

An Introduction to Binary Logistic Regression

Linda M. Quinn
QED Industries, Inc.

ABSTRACT

Modeling the relationship between a binary response variable and one or more independent variables can be done within the SAS System using PROC LOGISTIC. This tutorial will explain how to run a basic logistic model, will discuss the available options, and will review the output generated. A new feature of PROC LOGISTIC starting in version 6.10 is the ability to display ROC curves and to generate customized confidence intervals. Examples and coding will be provided.

INTRODUCTION

There are many PROCs that are available to perform logistic regression: CATMOD, GENMOD, LOGISTIC, and PROBIT. There are special features within each procedure that make it useful; however, for most applications PROC LOGISTIC is preferred.

The goal of binary logistic analysis is to find the best fitting and most parsimonious, yet reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions.

EXAMPLE

For illustration, let's use a dataset on treatment success for a new medical therapy. There are only two outcomes a study patient can experience, success or failure (OUTCOME). Success is coded as "S" and failure as "F".

There are several variables that may affect the outcome. The factor of most importance is whether the patient was on the test therapy or on the control therapy (GROUP). Three other potential variables were identified: (1) age of the

patient (AGE), (2) compliance of the patient to the therapy (COMPLY), and (3) gender (GENDER).

The sponsor of this study is interested in modelling how successful treatment is. Data were collected on 100 subjects, 50 patients in each of the two treatment arms.

Now, if OUTCOME were a continuous variable, like weight or blood pressure, the next step would be to construct a regression equation to predict OUTCOME from a linear combination of GROUP, AGE, COMPLY, and GENDER. The equation might then express the risk of being a successful OUTCOME as a weighted sum of the four factors.

But probabilities don't go in a straight line forever; they are bounded by 0 and 1. So we would like to transform things so the dependent variable ranges smoothly between 0 and 1. One such transformation is the logistic transformation.

$\text{logit}(p_i) = \log(p_i / (1-p_i)) = \alpha + \beta'X_i$ where p_i is the probability of event i .

The logit function has the advantage that it is easily interpreted and can be applied to data collected retrospectively or prospectively.

THE CODE

Let's start by determining whether OUTCOME is determined by GROUP and COMPLY. GROUP is a dichotomous variable where 1 indicates the patient is in the test group and 0 indicates the patient is in the control group. COMPLY is a continuous variable, representing the percent of compliance the patient had to the prescribed treatment therapy and was scored from 0 to 100%. Later we will want to use GENDER and AGE in the models. However GENDER is coded as M/F and will not work well as an independent variable. A variable MALE was constructed as a 0/1 variable based on GENDER.

Similar to PROC REG, the model statement would be MODEL OUTCOME = GROUP COMPLY; So the basic code would look like:

```
PROC LOGISTIC DATA=TESTDATA;  
  MODEL OUTCOME = GROUP COMPLY;  
RUN;
```

But before running this, by default, PROC LOGISTIC models the probability of the response that corresponds to the lowest ordered value. Notice that our outcome variable isn't even 0 or 1, it is character-based. So, in our case, the lowest value of OUTCOME is "F" and PROC LOGISTIC will model the probability of failure as an OUTCOME. Since OUTCOME is dichotomous, we could then figure out the probability of success of OUTCOME. But why not let SAS do that for us – the DESCENDING option. Using the DESCENDING option in the PROC LOGISTIC statement reverses the default ordering of the response variable and may be the simplest solution.

So now, the base code looks like:

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;  
  MODEL OUTCOME = GROUP COMPLY;  
RUN;
```

Output from PROC LOGISTIC includes (see Appendix A):

1. Response Profile which gives the ordered value, the value of the response variable, and the frequency. The probability being modeled is the value of the response variable that corresponds to the first ordered value. This would be the first thing to check. You need to verify that the first ordered value is the level of the response variable that you want to analyze.
2. The test of the global null hypothesis of whether the regression coefficients are all zero. These are also useful when comparing different models against each other. The AIC (Akaike Information Criterion) and the SC (Schwarz Criterion) adjust the $-2 \log L$ ($-2 \log$ likelihood statistic) score based on the number of predictor

variables in the model. All should give approximately similar results.

3. The parameter estimates are the estimated coefficients for the logistic regression equation. The standard error is the standard error of the parameter estimate and can be used for calculating confidence intervals. The Wald Chi-Square is the test used to check for the significance of each of the regression parameters. The $\text{Pr} > \text{Chi-Square}$ is the p-value associated with the Wald Chi-Square.
4. The odds ratio (OR) is computed by exponentiating the parameter estimate. The odds ratio represents the risk associated with a unit increase in the independent parameter. Consequently, PROC LOGISTIC always estimates the odds ratio for a unit increase in the independent variable, not an increase from a to b. If you want an estimate of the odds ratio for an increase from a to b, you would exponentiate the odds ratio estimate to the $(b-a)$ power (or use the UNITS statement, yet to come).

So an odds ratio of 1 means that for a one unit change in the independent variable, the odds of having the event of interest is 1 times larger than not having the event (so, essentially no difference). An odds ratio larger than 1, say 3.5, means that for a unit increase in the independent variable, the odds of having the event of interest is 3.5 times larger. Likewise, for an odds ratio less than 1, say 0.8, the odds of having the event of interest is 0.8 times.

5. The Association of Predicted Probabilities and Observed Responses is a table of values used to assess the quality of the logistic model.

For all possible pairs of observations with different response values, agreement is checked. In the dichotomous case, all observations with the event of interest are paired with all observations without the event of interest. Therefore if there are 85 cases of the event and 115 cases of non-event, there would be $85 \times 115 = 9,775$ pairs. A pair is concordant if the observation with the higher response value has a lower predicted

value than the corresponding observation. Likewise, a pair is discordant if the observation with the higher response value has a higher predicted value. You would like high concordance.

The various rank correlation measures (Somer's D, Gamma, Tau-a, and c) also indicate degree of agreement. In addition, the c statistic is equivalent to the area under the receiver operator characteristic (ROC) curve. The closer to one the better. Agreement by chance alone would be a c statistic of 0.5.

INTERPRETING THE OUTPUT

So what do you do when you get one of these outputs? All right, this isn't the order I really use; but, I do look at all these things at one time or another.

First thing I look at is the information at the very top. Make sure you are using the dataset and response variable you want. Check the number of observations used and number of response levels. Observations are deleted from the analysis whenever any variable in the model has a missing response. Notice that I am using the correct dataset and OUTCOME has 2 levels. It also tells me the frequency distribution between success and failure.

Next, I look at the response profile to make sure I'm modeling the outcome I want to model (otherwise I have to invert all the conclusions I'll be making or run the model over with the correct options to get what I want). So in this case the lowest ordered value corresponds to OUTCOME =S, which is the event that OUTCOME is successful, which is what I want. Remember that we used the DESCENDING option to achieve this. There are also 65 successes and 35 failures. A good rule of thumb for deciding how many variables to include in the model is 10-15 observations with each level of the response variable for each variable to be included. In this case we should probably not model more than 3 variables.

Next, I look at the -2 Log L to see if, overall, the predictor variables (GROUP and COMPLY) contribute to explaining the probability of having OUTCOME. In this case the p-value is 0.0001 which tells me I have a significant model.

Now, the important part - which variables are significant? Looking at the p-value column, I can tell that both variables, GROUP and COMPLY are significant. Now what? I rarely look at the parameter estimates and I usually jump straight to the odds ratio. An odds ratio of 3.321 tells me that a patient on the test therapy is 3 times more likely to have a successful outcome than someone who is on the control therapy. Odds ratios are pretty easy to interpret when your independent variable is dichotomous.

Consider COMPLY which is continuous. It is significant with an odds ratio of 1.094. This means that for every unit change in COMPLY, you are 1.094 times more likely to be successful than someone with a compliance value of one less. So, a person with a compliance of 95% is 1.094 times more likely than someone with 94% compliance of having a successful outcome. Similarly, a person with 85% compliance is 1.094 times more likely of having a successful outcome than someone with 84% compliance. However, a person with 95% compliance is $(1.094)^{20} = 6.03$ times more likely to have a successful outcome than someone with 75% compliance. Remember that there is a relationship between the odds ratio and the parameter estimate. $OR = e^{\beta}$. So, $e^{1.2002} = 3.321$ and $e^{0.0901} = 1.094$.

Now, I would look at how good the model is. You like high concordance. We have 74.7% and the c statistic is 0.754.

If this were my final model, I would want to report the odds ratios with confidence intervals. I have done those by hand many times, but SAS does them for you now.

CONFIDENCE INTERVALS

Now, suppose we want to add confidence intervals for the parameters. Since version 6.10 there are two methods for calculating confidence limits for logistic regression: the likelihood method and the Wald method. The likelihood method is iterative and based on the asymptotic Chi-Square distribution. The Wald method is simpler and assumes an underlying normal distribution for the parameters. These two methods produce approximately the same results for large samples, but may produce very different results for small samples or large parameter estimates. There are four different

options to be used on the model statement PLCL, WALDCL, PLRL, and WALDRL. The first two calculate confidence intervals for the coefficients and the last two calculate confidence intervals for the odds ratio. The WALDRL option is synonymous with RL in previous SAS versions. I usually request just the Wald Risk Limits. By default, 95% confidence limits are calculated. Specifying an additional option on the model statement of ALPHA = .01 will produce 99% confidence limits.

Now, the code looks like:

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;
  MODEL OUTCOME = GROUP COMPLY / WALDRL;
  RUN;
```

The only thing that has changed on the output is the addition of confidence limits for the odds ratios.

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits	
			Lower	Upper
GROUP	1.0000	3.321	1.273	8.661
COMPLY	1.0000	1.094	1.044	1.147

Now we can say that a person in the test group is from 1.2 to 8.7 times more likely to have a successful outcome and for each extra percentage of compliance, a person is 1.044 to 1.147 times more likely to have a successful outcome. Notice that the confidence limits are for a one unit change. What if we were more interested in a five unit change in compliance?

CUSTOMIZED ODDS RATIOS

Remember that the parameter estimate, and corresponding odds ratios reflect what happens for a unit change in the predictor variable. Sometimes it is useful to see what happens for a 10 unit change, whether increased or decreased (e.g. age). Since version 6.10, a UNITS statement is now available which will do exactly that.

The expanded code:

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;
  MODEL OUTCOME = GROUP COMPLY / WALDRL;
  UNITS COMPLY = 5 -5;
  RUN;
```

This will now give confidence limits for both a 5 percent increase and a 5 percent decrease in compliance. Note that these values are just reciprocals of each other.

Instead of the original confidence limits section, the new confidence limits are shown:

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits	
			Lower	Upper
COMPLY	5.0000	1.569	1.240	1.985
COMPLY	-5.0000	0.637	0.504	0.807

Once again, we could have figured out the 5 unit increase in COMPLY. It should be $1.094^5 = 1.569$. Similarly, a five unit decrease should correspond to an odds ratio of $1 / 1.569 = 0.637$. So even if we didn't have the UNITS statement, we could hand calculate the same (and prior to version 6.10, you will). What is especially nice is that SAS will also calculate the confidence intervals for these 5 unit changes. However, please notice that the confidence intervals for GROUP do not show up any more. If you use the UNITS statement, it will only give you the confidence intervals for the variables listed on the statement.

MODEL SELECTION

Modeling is the process of deriving the best fitting, parsimonious model that adequately captures the outcome of interest. While many modelers know what the structure of the final model will be, others are exploring relationships and may be interested in looking at a variety of models. Similar to PROC REG, there is a stepwise option and a best subsets option on PROC LOGISTIC. While results are not shown

here, an example of the code to do stepwise regression using all the available variables is:

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;
  MODEL OUTCOME = GROUP COMPLY AGE
    GENDER / WALDRL STEPWISE DETAILS;
RUN;
```

GOODNESS OF FIT

Goodness of fit measures are objective measures of how well your model fits the data. In PROC LOGISTIC, there are three such measures - the Hosmer and Lemeshow goodness of fit statistic, the R^2 (generalized coefficient of determination) and the adjusted R^2 .

To get these statistics, the modified code looks like:

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;
  MODEL OUTCOME = COMPLY GENDER
    /WALDRL LACKFIT RSQ;
RUN;
```

In Appendix B you can find the R^2 and the adjusted R^2 under the global model information. If the RSQ option is not used on the model statement, it isn't there. The LACKFIT option gives the Hosmer-Lemeshow Goodness of Fit Test. The Hosmer-Lemeshow test, by default, divides the observations into approximately 10 groups and compares actual to expected values using a Chi-Square test. The R^2 of 0.2022 (and the adjusted 0.2784) tells me I'm not doing very good and the Hosmer-Lemeshow statistic, since it is significant ($p=0.0031$), tells me I may have a lack of fit that I should explore further.

There are some warnings when interpreting the Hosmer-Lemeshow statistic:

- It is highly dependent on how the observations are grouped.
- It is a conservative test.
- It has low power to detect nonlinearity in the predictor variables.
- It tends to give significant results with very large datasets no matter how well the model fits.

ROC CURVES

The receiver operator characteristic (ROC) curve is a graphical representation of the predictive accuracy of a logistic regression. So a curve with high predictive accuracy rises quickly while a curve with low predictive accuracy rises gradually. The best way to check predictive accuracy is to run validation data not used in deriving the model in the fitted model.

The ROC is a plot of sensitivity (measured on the y-axis) versus 1-specificity (measured on the x-axis). Ideally, you would like a model with high sensitivity and high specificity; therefore, the ideal ROC curve would go through the point (0,1) and have an area under the curve of 1. If your model was no better than chance alone, the ROC curve would be the 45° line with an area under the curve of 0.5.

The area under the ROC curve is the c statistic which is in the basic PROC LOGISTIC procedure. To get the data to do the actual plot, we need to output the data and plot it using PROC PLOT (although you cannot draw the line) or PROC GPLOT. To do this, use the OUTROC= option on the model statement. The OUTROC dataset has several variables, the two of interest are `_SENSIT_` and `_1MSPEC_`.

I usually add a few data points and a variable to plot a 45° reference line. The extra code is shaded. (ROC Output is not provided.)

```
PROC LOGISTIC DATA=TESTDATA DESCENDING;
  MODEL OUTCOME = GROUP COMPLY
    /WALDRL OUTROC=ROCPLLOT;
RUN;
DATA REF45;
  _SENSIT_ = 0; _1MSPEC_ = 0; LINE = 1; OUTPUT;
  _SENSIT_ = 1; _1MSPEC_ = 1; LINE = 1; OUTPUT;
RUN;
DATA ROCPLLOT;
  SET ROCPLLOT (IN=IN1) REF45;
  IF IN1 THEN LINE = 0;
RUN;
PROC GPLOT DATA=ROCPLLOT;
  PLOT _SENSIT_ * _1MSPEC_ = LINE
    / VAXIS = 0 TO 1 BY .1
      HAXIS = 0 TO 1 BY .1;
  TITLE 'ROC CURVE';
RUN;
```

OTHER TOPICS

Here are some brief notes on other important topics concerning logistic regression.

Similar to regression analysis, regression diagnostics tells you something about the fit of the model, the validity of the underlying assumptions, and influential observations or outliers. There are several measures available in PROC LOGISTIC. INFLUENCE and IPLOTS are two additional options on the model statement which will provide you with a series of diagnostic measures.

I have only discussed the most common link function, the logit. There are two others available: the nomit (or probit) and the complementary log-log. These can be specified using the LINK= option on the MODEL statement. The default is the logit function.

There are options available which will help to correct for overdispersion of the model (when the variance of the response variable exceeds the nominal variance for binomial outcomes). Before correcting for overdispersion, eliminate any systematic components. If after that, you still wish to correct for overdispersion, you can use the AGGREGATE and SCALE= options on the model statement.

There is an option available to specify classification tables out of PROC LOGISTIC. A classification table categorizes the event or non-event status with the probability of event. This can be specified with the CTABLE option on the MODEL statement.

Sometimes, a problem with logistic regression is finding a finite solution. This is sometimes referred to as infinite parameters. PROC LOGISTIC will give somewhat ambiguous messages like: There is possibly a quasicomplete separation in the sample space. The maximum likelihood estimate may not exist. What now? For complete or quasicomplete separation you could check your data for errors, discretize continuous variables, get more data, build a model with stepwise selection, or redefine your outcome variable.

PROC LOGISTIC can also be expanded for use beyond the dichotomous outcome to ordinal logistic regression.

FINAL NOTES

Unfortunately, the documentation for PROC LOGISTIC is all over the place. The initial documentation is in the *SAS/STAT® User's Guide*. Then enhancements made in release 6.07 are documented in *SAS Technical Report P-229*. Finally, there are enhancements made in version 6.10 are in *SAS Software: Changes and Enhancements, Version 6.10*. through version 6.12. However, there is now a book specifically showing examples of logistic regression which I have cited in the references. It is an excellent resource and is up to date on procedure options.

REFERENCES

Hosmer, DW and Lemeshow, S (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

SAS Institute, Inc. (1995), *Logistic Regression Examples Using the SAS® System, Version 6, First Edition*, Cary, NC: SAS Institute, Inc.

PRODUCT INFORMATION

SAS and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

AUTHOR CONTACT

Linda M. Quinn
QED Industries, Inc.
1286 Gary Blvd.
Brunswick, OH 44212-2912

Phone: 330-225-4184
Fax: 330-220-9036
e-mail: lmq2@po.cwru.edu

Appendix A

The LOGISTIC Procedure

Data Set: WORK.TESTDATA
 Response Variable: OUTCOME
 Response Levels: 2
 Number of Observations: 100
 Link Function: Logit

Response Profile

Ordered Value	OUTCOME	Count
1	S	65
2	F	35

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	131.489	112.905	.
SC	134.094	120.720	.
-2 LOG L	129.489	106.905	22.585 with 2 DF (p=0.0001)
Score			20.872 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-7.2046	1.9908	13.0974	0.0003	.	.
GROUP	1	1.2002	0.4891	6.0225	0.0141	0.332529	3.321
COMPLY	1	0.0901	0.0240	14.0500	0.0002	0.540910	1.094

Association of Predicted Probabilities and Observed Responses

Concordant = 74.7%	Somers' D = 0.509
Discordant = 23.9%	Gamma = 0.516
Tied = 1.4%	Tau-a = 0.234
(2275 pairs)	c = 0.754

Appendix B

The LOGISTIC Procedure

Data Set: WORK.TESTDATA
 Response Variable: OUTCOME
 Response Levels: 2
 Number of Observations: 100
 Link Function: Logit

Response Profile

Ordered Value	OUTCOME	Count
1	S	65
2	F	35

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	131.489	112.905	.
SC	134.094	120.720	.
-2 LOG L Score	129.489	106.905	22.585 with 2 DF (p=0.0001) 20.672 with 2 DF (p=0.0001)

RSquare = 0.2022

Max-rescaled RSquare = 0.2784

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-7.2048	1.9908	13.0974	0.0003	.	.
GROUP	1	1.2002	0.4891	6.0225	0.0141	0.332529	3.321
COMPLY	1	0.0901	0.0240	14.0500	0.0002	0.540910	1.094

Association of Predicted Probabilities and Observed Responses

Concordant = 74.7%
 Discordant = 25.3%
 Tied = 1.4%
 (2275 pairs)

Somers' D = 0.509
 Gamma = 0.516
 Tau-a = 0.234
 c = 0.754

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits	
			Lower	Upper
GROUP	1.0000	3.321	1.273	8.661
COMPLY	1.0000	1.094	1.044	1.147

Hosmer and Lemeshow Goodness-of-Fit Test

Group	Total	OUTCOME = S		OUTCOME = F	
		Observed	Expected	Observed	Expected
1	12	0	2.90	12	9.10
2	10	2	3.87	8	6.13
3	9	6	4.71	3	4.29
4	12	11	7.35	1	4.64
5	11	10	7.40	1	3.60
6	10	8	7.42	2	2.58
7	8	7	6.53	1	1.47
8	10	8	8.40	2	1.60
9	11	8	9.85	3	1.15
10	7	5	6.56	2	0.44

Goodness-of-fit Statistic = 23.177 with 8 DF (p=0.0031)