

An Introduction to Binary Logistic Regression

Linda M. Quinn
Case Western Reserve University

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 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 acknowledge that one of the most pleasurable afflictions mankind has ever seen is Rock and Roll Fever (RRF), which involves multiple systems. Early signs include hand and foot movement whenever music is played, singing along with the car radio, and a slight decline in hearing possibly from attending too many concerts or playing the CD player a tiny bit too loud.

One brave SASologist (okay, it was me), ventured forth to determine risk factors for this

affliction. Four potential variables were identified: (1) number of concerts attended before age 25 (CONCERTS), (2) size of the CD collection (CDSIZE), (3) family history of RRF (FAMHX), and (4) a Minimal Wavelength Heard measurement (MWH).

To date, it is unknown how contagious RRF is. The SASologist set out to model the predictors of RRF. Data were collected on over 200 subjects, 100 with confirmed RRF and 100 without RRF. What a coincidence!

Now, if RRF were a continuous variable, the next step would be to construct a regression equation to predict RRF from a linear combination of CONCERTS, CDSIZE, FAMHX, and MWH. The equation might then express the risk of contracting RRF 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 RRF is determined by FAMHX and CDSIZE. FAMHX is a dichotomous variable where 1 indicates it is present and 0 indicates it is absent. CDSIZE is a continuous, integral variable. Similar to PROC REG, the model statement would be `MODEL RRF = FAMHX CDSIZE;`

So the basic code would look like:

```
PROC LOGISTIC DATA=QUINN.ROCK;  
  MODEL RRF = FAMHX CDSIZE;  
RUN;
```

But before running this, by default, PROC LOGISTIC models the probability of the response that corresponds to the lowest ordered value. So, as in our case, the lowest value of RRF is 0 and PROC LOGISTIC will model the probability of absence of RRF. Since RRF is dichotomous, we could then figure out the probability of presence of RRF. 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=QUINN.ROCK DESCENDING;  
  MODEL RRF = FAMHX CDSIZE;  
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 (Schwartz 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. So notice that I am using the correct dataset and RRF has 2 levels.

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 RRF=1, which is the event that RRF is present, which is what I want. Remember that we used the DESCENDING option to achieve this. There are also 100 of each of the levels of RRF. A good rule of thumb for deciding how many variables to include in the model is 10 observations with each level of the response variable for each variable to be included.

Next, I look at the -2 Log L to see if, overall, the predictor variables (FAMHX and CDSIZE) contribute to explaining the probability of having RRF. 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, FAMHX and CDSIZE are significant. Now what? I rarely look at the parameter estimates and I usually jump straight to the odds ratio. An odds ratio of 11.394 tells me that a person with a family history of RRF is 11 times more likely to have RRF than someone who does not have a family history of RRF. Odds ratios are pretty easy to interpret when your independent variable is dichotomous.

Consider CDSIZE which is continuous. It is significant with an odds ratio of 2.35. This means that for every unit change in CDSIZE, you are 2.4 times more likely to have RRF than the CDSIZE below. So, a person with 5 CDs is 2.4 times more likely than someone with 4 CDs of having RRF. Similarly, a person with 20 CDs is 2.4 times more likely of having RRF than someone with 19 CDs. However, a person with 20 CDs is $(2.4)^{20} = 40,199,887$ times more likely to have RRF than someone with no CDs. Remember that there is a relationship between the odds ratio and the parameter estimate. $OR = e^{\beta}$. So, $e^{2.4331} = 11.394$ and $e^{.8542} = 2.350$.

Now, I would look at how good the model is. You like high concordance which we have (95.9%) and the c statistic is .964. The high concordance and c statistic would make me wonder about the use of contrived data.

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. In 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=QUINN.ROCK DESCENDING;
MODEL RRF = FAMHX / WALDRL;
RUN;
```

The only thing that has changed on the output is the addition of confidence limits for the odds ratios (see Appendix B). Now we can say that a person with a family history of RRF is from 3 to 38 times more likely to have RRF and for each extra CD, a person is 1.4 to 3.8 times more likely to have RRF. Notice that the confidence limits are for a one unit change. What if we were more interested in a five unit change in CDs?

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). In version 6.10, a UNITS statement is now available which will do exactly that.

The expanded code:

```
PROC LOGISTIC DATA=QUINN.ROCK DESCENDING;
MODEL RRF = CDSIZE FAMHX / WALDRL;
UNITS CDSIZE = 5 -5;
RUN;
```

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

Instead of the original confidence limits section, the new confidence limits are shown below.

Once again, we could have figured out the 5 unit increase in CDSIZE. It should be $2.35^5 = 71.6$. Similarly, a five unit decrease should correspond to an odds ratio of $1 / 71.6 = 0.014$. 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 FAMHX 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=QUINN.ROCK DESCENDING;
MODEL RRF = CDSIZE FAMHX CONCERTS
MWH / 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=QUINN.ROCK DESCENDING;
MODEL RRF = CDSIZE FAMHX
/WALDRL LACKFIT RSQ;
RUN;
```

In Appendix C 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

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits	
			Lower	Upper
CDSIZE	5.0000	71.607	6.438	796.494
CDSIZE	-5.0000	0.014	0.001	0.155

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 .62 (and the adjusted .84) tells me I'm doing pretty good and the Hosmer-Lemeshow statistic, since it is significant, 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_. In our case, the code for the regression and plot looks like:

```
PROC LOGISTIC DATA=QUINN.ROCK DESCENDING;
  MODEL RRF = CDSIZE FAMHX
    /WALDRL OUTROC=ROCPLLOT;
RUN;

SYMBOL1 I=JOIN V=NONE L=1;
PROC GPLOT DATA=ROCPLLOT;
  PLOT _SENSIT_ * _1MSPEC_
    / VAXIS=0 TO 1 BY .1
      HAXIS=0 TO 1 BY .1;
  TITLE 'ROC CURVE';
RUN;
```

I usually add a few data points and a variable to plot a 45° reference line.

```
PROC LOGISTIC DATA=QUINN.ROCK DESCENDING;
  MODEL RRF = CDSIZE FAMHX
    /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;

SYMBOL1 I=JOIN V=NONE L=1;
SYMBOL2 I=JOIN V=NONE L=10;
PROC GPLOT DATA=ROCPLLOT;
  PLOT _SENSIT_ * _1MSPEC_ = LINE
    / VAXIS=0 TO 1 BY .1
      HAXIS=0 TO 1 BY .1;
  TITLE 'ROC CURVE';
RUN;
```

The resulting ROC curve is in Figure 1.

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 normit (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, enhancements made in version 6.10 are in *SAS Software: Changes and Enhancements, Version 6.10*. So, it can be

difficult to look up the syntax for what you are trying to do since you need all three parts to put together a complete document. 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 version 6.10 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

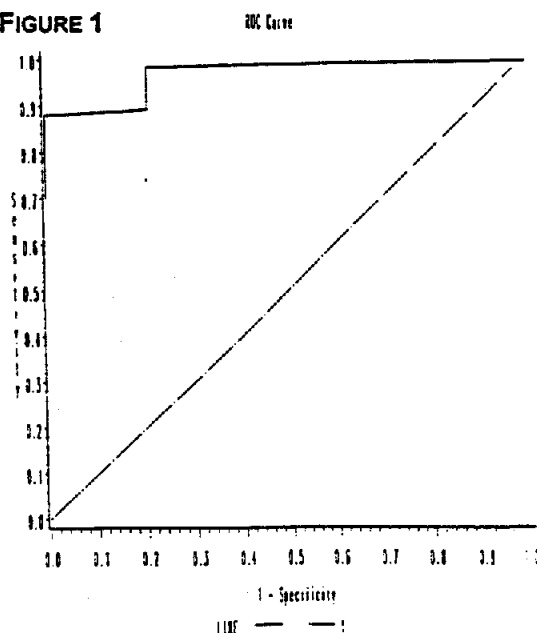
Linda M. Quinn
Research Associate
Case Western Reserve University
Division of General Internal Medicine and
The Program in Health Care Research
1286 Gary Blvd.
Brunswick, OH 44212-2912

Phone: 216-791-3800 x4814

Fax: 216-231-3476

e-mail: lmq2@po.cwru.edu

FIGURE 1



APPENDIX A

The LOGISTIC Procedure

Data Set: QUINN.ROCK
Response Variable: RRF
Response Levels: 2
Number of Observations: 200
Link Function: Logit

Response Profile

Ordered Value	RRF	Count
1	1	100
2	0	100

Testing Global Null Hypothesis: BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	279.259	85.330	.
SC	282.557	95.225	.
-2 LOG L Score	277.259	79.330	197.929 with 2 DF (p=0.0001) 117.153 with 2 DF (p=0.0001)

RSquare = 0.6283

Adjusted RSquare = 0.8377

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-3.7726	0.6120	38.0037	0.0001	.	.
FAMHX	1	2.4331	0.6134	15.7318	0.0001	0.671869	11.394
CDSIZE	1	0.8542	0.2458	12.0757	0.0005	8.058946	2.350

Association of Predicted Probabilities and Observed Responses

Concordant = 95.9%	Somers' D = 0.928
Discordant = 3.1%	Gamma = 0.937
Tied = 1.0%	Tau-a = 0.466
(10000 pairs)	c = 0.964

APPENDIX B

The LOGISTIC Procedure

Data Set: QUINN.ROCK
Response Variable: RRF
Response Levels: 2
Number of Observations: 200
Link Function: Logit

Response Profile

Ordered Value	RRF	Count
1	1	100
2	0	100

Testing Global Null Hypothesis: BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	279.259	85.330	.
SC	282.557	95.225	.
-2 LOG L Score	277.259	79.330	197.929 with 2 DF (p=0.0001) 117.153 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	-3.7726	0.6120	38.0037	0.0001	.	.
FAMHX	1	2.4331	0.6134	15.7318	0.0001	0.671869	11.394
CDSIZE	1	0.8542	0.2458	12.0757	0.0005	8.058946	2.350

Association of Predicted Probabilities and Observed Responses

Concordant = 95.9%	Somers' D = 0.928
Discordant = 3.1%	Gamma = 0.937
Tied = 1.0%	Tau-a = 0.466
(10000 pairs)	c = 0.964

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits Lower	Upper
FAMHX	1.0000	11.394	3.424	37.919
CDSIZE	1.0000	2.350	1.451	3.804

APPENDIX C

The SAS System

The LOGISTIC Procedure

Data Set: QUINN.ROCK
Response Variable: RRF
Response Levels: 2
Number of Observations: 200
Link Function: Logit

Response Profile

Ordered Value	RRF	Count
1	1	100
2	0	100

Testing Global Null Hypothesis: BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	279.259	85.330	.
SC	282.557	95.225	.
-2 LOG L	277.259	79.330	197.929 with 2 DF (p=0.0001)
Score	.	.	117.153 with 2 DF (p=0.0001)
RSquare = 0.6283		Adjusted RSquare = 0.8377	

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-3.7726	0.6120	38.0037	0.0001	.	.
FAMHX	1	2.4331	0.6134	15.7318	0.0001	0.671869	11.394
CDSIZE	1	0.8542	0.2458	12.0757	0.0005	8.058946	2.350

Association of Predicted Probabilities and Observed Responses

Concordant = 95.9%	Somers' D = 0.928
Discordant = 3.1%	Gamma = 0.937
Tied = 1.0%	Tau-a = 0.466
(10000 pairs)	c = 0.964

Conditional Odds Ratios and 95% Confidence Intervals

Variable	Unit	Odds Ratio	Wald Confidence Limits Lower	Upper
FAMHX	1.0000	11.394	3.424	37.919
CDSIZE	1.0000	2.350	1.451	3.804

Hosmer and Lemeshow Goodness-of-Fit Test

Group	Total	RRF = 1		RRF = 0	
		Observed	Expected	Observed	Expected
1	1	1	0.02	0	0.98
2	80	1	4.10	79	75.90
3	31	10	10.16	21	20.84
4	19	10	16.72	0	2.28
5	69	69	68.99	0	0.01

Goodness-of-fit Statistic = 48.562 with 3 DF (p=0.0001)

WARNING: In calculating the Expected values, predicted probabilities less than 0.0001 and greater than 0.9999 were changed to 0.0001 and 0.9999 respectively.