Paper P254-25

Performing Exact Logistic Regression with the SAS® System

Robert E. Derr, SAS Institute Inc., Cary, NC

ABSTRACT

Exact logistic regression has become an important analytical technique, especially in the pharmaceutical industry, since the usual asymptotic methods for analyzing small, skewed, or sparse data sets are unreliable. Inference based on enumerating the exact distributions of sufficient statistics for parameters of interest in a logistic regression model, conditional on the remaining parameters, is computationally infeasible for many problems. Hirji, Mehta, and Patel (1987) developed an efficient algorithm for generating the required conditional distributions, thus making these methods computationally available. This paper discusses the theory and methods for exact logistic regression and illustrates their application in Version 8 of the SAS® System with new facilities in the LOGISTIC procedure.

INTRODUCTION

Many clinical trials deal with the comparison of populations of subjects with categorical responses. Historically, statistical inference for such studies involve large-sample approximations, and fitting logistic regression models to such data is performed through the unconditional likelihood function. However, asymptotic methods may be inadequate when sample sizes are small or the data are sparse, skewed, or heavily tied. Exact conditional inference remains valid in such situations.

The LOGISTIC, GENMOD, PROBIT, and CATMOD procedures perform unconditional likelihood inference for logit models, and the PHREG procedure can perform asymptotic conditional likelihood inference for logit models. SAS users have requested the ability to perform exact tests for logistic regression modeling. Many exact statistical tests have already been added to the FREQ and NPAR1WAY procedures, and in Release 8.1, SAS/STAT® software includes exact logistic regression for binary (dichotomous) response variables in the LOGISTIC procedure.

The "METHODOLOGY" section in this paper presents the logistic regression model and the different likelihoods, then explains how the exact analysis algorithm implemented in PROC LOGISTIC works; details on the reported statistics are available in the appendix. The "SYNTAX" section describes the new statements and options in the LOGISTIC procedure for the exact methods. The "EXAMPLES" section provides several examples to illustrate the syntax and the usefulness of the method.

Dose-Response Study

First, consider a small dose-response study to motivate the usefulness of exact logistic regression. Researchers are interested in analyzing how mortality rates change with respect to dosage of a drug. The dose data set contains life/death outcomes for six levels of drug dosage (0 to 5). Three subjects are given each specific dose of the drug, and the number of deaths are recorded.

```
data dose;
   input Dose Deaths Total @@;
   datalines;
0 0 3   1 0 3   2 0 3   3 0 3
4 1 3   5 2 3
;
run;
```

All of the cells have counts that are less than 5, which makes the applicability of large sample theory questionable. For each subject i receiving dosage x_i, $i = 1, ..., 18$, let $Y_i = 1$ if the subject died, $Y_i = 0$ otherwise, and $\pi_i = \Pr(Y_i = 1|x_i)$. Then the linear logistic model for this problem is $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + x_i\beta$, which fits a common intercept and slope for the i subjects. In the PROC LOGISTIC invocation below, the EXACT statement requests an exact analysis and the ESTIMATE option produces exact parameter estimates.

```
proc logistic data=dose descending;
   model Deaths/Total = Dose;
   exact Dose / estimate=both;
run;
```

Figure 1 displays some of the unconditional asymptotic results that are produced by default. The likelihood ratio and score tests reject the null hypothesis that β is zero. However, the Wald test does not reject this null hypothesis. The seemingly conflicting conclusions of these tests are a telltale sign that the large-sample approximation is unreliable. The estimates for the intercept α and the slope β both have p-values greater than 0.05, indicating marginal influence. The confidence limits for the odds ratio of the dose parameter contains the value 1, from which you could conclude, if you accept the model, that there is no change in mortality with a change in dosage.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.1478	1	0.0043
Score	5.7943	1	0.0161
Wald	2.7249	1	0.0988

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-9.4745	5.5677	2.8958	0.0888
Dose	1	2.0804	1.2603	2.7249	0.0988

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dose	8.007	0.677	94.679

Figure 1. Output from Asymptotic Analysis

Figure 2 shows the results from the EXACT statement. The p-values in the "Conditional Exact Tests" table lead to rejecting the null hypothesis that β is zero (no conclusions can be made about α since it is "conditioned" away). Note that the p-values for the asymptotic estimates are larger than those for the exact estimates; however, Stokes, Davis, and Koch (1995) observe that, in general, the exact methods tend to produce more conservative results. The "Exact Parameter Estimates" table shows that the slope β is estimated to be $\widehat{\beta} = 1.8$, and since the 95% confidence interval for the odds ratio of $\widehat{\beta}$ does not contain 1, the odds of death increase significantly with dosage. Note that the exact tests do not produce standard errors for the estimates.

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Dose	Score	5.4724	0.0245	0.0190
	Probability	0.0110	0.0245	0.0190

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	1.8000	0.1157	5.8665	0.0245

Figure 2. Output from EXACT Analysis

Exact Conditional Analysis

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	6.049	1.123	353.000	0.0245

Figure 2. (continued)

The unconditional asymptotic and conditional exact results produce somewhat conflicting conclusions for this example. Stokes, Davis, and Koch (1995) recommend looking at the exact results when sample sizes are small and the approximate p-values are less than 0.10. For this example, the small sample size and the conflicting results for the asymptotic hypothesis tests indicate that an exact analysis would be more appropriate.

METHODOLOGY

The theory of exact conditional logistic regression analysis was originally laid out by Cox (1970), and the computational method employed in PROC LOGISTIC is described in Hirji, Mehta, and Patel (1987). Other references that provide useful summaries of the derivations include Cox and Snell (1989), Agresti (1990), and Mehta and Patel (1995).

This section summarizes the methodology behind logistic regression and explains how the algorithm for exact computations works.

Logistic Regression

Consider n independent Bernoulli random variables $Y_1, ..., Y_n$ having observed values $\boldsymbol{y}_0 = (y_{01}, ..., y_{0n})'$. For each observation $i = 1, ..., n$, let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip}, x_{i,p+1}, ..., x_{i,p+q})'$ be a $p + q$ vector of explanatory variables, and denote $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$. Let $\pi_i = \pi(\boldsymbol{x}_i) = \Pr(Y_i = 1|\boldsymbol{x}_i)$ be the event probability for each $i = 1, ..., n$, and denote $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)'$. Then the logistic regression model is $\mathrm{logit}(\boldsymbol{\pi}) = \boldsymbol{X\beta}$, or

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p+q})'$ is the unknown parameter vector.

The joint probability of the observed $\boldsymbol{y}_0$ is a product of n Bernoulli functions:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \pi_i^{y_{0,i}} (1-\pi_i)^{1-y_{0,i}} \\ &= \frac{\exp(\boldsymbol{y}_0'\boldsymbol{X\beta})}{\prod_{i=1}^{n}[1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})]} \end{aligned}$$

Unconditional likelihood inference is based on maximizing this likelihood function, and several asymptotic statistics (likelihood ratio, score, and Wald) can be used to perform hypothesis tests.

To perform conditional inference, first observe that the sufficient statistics for the β_j in the unconditional likelihood function are the corresponding $T_j = \sum_{i=1}^{n} y_i x_{ij}$, where y_i is a realization of Y_i. To create the probability density function (pdf) for $\boldsymbol{T} = (T_1, ..., T_{p+q})'$, sum over all binary sequences $\boldsymbol{y}$ that generate an observable $\boldsymbol{t}$

$$\Pr(\boldsymbol{T} = \boldsymbol{t}) = \frac{C(\boldsymbol{t}) \exp(\boldsymbol{t}'\boldsymbol{\beta})}{\prod_{i=1}^{n} [1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})]}$$

where $C(\boldsymbol{t}) = ||\{\boldsymbol{y} : \boldsymbol{y}'\boldsymbol{X} = \boldsymbol{t}'\}||$ is the number of sequences $\boldsymbol{y}$ that generate $\boldsymbol{t}$. Suppose the p parameters $\boldsymbol{\beta}_0 = (\beta_1, ..., \beta_p)'$ are *nuisance* parameters; that is, the current analysis is geared toward the last q parameters β_1. Denote the sufficient statistics for the nuisance parameters as $\boldsymbol{T}_0 = (T_1, ..., T_p)$, the corresponding observed values as $\boldsymbol{t}_0$, and the corresponding columns of $\boldsymbol{X}$ as $\boldsymbol{X}_0$. Similarly, define $\boldsymbol{T}_1$, $\boldsymbol{t}_1$, and $\boldsymbol{X}_1$ for the parameters of interest. The nuisance parameters can be removed from the analysis by conditioning on their sufficient statistics to create the conditional likelihood

$$\Pr(\boldsymbol{T}_1 = \boldsymbol{t}_1 | \boldsymbol{T}_0 = \boldsymbol{t}_0) = \frac{\Pr(\boldsymbol{T} = \boldsymbol{t})}{\Pr(\boldsymbol{T}_0 = \boldsymbol{t}_0)}$$
$$= \frac{C(\boldsymbol{t}) \exp(\boldsymbol{t}_1'\boldsymbol{\beta}_1)}{\sum_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{t}_0) \exp(\boldsymbol{u}'\boldsymbol{\beta}_1)}$$

where $C(\boldsymbol{u}, \boldsymbol{t}_0)$ is the number of vectors $\boldsymbol{y}$ such that $\boldsymbol{y}'\boldsymbol{X}_1 = \boldsymbol{u}$ and $\boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0$.

Conditional asymptotic inference is performed by maximizing the conditional likelihood and producing conditional statistics similar to the unconditional likelihood case.

Conditional exact inference is based on generating the conditional distribution for the parameters of interest. This distribution is called the *permutation* or *exact conditional* distribution. The conditional pdf $\Pr(\boldsymbol{T}_1 = \boldsymbol{t}_1 | \boldsymbol{T}_0 = \boldsymbol{t}_0)$ is denoted as $f_{\boldsymbol{\beta}_1}(\boldsymbol{t}_1 | \boldsymbol{t}_0)$. The following section describes the generation of this distribution, and details about the tests and inferences are provided in the appendix.

Exact Conditional Distribution

The goal of the exact conditional analysis is to determine how likely the observed response $\boldsymbol{y}_0$ is with respect to all 2^n possible responses $\boldsymbol{y} = (y_1, ..., y_n)'$. One way to proceed is to generate every $\boldsymbol{y}$ vector for which $\boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0$, and count the number of vectors $\boldsymbol{y}$ for which $\boldsymbol{y}'\boldsymbol{X}_1$ is equal to each unique $\boldsymbol{t}_1$.

Suppose you have the following data, and you want to find the permutation distribution of the sufficient statistics for x1 conditional on those for x0.

Observation	y	x0	x1
1	0	1	1
2	1	1	1
3	0	1	2
4	1	1	0

Here, the observed data are $\boldsymbol{y}_0 = (0, 1, 0, 1)'$, $\boldsymbol{X}_0 = (1, 1, 1, 1)'$, and $\boldsymbol{X}_1 = (1, 1, 2, 0)'$. The observed $\boldsymbol{t}$ is computed as $(t_0, t_1) = 0 \times (1, 1) + 1 \times (1, 1) + 0 \times (1, 2) + 1 \times (1, 0) = (2, 1)$, so you are conditioning on $t_0 = 2$. Tabulate the 16 possible $\boldsymbol{y} = (y_1, y_2, y_3, y_4)'$ vectors and their resulting $\boldsymbol{t} = (t_0, t_1)$ vectors:

	y_1	y_2	y_3	y_4	t_0	t_1
1	0	0	0	0	0	0
2	0	0	0	1	1	0
3	0	0	1	0	1	2
4	0	0	1	1	2	2
5	0	1	0	0	1	1
6	0	1	0	1	2	1
7	0	1	1	0	2	3
8	0	1	1	1	3	3
9	1	0	0	0	1	1
10	1	0	0	1	2	1
11	1	0	1	0	2	3
12	1	0	1	1	3	3
13	1	1	0	0	2	2
14	1	1	0	1	3	2
15	1	1	1	0	3	4
16	1	1	1	1	4	4

The conditional distribution is derived from this joint distribution by extracting every vector with $t_0 = 2$:

t_0	t_1	Frequency	Probability
2	1	2	2/6
2	2	2	2/6
2	3	2	2/6
total		6	1

Generating the conditional distribution from complete enumeration of the joint distribution is conceptually simple; however, this method becomes computationally infeasible very quickly. For example, if you had only 30 observations, you'd have to scan through 2^{30} different $\boldsymbol{y}$ vectors—more than a billion! You can reduce the number of vectors to look at if you are conditioning on the intercept by processing $\binom{30}{\sum_i y_{0,i}}$ vectors, but this does not improve the situation much.

The *multivariate shift algorithm* developed by Hirji, Mehta, and Patel (1987) is a faster method of generating and counting the $\boldsymbol{y}$ vectors for larger problems. The algorithm is based on the following observation. Given any $\boldsymbol{y} = (y_1, ..., y_n)'$ and a design $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$, let $\boldsymbol{y}_{(i)} = (y_1, ..., y_i)'$ and $\boldsymbol{X}_{(i)} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_i)' = \begin{pmatrix} x_{1,1} & \dots & x_{1,p+q} \\ \vdots & & \vdots \\ x_{i,1} & \dots & x_{i,p+q} \end{pmatrix}$ be the

first i rows of each matrix. Write the sufficient statistic based on these i rows as $\boldsymbol{t}'_{(i)} = \boldsymbol{y}'_{(i)}\boldsymbol{X}_{(i)}$. A recursion relation results: $\boldsymbol{t}_{(i+1)} = \boldsymbol{t}_{(i)} + y_{i+1}\boldsymbol{x}_{i+1}$.

The previous example is used to illustrate how this relation is exploited.

Figure 3 displays a tree diagram where each row (after the first) corresponds to an observation i, and each node of the tree is denoted by a pair of digits representing the value of $\boldsymbol{t}_{(i)}$. The top node in the tree is initially set to 00, and indicates that $t_{(0),0} = 0$ and $t_{(0),1} = 0$, or $\boldsymbol{t}_{(0)} = (0, 0)$. Each row of the tree is numbered; these numbers represent the *stages* of the algorithm. To move down the branches, add y times the next value of (x0,x1) to the current value of (t_0, t_1), for $y = 0$ and 1. For example, starting at the zeroth stage with $\boldsymbol{t}_{(0)} = (0, 0)$ =00, take $\boldsymbol{t}_{(0)} + y\boldsymbol{x}_1 = (0, 0) + 0(1, 1) = (0, 0)$ as the value of the left branch of the first stage, and $(0, 0)+1(1, 1) = (1, 1)$ for the right branch.

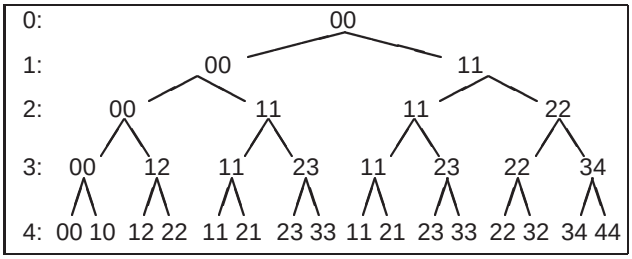

Figure 3. Stages of the Multivariate Shift Algorithm

The following table displays the distribution created from the frequency table of the $2^4 = 16$ possible $\boldsymbol{t}$ vectors from the final stage of Figure 3.

t_0	t_1	Frequency	Probability
0	0	1	1/16
1	0	1	1/16
1	1	2	2/16
1	2	1	1/16
2	1	2	2/16
2	2	2	2/16
2	3	2	2/16
3	2	1	1/16
3	3	2	2/16
3	4	1	1/16
4	4	1	1/16
total		16	1

The conditional distribution obtained for the observed $t_0 = 2$ is the same as previously generated.

There are five shortcuts you can observe from the example:

- There are two $(1, 1)$ nodes in the second stage of Figure 3, and the branches below those two values are identical. Computation time is significantly reduced if you process an entire stage and combine identical nodes; however, the trade-off is that a list of all valid nodes in a stage must be saved, increasing memory usage.

  Note that, in order to obtain the correct distribution, each node descended from this combined $(1, 1)$ node must count as 2 outcomes.
- In the third stage, there is no way to get from $(0, 0)$ to $(2, t_1)$ in one step by adding 0 or 1 times $(1, 0)$; similarly, if the value of t_0 in the third stage is 3, it cannot be reduced to the necessary value of 2. These illustrate what Hirji, Mehta, and Patel (1987) call *infeasibility criteria*.
- The infeasibility criterion is more effective when the larger covariate values are processed first. For example, if the value of x0 for the fourth observation was 2 instead of 1, then you could obtain a $(2, t_1)$ from the $(0, 0)$ third stage node, and hence you would have to process the extra nodes.
- Since the first two observations have the same covariate values, you can jump from stage 0 to stage 2 by combining the first two observations, incrementing the values in stage 0 along three branches with $i \times (1, 1)$ for $i = 0, 1, 2$, and modifying the counts by $\binom{2}{i}$. This saves search time at the expense of computing binomial coefficients.
- Once a distribution is computed for a set of effects, a distribution for any subset of these effects can be produced by scanning the larger distribution. In the example, the conditional distribution for $f_0(t_1|t_0 = 2)$ was produced from the joint distribution $f_0(t_0, t_1)$ by extracting members having $t_0 = 2$.

PROC LOGISTIC's implementation of the multivariate shift algorithm automatically utilizes these shortcuts to improve performance. The bulk of the computation time and memory is consumed by the creation of the exact joint distribution. After the joint distribution for a set of effects is created, the computational effort required to produce hypothesis tests and parameter estimates for any subset of the effects is (relatively) trivial.

EXACT CAPABILITIES OF PROC LOGISTIC

The exact conditional logistic regression analysis in PROC LOGISTIC provides

- two tests for the null hypothesis that the parameters for the effects specified in the EXACT statement are zero: the exact probability test and the exact conditional scores test. For each test, the "Conditional Exact Tests" table displays

  - a test statistic
  - an exact p-value, which is the probability of obtaining a more extreme statistic than the observed, assuming the null hypothesis
  - a mid p-value, which adjusts for the discreteness of the distribution
- parameter estimates and odds ratios for each effect in the EXACT statement conditional on the values of all the other parameters in the model. For each estimate, the "Exact Parameter Estimates" and "Exact Odds Ratios" tables display
  - the exact conditional maximum likelihood estimate (CMLE), or, in cases where the CMLE does not exist, the median unbiased estimate
  - one- or two-sided confidence limits
  - a one- or two-sided p-value for testing that the parameter estimate is zero or the odds ratio is one
- optionally, output data sets containing the derived distributions and summary statistics

Note that hypothesis tests can be generated for each individual effect in an EXACT statement or for all effects simultaneously. See the appendix for more detailed information about the reported tests and statistics.

SYNTAX

The following statements control the exact analyses in the LOGISTIC procedure. Items within the <> are optional.

PROC LOGISTIC <**EXACTONLY**>
<**EXACTOPTIONS(*options*)**>;
EXACT <**'*label*'**>*effects*** <**/*options***>**;**

Several EXACT statements may be specified in any program, but they must follow the MODEL statement. The new EXACTOPTIONS option in the PROC LOGISTIC statement affects every exact analysis requested, whereas options in an EXACT statement are local to that statement. For each EXACT statement, you can include an identifying *label* enclosed in quotes, and specify any *effects* in the MODEL statement or the keyword "intercept". The analysis conditions on any other effects (possibly including the intercept) not specified in the EXACT statement.

PROC LOGISTIC Options

The EXACTONLY option suppresses the unconditional likelihood analyses that PROC LOGISTIC usually performs, and only the exact analyses are executed. Input data sets can be in single-trial or events/trials form, but the response variable must have at most two levels. Options specified in parentheses after the EXACTOPTIONS option apply to every EXACT statement in the program. The following *options* are available:

MAXTIME=*seconds*
STATUSTIME=*seconds*

The MAXTIME= option specifies the maximum clock time (in seconds) that PROC LOGISTIC can use to calculate the permutation distributions. If the limit is exceeded, the procedure halts all computations and prints a note to the SAS LOG. The default maximum clock time is seven days.

The STATUSTIME= option specifies a time interval (in seconds) for printing a status line in the SAS LOG. You can use this status line to track the progress of the computation of the exact conditional distributions. The time interval you specify is approximate; the actual time intervals may vary for larger problems. By default, no status reports are produced.

EXACT Options

Several *options* can be specified in each EXACT statement. The available options are

ALPHA=*value*
ESTIMATE<*=keyword*>
JOINT
JOINTONLY
ONESIDED
OUTDIST=*SAS data set*

The ALPHA= option specifies the significance level for the confidence limits for the parameters; the (default) *value* of 0.05 results in 95% confidence limits.

The ESTIMATE option requests parameter estimates, confidence intervals, and tests for each individual parameter (conditional on all other parameters) specified in the EXACT statement. Optional keywords can be specified; the default ESTIMATE=PARM option requests parameter estimates, ESTIMATE=ODDS requests the odds ratios, and ESTIMATE=BOTH requests both parameter estimates and the odds ratios.

The JOINT option requests a test that all the parameters for the EXACT statement are simultaneously equal to zero in addition to the tests of the individual parameters, while the JOINTONLY option suppresses the default individual tests. The test is indicated in the "Conditional Exact Tests" table by the label "Joint."

The ONESIDED option requests one-sided confidence intervals and p-values for the individual parameter estimates and odds ratios. Note that the two-sided p-values are twice the one-sided p-values.

The OUTDIST= data set contains all of the exact conditional distributions requested in its EXACT statement. This data set contains the possible sufficient statistics for the effects specified in the EXACT statement, the counts derived from the multivariate shift algorithm, the probability of occurrence, and the score value for each sufficient statistic. When you request an OUTDIST= data set, the observed sufficient statistics are displayed in the "Sufficient Statistics" table.

Use with Other Statements and Options

Several existing options can be used in conjunction with the EXACT statement. You can define classification effects and strata using the CLASS statement, you can process the data using BY groups, and you can include a frequency variable with the FREQ statement. The NOINT option in the MODEL statement suppresses the intercept term.

If you receive messages indicating that the Newton-Raphson iterations for the parameter estimates or confidence intervals did not converge, specifying the ABSFCONV=, FCONV=, XCONV=, or MAXITER= options in the MODEL statement may help.

Exact analyses are not performed when you specify a WEIGHT statement, a non-logit link, an offset variable, the NOFIT option, or a model-selection method.

Output

PROC LOGISTIC presents the exact conditional analysis results in several tables:

- The "Conditional Exact Tests" table displays the score and probability statistics for testing that all parameters for the specified effects are zero. By default, tests for a single-effect model are produced, but tests for multiple-effect models can also be requested. Exact and mid p-values are also generated.
- The "Exact Parameter Estimates" table displays the individual parameter estimates (conditional on all other parameters in the model), confidence limits, and a p-value for testing that the parameter is zero.
- The "Exact Odds Ratios" table displays odds ratios for individual parameters, confidence limits, and a p-value for testing that the odds ratio is 1.
- The "Sufficient Statistics" table displays the sufficient statistic for each parameter in the model. This table is only generated when you also specify the OUTDIST= option to output the distribution to a SAS data set. The information is useful for certain further analyses.

As with all SAS procedure output, you can use ODS (Output Delivery System) to create output data sets of the values included in these tables by specifying a statement such as the following:

```
ods output SuffStats=suff ExactTests=test
    ExactParmEst=est ExactOddsRatio=odds;
```

Note that, at this writing, the exact facilities are still under development and the syntax and listing format may change.

EXAMPLES

The following examples illustrate different types of exact analysis. The data in these examples were constructed solely for illustrative purposes. The "Sparse Data" example illustrates that the MLE for the unconditional likelihood analysis may not exist, rendering the asymptotic inference impossible, while the exact conditional inference is still plausible. The "Stratified Analyses" example demonstrates how to use exact conditional analysis to adjust for within-strata correlation. The "Crossover Clinical Trial" example is a popular phase II analysis for the pharmaceutical industry.

Sparse Data

There are several types of data for which unconditional maximum likelihood estimates fail to exist, or for which the theory is not applicable. For data with small cell counts, tests based on the asymptotic normality of the maximum likelihood estimates may not be valid. For other data, the maximum likelihood estimates may not exist and the estimated dispersion matrix may be unbounded. In this example, the data set separate contains variables which perfectly predict the response, yielding a complete separation of data points.

```
data separate;
   input A B Response count @@;
   datalines;
0 0 1 1  0 1 0 2  1 0 1 8  1 1 1 21
;
```

The following statements fit the logistic regression model:

$$\text{logit}(\pi_i) = \alpha + \mathsf{A}\beta_1 + \mathsf{B}\beta_2$$

The JOINT option tests the joint hypothesis that $\beta_1 = \beta_2 = 0$ and the ESTIMATE option produces the individual parameter estimates of β_1 and β_2. The OUTDIST= option creates a data set containing all permutation distributions required for this analysis.

```
proc logistic data=separate;
   freq count;
   model Response=A B;
   exact A B / joint estimate
               outdist=dist;
proc print data=dist;
run;
```

Figure 4 shows that the usual asymptotic analysis indicates that complete separation has occurred. You can see that the parameter estimates do not converge if you specify both the ITPRINT and NOCHECK options in the MODEL statement. However, exact tests and estimates for the conditional analysis can still be computed and are displayed in Figure 5.

Model Convergence Status

Complete separation of data points detected.

Figure 4. Convergence Status

In Figure 5, the joint exact test of A and B is significant, but the B parameter appears insignificant. The median unbiased estimate is created instead of the CMLE because the value of the observed sufficient statistic lies at an extreme of the derived distribution, implying that the CMLE does not exist. Even though the asymptotic results are unreliable, the exact analysis allows you to conclude that there is a significant effect due to A.

Exact Conditional Analysis

Sufficient Statistics

Parameter	Value
Intercept	2
A	0
B	2

Conditional Exact Tests

			--- p-Value ---	
Effect	Test	Statistic	Exact	Mid
Joint	Score	21.1153	0.0020	0.0010
	Probability	0.00202	0.0020	0.0010
A	Score	22.0000	0.0040	0.0020
	Probability	0.00395	0.0040	0.0020
B	Score	2.0000	0.3333	0.1667
	Probability	0.3333	0.3333	0.1667

Exact Parameter Estimates

		95% Confidence		
Parameter	Estimate	Limits		p-Value
A	-3.8398*	-Infinity	-1.0718	0.0079
B	0.6931*	-2.9704	Infinity	0.6667

NOTE: * indicates a median unbiased estimate.

Figure 5. Output from EXACT Analysis

Figure 6 displays the three permutation distributions created with the OUTDIST= option; the joint distribution of A and B conditional on the intercept is contained in observations 1 through 8, the distribution for A conditional on the intercept and B is in observations 9 through 11, and the distribution for B conditional on the intercept and A is in observations 12 and 13. The "Sufficient Statistics" table in Figure 5 allows you to identify the row that contains the observed values. You can see that it is $(\text{intercept}, \text{A}, \text{B}) = (2, 0, 2)$, corresponding to the second, ninth, and thirteenth rows in Figure 6. Note that only the joint distribution for the A and B variables was computed from the multivariate shift algorithm; the univariate conditional distributions were extracted from the joint distribution to save CPU time. The OUTDIST= data set has three values in the distribution for the A variable and two for the B variable. If the permutation distribution is degenerate (has only one value), then the procedure does not produce any statistics and does not output the distribution. However, for small distributions, you have to decide whether there is enough information on which to base the estimates; in this simple example, there is probably too little information contained in the conditional distribution for the B variable.

Obs	A	B	Count	Score	Prob
1	0	1	2	20.2622	0.00403
2	0	2	1	21.1153	0.00202
3	1	0	8	8.9654	0.01613
4	1	1	37	4.4055	0.07460
5	1	2	42	4.9644	0.08468
6	2	0	28	5.5822	0.05645
7	2	1	168	0.7281	0.33871
8	2	2	210	0.9929	0.42339
9	0	.	1	22.0000	0.00395
10	1	.	42	4.5023	0.16601
11	2	.	210	0.1995	0.83004
12	.	1	2	0.5000	0.66667
13	.	2	1	2.0000	0.33333

Figure 6. OUTDIST= Data Set

Stratified Analyses

If your data are collected from different hospitals or different families, you can perform a stratified analysis to control for the within group correlation. The strata are treated as nuisance parameters and a conditional likelihood removes them from the analysis. Your model contains a different intercept term for each stratum:

$$\text{logit}(\pi_{hi}) = \alpha_h + \boldsymbol{x}_{hi}\boldsymbol{\beta}$$

where h indexes the strata, α_h are the strata intercepts, and i indexes the subjects within the strata.

With PROC LOGISTIC, you can specify a stratification variable by including it in the CLASS statement. For example, a stratification variable that has three levels can be parameterized as

Stratum	Level 1	Level 2
1	1	0
2	0	1
3	0	0

where the usual intercept term represents the last strata level, and the other strata levels are a combination of the intercept and the appropriate level term. This is defined in the CLASS statement with the PARAM=REF option. Alternatively, you can pa-

rameterize the stratum variable as

Stratum	Level 1	Level 2	Level 3
1	1	0	0
2	0	1	0
3	0	0	1

This is defined in the CLASS statement with the PARAM=GLM option. Since strata and intercepts are conditioned out of this analysis, either form is reasonable.

The stratified data set includes a response variable Y, two explanatory variables X1 and X2, and a stratification variable. The Z variable will be used in a later analysis.

```
data stratified;
   input Stratum Y X1 X2 count @@;
   Z = 2 - Y;
   datalines;
1 0 1 1 1  2 0 1 2 3  3 0 1 0 2
1 0 2 1 1  2 0 2 2 3  3 0 2 1 1
1 1 1 0 1  2 1 2 0 1  3 1 1 0 1
1 1 2 0 1  2 1 3 1 2  3 1 2 2 2
1 1 3 0 2             3 1 3 2 1
;
```

In the following statements, the stratification variable, which is defined in the CLASS statement, is included in the MODEL statement but left out of the EXACT statement, implying that it is a nuisance effect to be conditioned on for the analysis of the X1 and X2 effects of interest.

```
proc logistic descending exactonly;
   freq count;
   class Stratum / param=ref;
   model Y=Stratum X1 X2;
   exact X1 X2 / jointonly estimate;
run;
```

In Figure 7, the joint exact test for the X1 and X2 parameters rejects the null hypothesis. However, the X2 parameter appears insignificant.

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Joint	Score	7.9291	0.0165	0.0162
	Probability	0.000612	0.0077	0.0074

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
X1	1.9979	0.3140	5.2012	0.0126
X2	-1.0097	-2.9152	0.4142	0.1931

Figure 7. Exact Results

This exact analysis should be compared to an asymptotic conditional likelihood analysis, which is available with the PHREG procedure. First, define a variable Z to be 1 if the response is an event and 2 if the response is a nonevent. This variable is used as the time variable as well as the censoring indicator (with 2 as the censored value) in the MODEL statement of PROC PHREG. Also specify the TIES=DISCRETE option to request the discrete logistic model, and the STRATA statement to specify the strata to be conditioned on.

```
proc phreg;
   freq count;
   strata Stratum;
   model Z*Z(2)=X1 X2 / ties=discrete;
run;
```

The output of PROC PHREG is shown in Figure 8.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.6425	2	0.0081
Score	7.9291	2	0.0190
Wald	4.6510	2	0.0977

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
X1	1	2.32474	1.11585	4.3404	0.0372	10.224
X2	1	-1.11430	0.72917	2.3353	0.1265	0.328

Figure 8. PROC PHREG Results

Comparing Figure 7 with Figure 8, you can see that the value of the conditional score statistic for testing the overall null hypothesis $\beta_1 = \beta_2 = 0$ is 7.9291 for both the asymptotic conditional analysis in PROC PHREG and the exact analysis in PROC LOGISTIC. However, PROC PHREG computes a p-value of 0.019 by comparing the value of the conditional score statistic to a chi-squared distribution with 2 degrees of freedom (since there are two parameters), while PROC LOGISTIC derives a p-value of 0.0165 from the exact conditional distribution. Inference on individual parameters is often not the same between the exact conditional analysis and the asymptotic conditional likelihood results.

Crossover Clinical Trial

One common use of conditional logistic regression is in a crossover clinical trial. In this example, the subjects are given a sequence of drugs, and their response to each drug is recorded. Each subject is considered to be a separate stratum. The goal is to determine if the drugs have the same effect, adjusting for period and carryover effects. In this example, researchers give 15 different subjects three different drugs (A,B,P=placebo) in three consecutive periods (P1,P2,P3), and their response in each period is 1 for improvement and 0 for no improvement. The carryover effect is a classification variable indicating which drug was given in the preceding period.

```
data Crossover (drop=P1 P2 P3);
   input Subject P1$ P2$ P3$ Improve @@;
   Period=1; Drug=P1; Carry='0'; output;
   input Improve @@;
   Period=2; Drug=P2; Carry=P1;  output;
   input Improve @@;
   Period=3; Drug=P3; Carry=P2;  output;
   datalines;
1  A B P 0 0 0    8  B P A 0 0 1
2  A B P 1 1 0    9  B P A 1 0 1
3  A B P 0 1 1    10 B P A 0 1 0
4  A P B 1 0 1    11 P A B 0 1 0
5  A P B 1 0 0    12 P B A 1 0 1
6  B A P 0 0 0    13 P B A 0 0 1
7  B A P 1 1 0    14 P B A 0 1 0
                  15 P B A 0 1 1
;
```

The model to be fit is

$$\begin{aligned}\text{logit}(\pi_{hi}) &= \alpha_h + I(\text{Drug} = \text{A})\beta_1 \\ &\quad + I(\text{Drug} = \text{B})\beta_2 \\ &\quad + I(i = 1)\beta_3 + I(i = 2)\beta_4\end{aligned}$$

where h indexes the subject, α_h are the subject intercepts, i indexes the period, and the $I(\cdot)$ are indicator variables taking the value 1 when the condition is true. Note that this model ignores carryover effects.

```
proc logistic descending exactonly;
   class Subject Drug Period/ param=ref;
   model Improve=Subject Drug Period;
   exact 'one'   Drug Period/ jointonly;
   exact 'two'   Drug / jointonly;
   exact 'three' Period / jointonly;
run;
```

Even though three EXACT statements are invoked in this example, PROC LOGISTIC only computes the permutation distribution for the joint test of the drug and period parameters; the other two distributions are derived from the joint distribution.

The exact conditional score p-value for the test of significance of all the parameters is 0.1835; hence, you cannot reject the null hypothesis. However, the exact conditional score p-value for the test of no drug effects, $\beta_1 = \beta_2 = 0$, is 0.0583, while the p-value for the test of no period effects, $\beta_3 = \beta_4 = 0$, is 0.8605, which suggests that the period term should be dropped from this model.

APPENDIX

Hypothesis Tests

Using the same notation as in the "METHODOLOGY" section, consider testing the null hypothesis $H_0\colon \boldsymbol{\beta}_1 = \mathbf{0}$ against the alternative $H_A\colon \boldsymbol{\beta}_1 \neq \mathbf{0}$, conditional on $\boldsymbol{T}_0 = \boldsymbol{t}_0$. Under the null hypothesis, the test statistic for the *exact probability test* is just $f_{\beta_1=0}(\boldsymbol{t}_1|\boldsymbol{t}_0)$, while the corresponding p-value is the probability of getting a less likely (more extreme) statistic,

$$p(\boldsymbol{t}_1|\boldsymbol{t}_0) = \sum_{\boldsymbol{u}\in\Omega_p} f_0(\boldsymbol{u}|\boldsymbol{t}_0)$$

where $\Omega_p = \{\boldsymbol{u}\text{: there exist } \boldsymbol{y} \text{ with } \boldsymbol{y}'\boldsymbol{X}_1 = \boldsymbol{u}, \boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0, \text{ and } f_0(\boldsymbol{u}|\boldsymbol{t}_0) \leq f_0(\boldsymbol{t}_1|\boldsymbol{t}_0)\}$.

For the *exact conditional scores test*, the conditional mean $\boldsymbol{\mu}_1$ and variance matrix $\boldsymbol{\Sigma}_1$ of the $\boldsymbol{T}_1$ (conditional on $\boldsymbol{T}_0 = \boldsymbol{t}_0$) are calculated, and the score statistic for the observed value,

$$s = (\boldsymbol{t}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{t}_1 - \boldsymbol{\mu}_1)$$

is compared to the score for each member of the distribution

$$S(\boldsymbol{T}_1) = (\boldsymbol{T}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{T}_1 - \boldsymbol{\mu}_1)$$

The resulting p-value is

$$p(\boldsymbol{t}_1|\boldsymbol{t}_0) = Pr(S \geq s) = \sum_{\boldsymbol{u}\in\Omega_s} f_0(\boldsymbol{u}|\boldsymbol{t}_0)$$

where $\Omega_s = \{\boldsymbol{u}\text{: there exist } \boldsymbol{y} \text{ with } \boldsymbol{y}'\boldsymbol{X}_1 = \boldsymbol{u}, \boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0, \text{ and } S(\boldsymbol{u}) \geq s\}$.

The mid-p statistic, defined as

$$p(\boldsymbol{t}_1|\boldsymbol{t}_0) - \frac{1}{2} f_0(\boldsymbol{t}_1|\boldsymbol{t}_0)$$

was proposed by Lancaster (1961) to compensate for the discreteness of a distribution. Refer to Agresti (1992) for more information.

Inference for a Single Parameter

Exact parameter estimates are derived for a single parameter β_i by regarding all the other parameters $\boldsymbol{\beta}_0 = (\beta_1, ..., \beta_{i-1}, \beta_{i+1}, ..., \beta_{p+q})'$ as nuisance parameters. The appropriate sufficient statistics are $\boldsymbol{T}_1 = T_i$ and $\boldsymbol{T}_0 = (T_1, ..., T_{i-1}, T_{i+1}, ..., T_{p+q})'$, with their observed values denoted by the lowercase t. Hence, the conditional pdf used to create the parameter estimate for β_i is

$$f_{\beta_i}(t_i|\boldsymbol{t}_0) = \frac{C(t_i, \boldsymbol{t}_0)\exp(t_i\beta_i)}{\sum_{u\in\Omega} C(u, \boldsymbol{t}_0)\exp(u\beta_i)}$$

for $\Omega = \{u\text{: there exist } \boldsymbol{y} \text{ with } T_i = u \text{ and } \boldsymbol{T}_0 = \boldsymbol{t}_0\}$.

The maximum exact conditional likelihood estimate is the quantity $\widehat{\beta}_i$ which maximizes the conditional pdf.

A Newton-Raphson algorithm is used to perform this search. However, if the observed t_i attains either its minimum or maximum value in the permutation distribution (that is, either $t_i = \min\{u : u \in \Omega\}$ or $t_i = \max\{u : u \in \Omega\}$), then the conditional pdf is monotonically increasing in β_i and cannot be maximized. In this case, a median unbiased estimate (Hirji, Tsiatis, and Mehta 1989; Hirji and Tang 1998) $\widehat{\beta}_i$ is produced that satisfies $f_{\widehat{\beta}_i}(t_i|\mathbf{t}_0) = \frac{1}{2}$, and a Newton-Raphson-type algorithm is used to perform the search.

Likelihood ratio tests based on the conditional pdf are used to test the null $H_0\colon \beta_i = 0$ against various alternatives. For testing against the alternative $H_A\colon \beta_i > 0$, the critical region for the UMP test consists of the upper tail of values for T_i in the permutation distribution. Thus, the one-sided significance level $p_G(t_i; 0)$ is the probability of a more extreme (greater) value:

$$p_G(t_i; 0) = \sum_{u \geq t_i} f_0(u|\mathbf{t}_0)$$

The one-sided significance level $p_L(t_i; 0)$ against $H_A\colon \beta_i < 0$ is

$$p_L(t_i; 0) = \sum_{u \leq t_i} f_0(u|\mathbf{t}_0)$$

The minimum of these one-sided levels is reported when the ONESIDED option is specified. The two-sided significance level $p(t_i; 0)$ against $H_A\colon \beta_i \neq 0$ is calculated as

$$p(t_i; 0) = 2\min[p_L(t_i; 0), p_G(t_i; 0)]$$

An upper $100(1 - 2\epsilon)\%$ confidence limit for $\widehat{\beta}_i$ corresponding to the observed t_i is the solution $\beta_U(t_i)$ of $\epsilon = p_L(t_i, \beta_U(t_i))$, while the lower confidence limit is the solution $\beta_L(t_i)$ of $\epsilon = p_G(t_i, \beta_L(t_i))$. A Newton-Raphson procedure is used to search for the solutions.

ACKNOWLEDGMENTS

I am grateful to Virginia Clark, Greg Goodwin, Ying So, Maura Stokes, and Randy Tobias of the Applications Division at SAS Institute for their valuable assistance in the preparation of this manuscript.

REFERENCES

Agresti, Alan (1990), *Categorical Data Analysis,* New York: John Wiley & Sons, Inc.

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–177.

Cox, D.R. (1970), *Analysis of Binary Data,* New York: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1989), *Analysis of Binary Data,* Second Edition, New York: Chapman and Hall.

Hirji, Karim F., Mehta, Cyrus R., and Patel, Nitin R. (1987), "Computing Distributions for Exact Logistic Regression," *JASA*, 82, 1110–1117.

Hirji, Karim F. and Tang, Man-Lai (1998), "A Comparison of Tests for Trend," *Communications in Statistics—Theory and Methods*, 27, 943–963.

Hirji, Karim F., Tsiatis, Anastasios A., and Mehta, Cyrus R. (1989), "Median Unbiased Estimation for Binary Data," *American Statistician*, 43, 7–11.

Lancaster, H. O., (1961), "Significance Tests in Discrete Distributions," *JASA*, 56, 223–234.

Mehta, Cyrus R. and Patel, Nitin R. (1995), "Exact Logistic Regression: Theory and Examples," *Statistics in Medicine*, 14, 2143–2160.

Stokes, Maura E., Davis, Charles S., and Koch, Gary G. (1995), *Categorical Data Analysis Using the SAS System,* Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at

Robert E. Derr, SAS Institute Inc., SAS Campus Drive, R5245, Cary, NC 27513. Phone (919) 677-8000 ext 6137. FAX (919) 677-4444. E-mail Bob.Derr@sas.com

Version 3.0