A Tutorial on Logistic Regression

Ying So, SAS Institute Inc., Cary, NC

ABSTRACT

Many procedures in SAS/STAT® can be used to perform logistic regression analysis: CATMOD, GENMOD,LOGISTIC, and PROBIT. Each procedure has special features that make it useful for certain applications. For most applications, PROC LOGISTIC is the preferred choice. It fits binary response or proportional odds models, provides various model-selection methods to identify important prognostic variables from a large number of candidate variables, and computes regression diagnostic statistics. This tutorial discusses some of the problems users encountered when they used the LOGISTIC procedure.

INTRODUCTION

PROC LOGISTIC can be used to analyze binary response as well as ordinal response data.

Binary Response

The response, Y, of a subject can take one of two possible values, denoted by 1 and 2 (for example, Y=1 if a disease is present; otherwise Y=2). Let $\mathbf{x} = (x_1, \ldots, x_k)'$ be the vector of explanatory variables. The logistic regression model is used to explain the effects of the explanatory variables on the binary response.

$$\text{logit}\{\Pr(Y = 1|\mathbf{x})\} = \log\left\{\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})}\right\} = \beta_0 + \mathbf{x}'\boldsymbol{\beta}$$

where β_0 is the intercept parameter, and $\boldsymbol{\beta}$ is the vector of slope parameters (Hosmer and Lameshow, 1989).

Ordinal Response

The response, Y, of a subject can take one of m ordinal values, denoted by $1, 2, \ldots, m$. PROC LOGISTIC fits the following cumulative logit* model:

$$\text{logit}\{\Pr(Y \le r|\mathbf{x})\} = \alpha_r + \mathbf{x}'\boldsymbol{\beta} \qquad 1 \le r < m$$

where $\alpha_1, \ldots, \alpha_{m-1}$ are (m-1) intercept parameters. This model is also called the proportional odds model because the odds of making response $\le r$ are exp($\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)$) times higher at $\mathbf{x} = \mathbf{x}_1$ than at $\mathbf{x} = \mathbf{x}_2$ (Agresti, 1990).

This ordinal model is especially appropriate if the ordinal nature of the response is due to methodological limitations in collecting the data in which the researchers are forced to lump together and identify various portions of an otherwise continuous variable. Let T be the underlying continuous variable and suppose that

$$Y = r \quad \text{if} \quad \gamma_{r-1} < T \le \gamma_r$$

for some $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_m = \infty$. Let x_0 = 1. Consider the regression model

$$T = \sum_{i=0}^{k} \beta_i^* x_i + e$$

where $\beta_0^*, \beta_1^*, \ldots, \beta_{m-1}^*$ are regression parameters and e is the error term with a logistic distribution F. Then

$$\Pr(Y \le r) = \Pr(T \le \gamma_r) = F(\gamma_r - \sum_{i=0}^{k} \beta_i^* x_i)$$

or

$$\text{logit}\{\Pr(Y \le r|\mathbf{x})\} = \gamma_r - \sum_{i=0}^{k} \beta_i^* x_i$$

This is equivalent to the proportional odds model given earlier.

INFINITE PARAMETERS

The term *infinite parameters* refers to the situation when the likelihood equation does not have a finite solution (or in other words, the maximum likelihood estimate does not exist). The existence of maximum likelihood estimates for the logistic model depends on the configurations of the sample points in the observation space (Albert and Anderson, 1984, and Santner and Duffy, 1985). There are three mutually exclusive and exhaustive categories: complete separation, quasicomplete separation, and overlap.

Consider a binary response model. Let Y_i be the response of the i^{th} subject and let $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ be the vector of explanatory variables (including the constant 1).

Complete Separation

There is a complete separation of data points if there exists a vector $\mathbf{b}$ that correctly allocates all observations to their response groups; that is,

$$\begin{cases} \mathbf{b}'\mathbf{x}_i > 0 & Y_i = 1 \\ \mathbf{b}'\mathbf{x}_i < 0 & Y_i = 2 \end{cases}$$

The maximum likelihood estimate does not exist. The loglikelihood goes to 0 as iteration increases.

*logit of the cumulative probabilities

The following example illustrates such situation. Consider the data set DATA1 (Table 1) with 10 observations. Y is the response and x_1 and x_2 are two explanatory variables.

Table 1. Complete Separation Data (DATA1)

Observation	Y	x_1	x_2
1	1	29	62
2	1	30	83
3	1	31	74
4	1	31	88
5	1	32	68
6	2	29	41
7	2	30	44
8	2	31	21
9	2	32	50
10	2	33	33

Figure 1 shows that the vector $\mathbf{b} = (6, -2, 1)'$ completely separates the observations into their response groups; that is, all observations of the same response lie on the same side of the line $x_2 = 2x_1 - 6$.

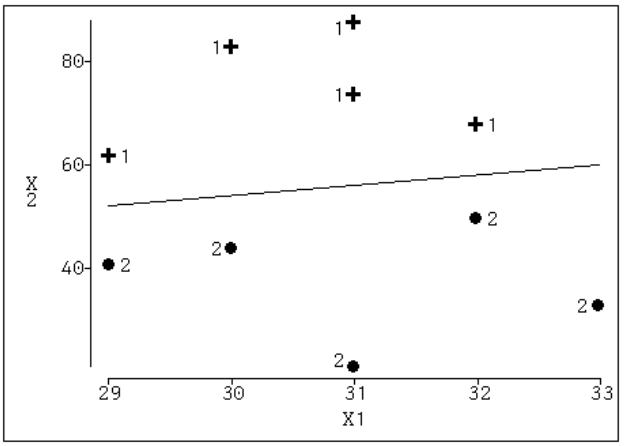

Figure 1. Scatterplot of Sample Points in DATA1

The iterative history of fitting a logistic regression model to the given data is shown in Output 1. Note that the negative loglikehood decreases to 0 --- a perfect fit.

Quasicomplete Separation

If the data are not completely separated and there exists a vector $\mathbf{b}$ such that

$$\begin{cases} \mathbf{b}'\mathbf{x}_i \geq 0 & Y_i = 1 \\ \mathbf{b}'\mathbf{x}_i \leq 0 & Y_i = 2 \end{cases}$$

with equality holds for at least one subject in each response group, there is a quasicomplete separation. The maximum likelihood estimate does not exist. The loglikelihood does not diminish to 0 as in the case of complete separation, but the dispersion matrix becomes unbound.

Output 1. Partial LOGISTIC Printout for DATA1

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCPT	X1	X2
0	INITIAL	13.862944	0	0	0
1	IRLS	4.691312	-2.813220	-0.062042	0.083761
2	IRLS	2.280691	-2.773158	-0.187259	0.150942
3	IRLS	0.964403	-0.425345	-0.423977	0.238202
4	IRLS	0.361717	2.114730	-0.692202	0.339763
5	IRLS	0.133505	4.250789	-0.950753	0.443518
6	IRLS	0.049378	6.201510	-1.203505	0.547490
7	IRLS	0.018287	8.079876	-1.454499	0.651812
8	IRLS	0.006774	9.925139	-1.705284	0.756610
9	IRLS	0.002509	11.748893	-1.956323	0.861916
10	IRLS	0.000929	13.552650	-2.207666	0.967727
11	IRLS	0.000344	15.334133	-2.459215	1.074024
12	IRLS	0.000127	17.089516	-2.710811	1.180784
13	IRLS	0.000047030	18.814237	-2.962266	1.287983
14	IRLS	0.000017384	20.503310	-3.213375	1.395594
15	IRLS	0.000006423	22.151492	-3.463924	1.503590
16	IRLS	0.000002372	23.753408	-3.713693	1.611943
17	IRLS	0.000000876	25.303703	-3.962463	1.720626
18	IRLS	0.000000323	26.797224	-4.210021	1.829610
19	IRLS	0.000000119	28.229241	-4.456170	1.938869
20	IRLS	4.3956397E-8	29.595692	-4.700735	2.048377
21	IRLS	1.620409E-8	30.893457	-4.943572	2.158109
22	IRLS	5.9717453E-9	32.120599	-5.184576	2.268042
23	IRLS	2.2002107E-9	33.276570	-5.423689	2.378153
24	IRLS	8.10449E-10	34.362317	-5.660901	2.488421
25	IRLS	2.984679E-10	35.380281	-5.896252	2.598826

WARNING: Convergence was not attained in 25 iterations.
Iteration control is available with the MAXITER and the
CONVERGE options on the MODEL statement.

You can modify DATA1 to create a situation of quasicomplete separation, for instance, change $x_2 = 44$ to $x_2 = 64$ in observation 6. Let the modified data set be DATA2. With $\mathbf{b} = (-4, -2, 1)'$, the equality holds for observations 1, 5, and 7, and the rest of the observations are separated into their response groups (Figure 2). It is easy to see that there is no straight line that can completely separate the two response groups.

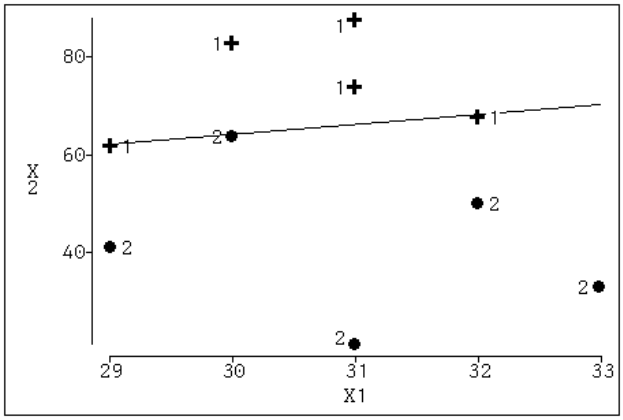

Figure 2. Scatterplot of Sample Points in DATA2

The parameter estimates during the iterative phase are displayed in Output 2 and the dispersion matrices for iterations 0, 5, 10, 15, and 25 are shown in Output 3. The log-likelihood approaches a nonzero constant . The seemingly large variances of pseudoestimates are typical of a quasicomplete separation of data.

Output 2. Partial LOGISTIC Printout for DATA2

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCPT	X1	X2
0	INITIAL	13.862944	0	0	0
1	IRLS	6.428374	-4.638506	0.003387	0.077640
2	IRLS	4.856439	-7.539932	-0.011060	0.131865
3	IRLS	4.190154	-9.533783	-0.066638	0.190242
4	IRLS	3.912968	-11.081432	-0.146953	0.252400
5	IRLS	3.800380	-11.670780	-0.265281	0.316912
6	IRLS	3.751126	-11.666819	-0.417929	0.388135
7	IRLS	3.727865	-11.697310	-0.597641	0.472639
8	IRLS	3.716764	-11.923095	-0.806371	0.573891
9	IRLS	3.711850	-12.316216	-1.038254	0.688687
10	IRLS	3.709877	-12.788230	-1.281868	0.810247
11	IRLS	3.709130	-13.282112	-1.529890	0.934224
12	IRLS	3.708852	-13.780722	-1.779320	1.058935
13	IRLS	3.708750	-14.280378	-2.029162	1.183855
14	IRLS	3.708712	-14.780288	-2.279118	1.308833
15	IRLS	3.708698	-15.280265	-2.529107	1.433827
16	IRLS	3.708693	-15.780258	-2.779104	1.558826
17	IRLS	3.708691	-16.280257	-3.029103	1.683825
18	IRLS	3.708691	-16.780256	-3.279103	1.808825
19	IRLS	3.708690	-17.280256	-3.529103	1.933825
20	IRLS	3.708690	-17.780256	-3.779103	2.058825
21	IRLS	3.708690	-18.280256	-4.029102	2.183825
22	IRLS	3.708690	-18.780255	-4.279102	2.308825
23	IRLS	3.708690	-19.280256	-4.529102	2.433825
24	IRLS	3.708690	-19.780257	-4.779103	2.558825
25	IRLS	3.708690	-20.280250	-5.029099	2.683824

WARNING: Convergence was not attained in 25 iterations.
Iteration control is available with the MAXITER and the CONVERGE options on the MODEL statement.

Output 3. Dispersion Matrices on Selected Iterations (DATA2)

Iter= 0 -2 Log L = 13.862944

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	269.05188212	-8.42405441	-0.157380245	0
Z1	-8.42405441	0.2673239797	0.0032615725	0
Z2	-0.157380245	0.0032615725	0.0009747228	0

Iter=5 -2 Log L = 3.800380

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	985.12006548	-29.47104673	-1.460819309	-11.670780
Z1	-29.47104673	1.4922999204	-0.242120428	-0.265281
Z2	-1.460819309	-0.242120428	0.1363093424	0.316912

Iter= 10 -2 Log L = 3.709877

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	1391.583624	169.160036	-100.9268654	-12.788230
Z1	169.160036	105.7305273	-52.20138038	-1.281868
Z2	-100.9268654	-52.20138038	26.043666498	0.810247

Iter= 15 -2 Log L = 3.708698

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	62940.299541	30943.762505	-15488.22021	-15.280265
Z1	30943.762505	15493.136539	-7745.900995	-2.529107
Z2	-15488.22021	-7745.900995	3872.8917604	1.433827

Iter=20 -2 Log L = 3.708690

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	9197536.1382	4598241.6822	-2299137.18	-17.780256
Z1	4598241.6822	2299142.0966	-1149570.381	-3.779103
Z2	-2299137.18	-1149570.381	574785.13177	2.058825

Iter=25 -2 Log L = 3.708690

Variable	INTERCPT	Z1	Z2	ESTIMATE
INTERCPT	502111231.75	251055089.49	-125527561.1	-20.280250
Z1	251055089.49	125527566	-62763782.33	-5.029099
Z2	-125527561.1	-62763782.33	31381891.107	2.683824

Overlap

If neither complete nor quasicomplete separation exists in the sample points, there is an overlap of sample points. The maximum likelihood estimate exists and is unique.

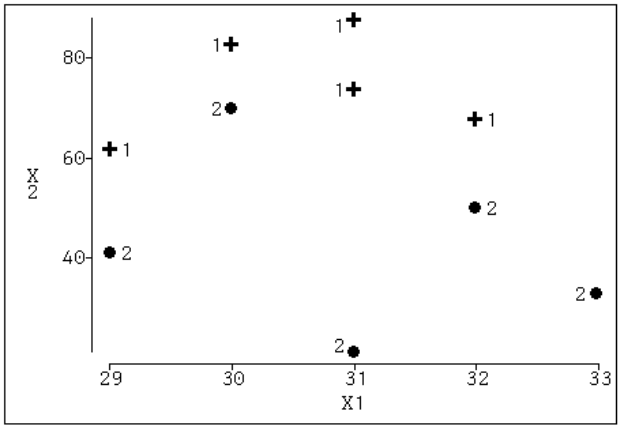

Figure 3. Scatterplot of Sample Points in DATA3

If you change $x_2 = 44$ to $x_2 = 74$ in observation 6 of DATA1, the modified data set (DATA3) has overlapped sample points. A scatterplot of the sample points in DATA3 is shown in Figure 3. For every straight line on the drawn on the plot, there is always a sample point from each response group on same side of the line. The maximum likelihood estimates are finite (Output 4).

Output 4. PROC LOGISTIC Printout for DATA3

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCPT	X1	X2
0	INITIAL	13.862944	0	0	0
1	IRLS	7.192759	-4.665775	0.011192	0.073238
2	IRLS	6.110729	-7.383116	0.010621	0.116549
3	IRLS	5.847544	-8.760124	-0.013538	0.148942
4	IRLS	5.816454	-9.185086	-0.033276	0.164399
5	IRLS	5.815754	-9.228343	-0.037848	0.167125
6	IRLS	5.815754	-9.228973	-0.037987	0.167197
7	IRLS	5.815754	-9.228973	-0.037987	0.167197

Last Change in -2 Log L: 2.282619E-13

Last Evaluation of Gradient

INTERCPT	X1	X2
-1.109604E-7	-3.519319E-6	-3.163568E-6

Empirical Approach to Detect Separation

Complete separation and quasicomplete separation are problems typical for small sample. Although complete separation can occur with any type of data, quasicomplete separation is not likely with truly continuous data.

At the j^{th} iteration, let $\mathbf{b}^j$ be the vector of pseudoestimates. The probability of correct allocation based on $\mathbf{b}^j$ is given by

$$\begin{cases} \frac{\exp(\mathbf{x}'\mathbf{b}^j)}{1+\exp(\mathbf{x}'\mathbf{b}^j)} & Y = 1 \\ \\ \frac{1}{1+\exp(\mathbf{x}'\mathbf{b}^j)} & Y = 2 \end{cases}$$

- Stop at the iteration when the probability of correct allocation is 1 for all observations. There is a complete separation of data points. For DATA1, correct allocation of all data points is achieved at iteration 13 (Table 2).
- At each iteration, look for the observation with the largest probability of correct allocation. If this probability has become extremely close to 1, and any diagonal element of the dispersion matrix becomes very large, stop the iteration. It is very likely there is a quasicomplete separation in the data set. Table 3 displays the maximum probability of correct allocation for DATA2. The dispersion matrix should be examined after the 5^{th} iteration.

Table 2. Percentage of Correct Allocation (DATA1)

j	-2 Log L	b_0^j	b_1^j	b_2^j	% of Correct Allocation
0	13.8629	0.0000	0.00000	0.00000	0
1	4.6913	-2.8132	-0.06204	0.08376	0
2	2.2807	-2.7732	-0.18726	0.15094	0
3	0.9644	-0.4253	-0.42398	0.23820	0
4	0.3617	2.1147	-0.69220	0.33976	10
5	0.1335	4.2508	-0.95075	0.44352	40
6	0.0494	6.2015	-1.20351	0.54749	40
7	0.0183	8.0799	-1.45450	0.65181	40
8	0.0068	9.9251	-1.70528	0.75661	50
9	0.0025	11.7489	-1.95632	0.86192	50
10	0.0009	13.5527	-2.20767	0.96773	50
11	0.0003	15.3341	-2.45922	1.07402	60
12	0.0001	17.0895	-2.71081	1.18078	80
13	0.0000	18.8142	-2.96227	1.28798	100
14	0.0000	20.5033	-3.21338	1.39559	100
15	0.0000	22.1515	-3.46392	1.50359	100

Table 3. Maximum Probability of Correct Allocation (DATA2)

j	-2 Log L	b_0^j	b_1^j	b_2^j	Maximum Probability
0	13.8629	0.0000	0.00000	0.00000	0.50000
1	6.4284	-4.6385	0.00339	0.07764	0.87703
2	4.8564	-7.5399	-0.01106	0.13187	0.97217
3	4.1902	-9.5338	-0.06664	0.19024	0.99574
4	3.9130	-11.0814	-0.14695	0.25240	0.99950
5	3.8004	-11.6708	-0.26528	0.31691	0.99995
6	3.7511	-11.6668	-0.41793	0.38814	1.00000
7	3.7279	-11.6973	-0.59764	0.47264	1.00000
8	3.7168	-11.9231	-0.80637	0.57389	1.00000
9	3.7119	-12.3162	-1.03825	0.68869	1.00000
10	3.7099	-12.7882	-1.28187	0.81025	1.00000
11	3.7091	-13.2821	-1.52989	0.93422	1.00000
12	3.7089	-13.7807	-1.77932	1.05894	1.00000
13	3.7088	-14.2804	-2.02916	1.18386	1.00000
14	3.7087	-14.7803	-2.27912	1.30883	1.00000
15	3.7087	-15.2803	-2.52911	1.43383	1.00000

ORDERING OF THE BINARY RESPONSE LEVELS

If the binary response is 0 and 1, PROC LOGISTIC, by default, models the probability of 0 instead of 1; that is,

$$\log\left\{\frac{\Pr(Y=0|\mathbf{x})}{\Pr(Y=1|\mathbf{x})}\right\} = \beta_0 + \mathbf{x}'\boldsymbol{\beta}$$

This is consistent with the cumulative logit model, though this may not always be desirable because 1 is often used to denote the response of the event of interest. Consider the following logistic regression example. Y is the response variable with value 1 if the disease is present and 0 otherwise. EXPOSURE is the only explanatory variable with value 1 if the subject is exposed and 0 otherwise.

```
data disease;
   input y exposure freq;
   cards;
   1 0 10
   1 1 40
   0 0 45
   0 1  5
   ;
   run;
proc logistic data=disease;
  model y=exposure;
  freq freq;
  run;
```

Output 5. Logistic Regression of Disease on Exposure

```
                    Response Profile

                Ordered
                  Value       Y     Count

                      1       0        50
                      2       1        50



              Criteria for Assessing Model Fit

                          Intercept
                Intercept     and
Criterion         Only    Covariates Chi-Square for Covariates

AIC              140.629     87.550      .
SC               143.235     92.761      .
-2 LOG L         138.629     83.550    55.079 with 1 DF (p=0.0001)
Score               .          .       49.495 with 1 DF (p=0.0001)


           Analysis of Maximum Likelihood Estimates

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

INTERCPT   1          1.5041     0.3496      18.5093       0.0001
EXPOSURE   1         -3.5835     0.5893      36.9839       0.0001

        Analysis of Maximum
       Likelihood Estimates

              Standardized        Odds
Variable        Estimate          Ratio

INTERCPT                .         4.500
EXPOSURE         -0.987849        0.028
```

Results of the analysis are displayed in Output 5. Since the coefficient for EXPOSURE is negative, as EXPOSURE

changes from 0 to 1, the probability of "no disease" decreases. This is a less direct way of saying that the probability of "disease" increases with EXPOSURE.

Since

$$\log\left\{\frac{\Pr(Y=1|\mathbf{x})}{\Pr(Y=0|\mathbf{x})}\right\} = -\log\left\{\frac{\Pr(Y=0|\mathbf{x})}{\Pr(Y=1|\mathbf{x})}\right\}$$

the probability of response 1 is given by

$$\log\left\{\frac{\Pr(Y=1|\mathbf{x})}{1-\Pr(Y=1|\mathbf{x})}\right\} = -\beta_0 - \mathbf{x}'\boldsymbol{\beta}$$

That is, the regression coefficients for modeling the probability of 1 will have the same magnitude but opposite sign as those of modeling the probability of 0. In order to have a more direct interpretation of the regression coefficient, it is desirable to model the probability of the event of interest. In the LOGISTIC procedure, the response levels are sorted according to the ORDER= option (the Response Profiles table lists the ordering of the responses). PROC LOGISTIC then models the probability of the response that corresponds to the lower ordered value.

Note that the first observation in the given input data has response 1. By using the option ORDER=DATA, the response 1 will have ordered value 1 and response 0 will have ordered value 2. As such the probability modeled is the probability of response 1. There are several other ways that you can reverse the response level ordering in the given example (Schlotzhauer, 1993).

- The simplest method, available in Release 6.07 TS301 and later, uses the option DESCENDING. Specify the DESCENDING option on the PROC LOGISTIC statement to reverse the ordering of Y.

```
proc logistic data=disease descending;
   model y=exposure;
   freq freq;
   run;
```

- Assign a format to Y such that the first formatted value (when the formatted values are put in sorted order) corresponds to the presence of the disease. For this example, Y=0 could be assigned the formatted value 'no disease' and Y=1 could be assigned the formatted value 'disease'.

```
proc format;
   value disfmt 1='disease'
                0='no disease';
   run;
proc logistic data=disease;
   model y=exposure;
   freq freq;
   format y disfmt.;
   run;
```

- Create a new variable to replace Y as the response variable in the MODEL statement such that observation Y=1 takes on the smaller value of the new variable.

```
data disease2;
   set disease;
   if y=0 then y1='no disease';
   else 'disease';
   run;
proc logistic data=disease2;
   model y1=exposure;
   freq freq;
   run;
```

- Create a new variable (N, for example) with constant value 1 for each observation. Use the *event/trial* MODEL statement syntax with Y as the *event* variable and N as the *trial* variable.

```
data disease3;
   set disease;
   n=1;
   run;
proc logistic data=disease;
   model y/n=exposure;
   freq freq;
   run;
```

OTHER LOGISTIC REGRESSION APPLICATIONS

There are many logistic regression models that are not of the standard form as given earlier (Agresti, 1990, and Strauss, 1992). For some of them you could "trick" PROC LOGISTIC to do the estimation, for others you may have to resort to other means. The following sections discuss some of the models that are often inquired by SAS users.

Conditional Logistic Regression

Conditional logistic regression is useful in investigating the relationship between an outcome and a set of prognostic factors in a matched case-control studies, the outcome being whether the subject is a case or a control. When there is one case and one control in a matched set, the matching is 1:1. 1:n matching refers to the situation when there is one case and a varying number of controls in a matched set. For the i^{th} set, let $\mathbf{u}_i$ the covariate vector for the case and let $\mathbf{v}_{i1}, \ldots, \mathbf{v}_{in_i}$ be the covariate vectors for the n_i controls. The likelihood for the *N* matched sets is given by

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{\exp(\mathbf{u}_i'\boldsymbol{\beta})}{\sum_{j=1}^{n_i}\exp(\mathbf{v}_{ij}'\boldsymbol{\beta})}$$

For the 1-1 matching, the likelihood is reduced to

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{\exp(\mathbf{u}_i'\boldsymbol{\beta})}{\exp(\mathbf{u}_i'\boldsymbol{\beta}) + \exp(\mathbf{v}_{i1}'\boldsymbol{\beta})}$$

By dividing the numerator and the denominator by $\exp(\mathbf{v}_{i1}'\beta)$, one obtains

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{\exp((\mathbf{u}_i - \mathbf{v}_{i1})'\boldsymbol{\beta})}{1 + \exp((\mathbf{u}_i - \mathbf{v}_{i1})'\boldsymbol{\beta})}$$

Thus the likelihood is identical to that of the binary logistic model with $\mathbf{d}_i = \mathbf{u}_i - \mathbf{v}_{i1}$ as covariates, no intercept, and a constant response. Therefore, you can "trick" PROC LOGISTIC to perform the conditional logistic regression for 1-1 matching (See Example 5 of the LOGISTIC documentation). For 1:n matching, it is more convenient to use PROC PHREG (see Example 3 of the PHREG documentation).

Bradley-Terry Model for Paired Comparison

The Bradley-Terry Model is useful in establishing the overall ranking of n items through paired comparisons. For instance, it is difficult for a panelist to rate all 9 brands of beer at the same occasion; rather it is preferable to compare the brands in a pairwise manner. For a given pair of products, the panelist would state his preference after tasting them at the same occasion. Let $\beta_1, \beta_2, \ldots, \beta_n$ be regression coefficients associated with the n items $I_1, \ldots, I_n$, respectively. The probability that I_i is preferred to I_j is

$$\begin{aligned} \pi_{ij} &= \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)} \\ &= \frac{\exp(\beta_i - \beta_j)}{1 + \exp(\beta_i - \beta_j)} \end{aligned}$$

and, therefore, the likelihood function for the paired comparison model is

$$\mathcal{L}(\beta_1, \ldots, \beta_n) = \prod_{(i,j) \in A} \pi_{ij}$$

where A is the sample collection of all the test pairs. For the l^{th} pair of comparison, if I_i is preferable to I_j, let the vector $\mathbf{d}_l = (d_{l1}, \ldots, d_{ln})$ be such that

$$d_{lk} = \begin{cases} 1 & k = i \\ -1 & k = j \\ 0 & \text{otherwise} \end{cases}$$

The likelihood for the Bradley-Terry model is identical to the binary logistic model with $\mathbf{d}_l$ as covariates, no intercept, and a constant response.

Multinormial Logit Choice Model

The multinormial logit model is useful in investigating consumer choice behavior and has become increasingly popular in marketing research. Let C be a set of n choices, denoted by $\{1, 2, \ldots, n\}$. A subject is present with alternatives in C and is asked to choose the most preferred alternative. Let $\mathbf{x}_i$ be a covariate vector associated with the alternative i. The multinomial logit model for the choice probabilities is given by

$$\Pr(i|\mathrm{C}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{j=1}^{n} \exp(\mathbf{x}_j'\boldsymbol{\beta})}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression parameters. It is difficult to use PROC LOGISTIC to fit such a model. Instead, by defining a proper time and a proper censoring variable, you can trick PROC PHREG to provide the maximum likelihood estimates of the parameters. For details on using PROC PHREG to analyse discrete choice studies, write to Warren Kuhfeld at SAS Institute Inc. (email: saswfk@unx.sas.com) for a copy of the article "Multinormial Logit, Discrete Choice Model."

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*. Wiley, New York.

Albert A. and Anderson, J.A. (1984), "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika*, **71**, pp. 1-10.

Hosmer, D.W., Jr. and Lameshow, S. (1989), *Applied Logistic Regression*. Wiley, New York.

Santner T.J. and Duffy, E.D. (1986), "A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models." *Biometrika*, **73**, pp. 755-758.

SAS Institute Inc. (1990), *SAS/STAT User's Guide*, Vol. 1 & 2, Version 6, Fourth Edition, Cary, NC. (The CATMOD, LOGISTIC, PROBIT procedures.)

SAS Institute Inc. (1992), SAS Technical Report P-229. SAS/STAT Software: Changes and Enhancements. Cary, NC. (The PHREG Procedure.)

SAS Institute Inc. (1993), SAS Technical Report P-243. SAS/STAT Software: The GENMOD Procedure. Cary, NC.

Schlotzhauer, D.C (1993), "Some issues in using PROC LOGISTIC for binary logistic regression". *Observations: The Technical Journal for SAS Software Users.* Vol. 2, No. 4.

Strauss, D. (1992), "The many faces of logistic regression." *The American Statistician*, Vol. 46, No. 4, pp. 321-326.