VARIOUS WAYS OF PERFORMING LOGISTIC REGRESSION IN *SAS*®

Michael Lerner
Ontario Ministry of Education
September 2010

INTRODUCTON

CATMOD, GENMOD, PROBIT and LOGISTIC perform 'ordinary' logistic regression in SAS STAT.

But even the simplest possible analyses that use discrete predictors can produce different looking results. This presentation discusses why these differences exist and how to produce the same results.

Once multiple outcomes are introduced, the default analyses produced by different procedures are genuinely different. This presentation discusses what these differences are and how to produce the type analysis that you need.

But first a few words about logistic regression.

LOGISTIC REGRESSION

Logistic regression is a statistical technique that estimates the natural base logarithm of the probability of one discrete event (e.g., passing) occurring as opposed to another event (failing) or more other events. The log-odds of the event (broadly referred to as the logit here) are the predicted values.

Exponents of parameters in a logistic regression yield the odds of an event occurring. The probability of an event occurring is equal to the odds divided by the sum of the odds plus 1.

- Odds below 1 mean that there is less than a 50% chance of the event occurring.
- Odds above 1 mean that there is more than a 50% chance of the event occurring.

Odd ratios compare the odds across different values of a predictor.

- When the predictor is discrete such as gender the odds ratio equals the ratio of the odds of the event occurring ('pass') divided by the odds of the event not occurring ('fail') for one gender (female) to the odds of the event occurring divided by the odds of the event not occurring for the other gender (male).

- When the predictor is continuous, the odds ratio is equal to the odds raised to the power of the increment of interest. If there is interest in the effect of a change in age of two years on the odds of passing, then the odds for age should be squared. A 3-year increment requires cubing and so forth.

Odds ratios greater than 1 mean that the event is more relatively likely to occur than not for one group on a discrete predictor as opposed to another or on one setting on a continuous predictor as opposed to another.

DEFAULT DISCRETE PREDICTORS

Consider the following simple data set, example, with a pass/fail outcome, a 3-value discrete predictor, and 114 cases, where there is a 53.5% chance of passing

Outcome	Predictor1	Number
Pass	A	25
Fail	A	12
Pass	B	17
Fail	B	23
Pass	C	19
Fail	C	18

Here are the estimated effects of predictor1 in each procedure for the probability of 'fail':

Estimate	Catmod & Logistic	Genmod & Probit
Intercept	-.1619	-.0541
A	-.5721	.6799
B	+.4642	.3563
C		0

Syntax provided at end of paper.

DEFAULT DISCRETE PREDICTORS

Warning, CATMOD is different:
Assumes that all variables are discrete (no CLASS) and use a DIRECT statement for continuous variables.
Cannot use event/trial syntax.
Requires fixing the margins (include all interaction effects between discrete predictors) if more than 1 discrete predictor is used.
To ensure that 0's are treated as sampling 0's as in LOGISTIC and GENMOD, set the missing keyword in the MODEL options to SAMPLING.

Differences are due to defaults:

In the case of CATMOD and LOGISTIC, the default is effect parameterization. If there is a 3-level discrete predictor, this parameterization estimates the difference in the effect of each nonreference level (A and B in our case; C is the reference) compared to the average effect over all 3 levels.

GENMOD and PROBIT use GLM parameterization by default. If there is a 3-level discrete predictor, this parameterization estimates the difference in the effects of each level compared to the last level.

GLM parameterization has 3 columns of contrasts (parameters) rather than 2 as with effect parameterization (i.e., it is singular since it has more parameters than degrees of freedom). *The intercept is equal to the omitted category. In the case of all other parameterizations discussed here, the omitted category is equal to the sum of the effects.*

The default ordering of discrete variables in these procedures (order=data|freq|formatted|internal) can affect results. *The default 'order=' is formatted in LOGISTIC, PROBIT and GENMOD but is internal in CATMOD. It is always a good idea to specify the option required.*

OPTIONAL EFFECTS FOR DISCRETE PREDICTORS

In GENMOD and LOGISTIC, parameterization of discrete predictors is controlled via the PARAM option of the CLASS statement. Some of the other parameterizations are:

- Reference, estimates the difference in the effect of each nonreference level compared to the effect of the reference level. First or last level (default) can be designated as the reference as is the case for effect parameterization.

- Ordinal thermometer parameterization that estimates the differences between effects of successive levels is available in LOGISTIC and GENMOD. The keyword is ORDINAL. The 1st level is the control or baseline.

EXAMPLES OF CLASS STATEMENT PARAM OPTIONS

Global glm coding (only possibility for glm parameterization)
```
CLASS A B/param=glm;
```

Separate coding of A as reference coding and B as effect coding
```
CLASS A (param=ref) B (param=effect);
```

Separate coding of A as reference coding with 1st level as reference and B as effect coding
```
CLASS A (param=ref ref=first) B (param=effect);
```

In CATMOD, PARAM=REF in the options of the MODEL statement will produce reference category paramaterizaton; the default reference category is the last.

But, PROBIT does not allow any changes to the parameterization of discrete predictors.

DIFFERENT TREATMENTS OF MULTIPLE OUTCOMES

It is possible to have multiple outcomes. The outcomes may have no ordering (nominal level of measurement) such as favourite foods. Outcomes may have ordering (ordinal level of measurement) such as very happy to very unhappy; or none, some or many; or educational attainment.

The default ordering of discrete variables in these procedures (order=data|freq|formatted|internal) can definitely affect results.

Nominal Outcomes:

The generalized logit that compares probability of each outcome to a reference outcome (category) is the only way of looking at unordered outcomes. *They are contrasts between pairs of categories, requiring careful interpretation* (Allison, 1999: 131).

If there is a 3-value outcome and the 3rd outcome is the reference, parameters for two logits (1st vs 3rd and 2nd vs 3rd) are produced.

CATMOD produces generalized logits as the default when there are more than 2 outcomes. The last outcome as defined by 'ORDER=' in the CATMOD statement.

LOGISTIC produces generalized logits when 'link=glogit' is specified as an option in the MODEL statement. Its ordering of the categories can also be specified via
MODEL OUTCOME(REF=FIRST|LAST)=…;

DIFFERENT TREATMENTS OF MULTIPLE OUTCOMES

Ordinal Outcomes:

There 3 basic ways of looking at ordinal outcomes.

1. Cumulative Logit:

In a 3-level *ordered* outcome, these are the 1st (lowest rank outcome) versus the all others; the 1st and 2nd versus the 3rd.

This is the default in PROC LOGISTIC *with the assumption of proportional odds being tested.* A separate intercept for each logit is estimated but all predictors have one common effect. A test for the null hypothesis of a common effect, proportional odds, not being rejected is presented.

The same functional form of cumulative logistic regression is an option in GENMOD by specifying 'link=cumlogit dist=multinomial' in the options portion of the MODEL statement. It is the default in PROBIT but the second and subsequent intercepts are shown as deviations from the first.

CATMOD also produces a different cumulative logit that estimates separate intercepts and slopes through the use of the RESPONSE CLOGLIT statement. Using _RESPONSE_ on the right-hand side of the '='-sign in the MODEL statement will force the estimated model to have a common intercept and effects which is not the same as the proportional odds model.

DIFFERENT TREATMENTS OF MULTIPLE OUTCOMES

2. Adjacent Categories Logit:

In a 3-level ordered outcome, these are the logits for the 1st level versus the 2nd, the 2nd versus the 3rd.

This can be estimated by CATMOD by using the RESPONSE ALOGIT statement. As the model cannot be estimated by maximum likelihood in CATMOD, use weighted least squares by entering METHOD=WLS in the options portion of the MODEL statement.

According to Allison (1999: 149) the data must be in a cross tabulation list output (e.g., the simple example used at the beginning of this presentation). However, experimentation with Allison's data using version 9.22 reveals that this does not seem to be necessary.

Using _RESPONSE_ on the right-hand side of the '='-sign in the model statement will force the estimated model to have a common intercept and effects.

CATMOD also has cumulative (requires WLS) and generalized logit capacity (maximum likelihood).

Maximum likelihood estimates of adjacent logits can be accomplished in GENMOD where the data are Poisson distributed with a log-link function. (Allison, 1999: 250-52).

DIFFERENT TREATMENTS OF MULTIPLE OUTCOMES

3. Continuation Ratio Logit:

In a 3-level outcome each level is thought of as a stage that must be attained before passing onto the next stage. The outcome of interest is whether the progression occurred or not e.g., 1st versus 2nd and 3rd; 2nd versus 3rd).

This can be thought of as a series of binary logits where each logit excludes cases that failed to progress to the preceding stage.

Allison (1999: 151-158) shows how to analyze a data set that incorporates all of these logits using GENMOD where there is a single intercept and set of effects.

It is possible to test whether the single intercept and effects are (constrained model) is suitable by seeing if the estimated parameters vary by stage. The LOGISTIC procedure could be used as well.

UNTANGLING ESTIMATED COEFFICIENTS

Multiple outcome logits can generate many estimated coefficients. Untangling them is important.

In general, the first (lowest subscripted, such as 'intercept_1') coefficient is for the 1st equation, the second for the 2nd equation and so forth in LOGISTIC, PROBIT and GENMOD. In CATMOD, the function number serves as the subscript.

Thus, the 1st intercept refers to the 1st equation, the 2nd to the second equation and so forth.

Thus, the 1st coefficient for the first predictor refers to the 1st equation, the 2nd to the second equation and so forth.

If a common intercept or slope model is being estimated, this does not apply as there will be a set of common parameters.

SELECTING THE 'RIGHT' PROCEDURE

Selecting the right procedure depends on what you want to do and upon what you are most comfortable with in terms of the interpretation of parameters produced by discrete predictors.

For instance, those who are most comfortable with GLM parameterization and wish to have it without having to select it would prefer GENMOD. This provides continuity with GLM. However, if more than a GLM-style parameterization is desired, then GENMOD or LOGISTIC are available.

Selection of the appropriate procedure and options will yield generalized and cumulative logits.

Adjacent category logits require CATMOD or GENMOD. If GENMOD is used the data must be modified and the log-linear analysis of the cross-tabulation through the analysis of Poisson-distributed variables equivalent to the logistic regression of interest must be undertaken.

Continuation ratio logits require a modified data set that can be analyzed through GENMOD or LOGISTIC.

SELECTING THE 'RIGHT' PROCEDURE

Additional considerations that may affect choice of procedure are the particular insights that they provide:

Both GENMOD and LOGISTIC have a 'least square means' statement. GLM parameterization must be used. In version 9.22 an imbalanced design as may exist in the data being analyzed can be selected (OM keyword in the options). In version 9.21 there is no such option and a balanced design (equal number of cases in each cell) is assumed.

LOGISTIC offers enhanced analysis of odds ratio (ODDSRATIO statement) and the UNITS statement. Receiver operator characteristic curves can be analyzed (OUTROC= keyword in the MODEL options) and ODS graphics.

PROBIT provides analyses of natural response rates (i.e., what is the estimated proportion of cases that that 'responded' at various levels of a continuous predictor).

It is quite possible that several different procedures may have to be used in order to gain all of the needed insights required from a 'simple' logistic regression.

REFERENCES

Allison, Paul D. 1999. *Logistic Regression Using the SAS® System: Theory and Application*. Cary, N.C.: SAS Institute Inc.

Friendly, Michael. 2000. *Visualizing Categorical Data.* Cary, N.C.: SAS Institute Inc.

SAS Instutute. 2010. *SAS/STAT 9.22 User's Guide.* Cary, N.C.: SAS Institute Inc. (http://support.sas.com/documentation/onlinedoc/stat/indexproc.html#stat922, accessed August 3-5, 2010).

SAMPLE SIMPLE CODE FOR CATMOD, LOGISTIC AND GENMOD

```
data example;
input Outcome $ Predictor1 $ Number;
datalines ;
Pass  A     25
Fail  A     12
Pass  B     17
Fail  B     23
Pass  C     19
Fail  C     18
;
*DEFAULT CATMOD;
PROC CATMOD DATA=EXAMPLE;
MODEL OUTCOME=PREDICTOR1;
WEIGHT NUMBER;
RUN;
*DEFAULT LOGISTIC;
PROC LOGISTIC DATA=EXAMPLE;
CLASS OUTCOME PREDICTOR1;
MODEL OUTCOME=PREDICTOR1;
FREQ NUMBER;
RUN;
*DEFAULT GENMOD;
PROC GENMOD DATA=EXAMPLE;
CLASS OUTCOME PREDICTOR1;
MODEL OUTCOME=PREDICTOR1/DIST=BIN LINK=LOGIT;
FREQ NUMBER;
RUN;
*DEFAULT PROBIT;
PROC PROBIT DATA=EXAMPLE;
CLASS OUTCOME PREDICTOR1;
MODEL OUTCOME=PREDICTOR1/DIST=LOGISTIC;
WEIGHT NUMBER;
RUN;
```