statement. The Hosmer-Lemeshow test evaluates the null hypothesis that the specified model fits the data well. It does so by ranking the observations by predicted value into deciles and stratifying by true outcome status. It then evaluates the observed and the expected number of observations in each decile and calculates an overall chi-square statistic with 8 degrees of freedom.

Partition for the Hosmer and Lemeshow Test

		PLFO = 1		PLFO = 0	
Group	Total	Observed	Expected	Observed	Expected
1	101	12	16.93	89	84.07
2	101	30	29.27	71	71.73
3	101	32	40.58	69	60.42
4	101	61	50.40	40	50.60
5	101	62	59.38	39	41.62
6	101	69	67.77	32	33.23
7	101	81	75.40	20	25.60
8	101	81	82.73	20	18.27
9	101	88	89.07	13	11.93
10	99	89	93.47	10	5.53

Hosmer and Lemeshow Goodness-of-Fit Test

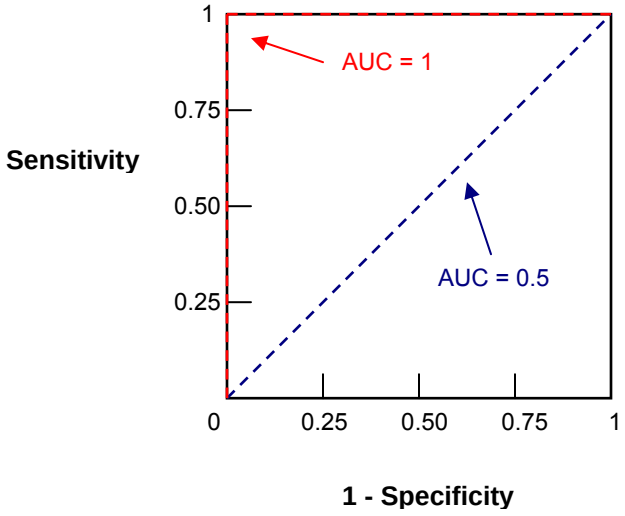
Chi-Square	DF	Pr > ChiSq
15.3499	8	0.0527

Since the p-value is not statistically significant, we fail to reject goodness of fit; however, the nearly significant p-value suggests some reason for concern. The observed and expected values differ most in deciles 3 and 4. This indicates that there may be an unusually sharp rise in risk between those with few risk factors and those with a moderate number of risk factors. Because major deviation is only seen in these two adjacent categories, it is quite possible that it is due only to random chance.

Receiver-Operating Characteristic (ROC) Analysis

Receiver-Operating Characteristic (ROC) analysis is a method in which a continuous measure is evaluated as a diagnostic tool in predicting a dichotomous outcome. Some common examples include using chance of rain (continuous measure) to predict whether rain will actually occur (dichotomous outcome) and using a credit score (continuous measure) to predict whether a card holder will have a delinquent payment (dichotomous outcome).

In our analysis, we want to evaluate our final model as a diagnostic tool for PLFO. In order to do so, the predicted possibility of PLFO, as specified by the final model, will be calculated for each patient (giving each patient a continuous "score"). Different cutoff values of this score are then assessed among patients stratified by true PLFO status. For each of these various cutoff points, sensitivity and specificity can be calculated. After several cutoff points have been evaluated, we plot sensitivity versus 1 - specificity (the true positive percentage verses the false positive percentage). The resulting plot is called the ROC curve, and it characterizes the ability of the final model to "diagnose" PLFO. A summary measure of how well the final model predicts PLFO is the area under the curve (AUC). If it is possible to obtain a high true positive percentage without rapidly increasing the false positive percentage, the curve rises sharply, and the total area under the curve is close to 1. When AUC=0.5, the continuous predictor is non-informative (essentially like flipping a coin); when AUC=1, the predictor is a perfect diagnostic measure. On the ROC curve, an AUC of 0.5 would be represented by the diagonal of the plot area staring at the origin, and an AUC of 1 would be represented by a step function of unity:

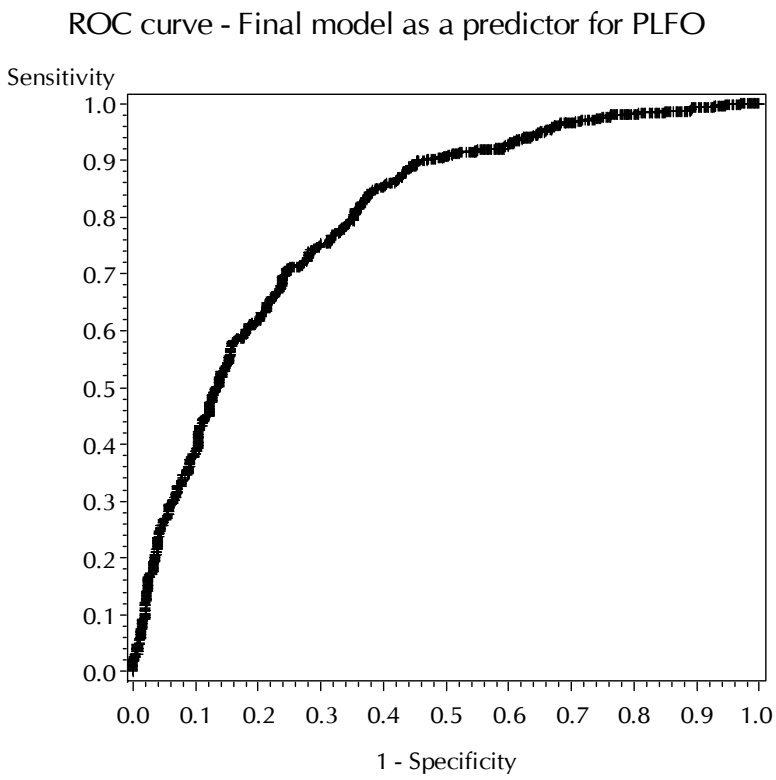


ROC data in PROC LOGISTIC is generated via the `outroc=` option in the model statement. SAS® basically runs through multiple cutoff points of the predictor and calculates the corresponding sensitivity and specificity, writing the resulting data to a specified dataset. To generate the ROC curve, we must plot the sensitivity versus 1 - specificity from the resulting dataset:

```
* Generating the ROC data;
proc logistic data=saslib.sad_plfo;
  where PLFO in (0,1);
  class sexmf (ref='1') / param=ref;
  model PLFO (event='1') = agebl_dec sexmf raceblack racehisp presentsmk
                           pastsmk ihxad_fam ipet sxduyn  sxasyn collgrad
                           sxasyn amdurbl_dec_male amdurbl_dec_female
                           / outroc=rocdata;
run;

* Plotting the ROC data;
proc gplot data=rocdata;
  title1 "ROC curve - Final model as a predictor for PLFO";
  plot _sensit_*_1mspec_;
run;
quit;
```

This code generates the following plot:



The ROC analysis is also closely associated with a set of summary statistics that are part of the default SAS® output:

```
Association of Predicted Probabilities and Observed Responses

Percent Concordant      79.6    Somers' D    0.594
Percent Discordant      20.2    Gamma        0.595
Percent Tied             0.2    Tau-a        0.286
Pairs                243815    c            0.797
```

These statistics are generated from pairing every single patient where PLFO=1 with every single patient where PLFO=0. The number of pairs in the output is simply the product of the group sample sizes (i.e., 243815 = 403 X 605). For every single pair, SAS® looks at whether the model score for the PLFO=1 patient is greater than that of the PLFO=0 patient. If so, then the pair is considered concordant; if not, then the pair is considered discordant. The percent concordant and percent discordant (and percent tied) are then calculated and displayed in the output. The value c represents the concordance index and is the percent concordant adjusted for ties. The concordance index also happens to be equivalent to the AUC. Therefore, in this example the AUC=0.797, indicating that the model has some diagnostic ability to predict PLFO. The Somers' D statistic is similar to the concordance index, except it resides on a -1 to 1 scale (rather than a 0 to 1 scale) and is equal to 2(c - 0.5). The values for Gamma and Tau-a also indicate how much better the model is compared to random chance, but they handle ties differently, and Tau-a ranges from 0 to 0.5 for the range of concordance values for which the concordance index ranges from 0.5 to 1.

CONCLUSION

Logistic regression is a very powerful method for determining which factors are associated with a dichotomous outcome while controlling for extraneous variables. PROC LOGISTIC generates a multitude of statistics that can be very helpful tools in constructing models, assessing risk, identifying effect modification, determining goodness of fit, and characterizing diagnostic ability. Through the various options, one can gain further insight into the data and ultimately make stronger analytical inferences.

REFERENCES

Dolan CM, Faher KE, Bleecker ER, Borish L, Chipps B, Hayden ML, Weiss S, Zheng B, Johnson C, Wenzel S. Design and baseline characteristics of The Epidemiology and Natural History of Asthma: Outcomes and Treatment Regimens (TENOR study): a large cohort of patients with severe or difficult-to-treat asthma. *Ann. Allergy Asthma Immunol.* 2004. 92(1): 32-9.

Pasta DJ. Parameterizing models to test the hypotheses you want: Coding indicator variables and modified continuous variables. Proceedings of the Western Users of SAS Software Twelfth Annual Conference.

SAS Institute, Inc. SAS/STAT®. User's Guide, Version 9. Cary, NC: SAS Institute Inc. 2004.

CONTACT INFORMATION

The authors welcome questions and comments. Please direct inquiries to:

Lawrence Rasouliyan
Research Manager, Statistical Analysis
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
lrasouliyan@ovation.org

Dave P. Miller
Senior Director, Statistical Analysis
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
dmiller@ovation.org