Paper 268

Recent Advances in Categorical Data Analysis

Maura E. Stokes
SAS Institute Inc.
Cary, North Carolina, USA

Abstract

The last fifteen years have brought many changes to the practice of categorical data analysis. This paper reviews some of the major changes and shifts of emphasis and discusses several examples using SAS® software procedures. Topics include the use of exact methods, Generalized Estimating Equations, conditional logistic regression, and current uses of weighted least squares modeling. Applications provide illustrations for many topics. This paper describes software currently available in the SAS System and indicates the areas where users can expect to find strategies implemented in the next few releases. The references include some pertinent methodology and review papers.

Introduction

Fifteen years ago, the SUGI paper describing the categorical data analysis enhancements in the upcoming Version 5 release described the new CATMOD procedure that would replace PROC FUNCAT and also described how the FREQ procedure would now include the Mantel-Haenszel nonparametric statistics then present in PROC TFREQ. The paper then continued with a discussion directed at weighted least squares methods for statistical modeling and Mantel-Haenszel methods for testing association in a contingency table or sets of contingency tables.

Since that time, new developments in categorical data analysis and new computing strategies have changed the typical strategies employed by the data analyst facing categorical data. Once safely armed with her CATMOD documentation and PROC FREQ, then patiently learning the difference between the supplemental procedure PROC LOGIST contributed by Frank Harrell and PROC LOGISTIC, the analyst now contends with a number of different procedures such as LOGISTIC, GENMOD, and PHREG, and keeps a scorecard of the growing number of exact tests available in the FREQ and NPAR1WAY procedures.

Mantel-Haenszel strategies are now employed in the analysis of repeated measurements and crossover studies. GEE methods provide a convenient way of modeling repeated measurements data that can include continuous covariates, missing data, and time-dependent covariates. Highly stratified data can be handled with conditional logistic methods, which also provide a way of analyzing crossover data. When asymptotic assumptions are not appropriate due to sparse data or small sample sizes, exact methods can provide a way to produce appropriate p-values for tests, valid confidence limits for odds ratios, and parameter estimates and standard errors in logistic regression. Weighted least squares, a very important strategy in the 70s and 80s, still provides a useful way of modeling functions of interest such as rank measures of association and incidence densities.

What follows are descriptions of these newer strategies and illustrations of their application.

Exact p-Values

Exact p-values provide an alternative strategy when data are sparse, skewed, or unbalanced so that the assumptions required for standard asymptotic tests are violated. Advances in computer performance and developments in network algorithms over the last decade have made exact p-values accessible for a number of statistical tests. In Release 6.11, exact p-values were added for the simple linear rank statistics produced by the NPAR1WAY procedure. In Release 6.12, exact p-values are produced for many of the statistics computed by the FREQ procedure. You are now able to request exact p-values for the following chi-square statistics: Pearson's chi-square, likelihood-ratio chi-square, Mantel-Haenszel chi-square, Fisher's exact test and r by c exact test, Jonckheere-Terpstra test, and McNemar's test. In addition, you can also obtain exact p-values for hypothesis tests that the following statistics are equal to 0: Pearson correlation coefficient, Spearman correlation coefficient, simple kappa statistic, and weighted kappa statistic. Exact confidence bounds are also available for the odds ratios produced for 2 by 2 tables.

In Version 7, a test for the binomial proportion is available along with an exact p-value, and exact p-values are available for the Pearson chi-square statis-

tic for two-way tables and the goodness-of-fit statistic for one-way tables. In addition, the Monte Carlo method of computing exact p-values is included in the NPAR1WAY procedure and will be included in the FREQ procedure in Version 8.

The following example illustrates the use of the new EXACT statement to produce an exact p-value for the simple kappa statistic. Researchers studied two scoring systems for evaluating fitness in fifth grade students. Forty-three students were classified into one of four fitness categories. Interest lies in determining whether there is agreement between the two scoring systems, which can be assessed by testing whether the kappa coefficient is equal to 0.

```
data fitness;
   input score1 $ score2 $ count;
   datalines;
   poor poor 5
   average average 4
   good good 4
   superior superior 3
   poor average 3
   average poor 1
   average good 6
   good average 5
   good superior 1
   superior average 10
   superior good 1
   ;
```

To request the exact p-value for the kappa statistic, you specify the keyword KAPPA in the EXACT statement. The AGREE option in the MODEL statement requests the measures of agreement.

```
proc freq;
   weight count;
   tables score1 * score2 / agree;
   exact kappa;
run;
```

The following figure displays the contingency table form of the data. Note the number of zero cells, which makes the use of the asymptotic test questionable.

TABLE OF SCORE_1 BY SCORE_2

SCORE_1 \ SCORE_2	average	good	poor	superior	Total
average	4	6	1	0	11
	9.30	13.95	2.33	0.00	25.58
	36.36	54.55	9.09	0.00	
	18.18	54.55	16.67	0.00	
good	5	4	0	1	10
	11.63	9.30	0.00	2.33	23.26
	50.00	40.00	0.00	10.00	
	22.73	36.36	0.00	25.00	
poor	3	0	5	0	8
	6.98	0.00	11.63	0.00	18.60
	37.50	0.00	62.50	0.00	
	13.64	0.00	83.33	0.00	
superior	10	1	0	3	14
	23.26	2.33	0.00	6.98	32.56
	71.43	7.14	0.00	21.43	
	45.45	9.09	0.00	75.00	
Total	22	11	6	4	43
	51.16	25.58	13.95	9.30	100.00

Figure 1. Exact Test for Simple Kappa

The resulting exact p-value for the hypothesis that the simple kappa statistic is equal to 0 is p=0.055, which may be considered to have marginal significance at best. Note the value p=0.038 for the asymptotic test. Using exact p-values for this analysis leads to a very different conclusion than using the asymptotic test.

STATISTICS FOR TABLE OF SCORE_1 BY SCORE_2

Test of Symmetry

Statistic = 11.091 DF = 6 Prob = 0.086

Simple Kappa Coefficient

		95% Confidence Bounds	
Kappa = 0.167	ASE = 0.102	-0.032	0.366

Asymptotic P-Values	Exact P-Values
(Right-sided) = 0.019	(Right-sided) = 0.034
(Two-sided) = 0.038	(Two-sided) = 0.055

Weighted Kappa Coefficient

		95% Confidence Bounds	
Kappa = 0.100	ASE = 0.100	-0.096	0.297

Sample Size = 43

Figure 2. Exact Test for Simple Kappa

Generalized Estimating Equations

Weighted least squares (WLS) modeling of repeated categorical data was described by Landis et al. (1977) and provides a large sample asymptotic method that works nicely for data that have adequate sample size, a small number of response points, a small number of categorical explanatory variables measured at the subject level, and no missing data. The CATMOD procedure introduced the REPEATED statement to provide this analysis.

While this method is still useful for data that meet these conditions, most data that are collected with clustered or repeated responses do not. Longitudinal data are usually plagued with missing responses,

explanatory variables include continuous variables as well as categorical, and there is often interest in time-dependent covariates such as blood pressure in a clinical trial. In addition, as you start to increase the number of explanatory variables, you often don't meet the asymptotic requirements for the WLS approach.

Generalized Estimating Equations (GEE) provides a nonlikelihood based approach to modeling repeated or clustered data that applies to a broader set of data situations that are frequently encountered. It handles missing data, continous explantory variables, and time-dependent explanatory variables. While responses can be either continuous or categorical, it is especially useful for data that are binary or discrete counts. An extension of the generalized linear model (GLM) first suggested by Nelder and Wedderburn (1972), the GEE approach was outlined in work by Zeger and Liang (1986) and Liang and Zeger (1986) that describe a quasi-likelihood approach for modeling correlated responses. Besides using the linear predictor set-up of the GLM, you model the covariance matrix of the responses. The GEE approach produces population-averaged estimates. With quasi-likelihood, you can pursue statistical models by making assumptions about the link function and the relationship between the first two moments, but without specifying the complete distribution of the response.

Say that $Y_{ij}(j = 1, \ldots, n_i, i = 1, \ldots, K)$ represent the jth measurement on the ith subject. There are n_i measurements on subject i and $\sum_{i=1}^{K} n_i$ total measurements.

The generalized estimating equation for estimating $\boldsymbol{\beta}$ is an extension of the GLM estimating equation:

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

where $\boldsymbol{\mu}$ is the corresponding vector of means $\boldsymbol{\mu} = [\mu_{i1}, \ldots, \mu_{in_i}]'$ and $\mathbf{V}_i$ is an estimate of the covariance matrix of $\mathbf{Y}$.

The covariance matrix of $\mathbf{Y}_i$ is modeled as

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element.

The working correlation matrix $\mathbf{R}_i(\alpha)$ is estimated as

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual.

There are many choices for the working correlation matrix. The independent working correlation matrix includes 1s on the diagonals and 0s on the off-diagonals. Other choices are the unstructured, exchangeable (compound symmetry), autoregressive(1), and m-dependent.

Finding the GEE solution requires these steps:

- Relate the marginal response μ_{ij}=E(Y_{ij}) to $\mathbf{x}'_{ij}\boldsymbol{\beta}$ with a link function. For example, the logit.
- Specify the variance function.
- Choose a working correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$.
- Compute an initial estimate of $\boldsymbol{\beta}$, for example with an ordinary generalized linear model assuming independence.
- Compute the working correlation matrix $\mathbf{R}_i$.
- Compute an estimate of the covariance matrix:

  $$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \hat{\mathbf{R}}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$$

- Update $\boldsymbol{\beta}$:

  $$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r - \left[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} (\mathbf{Y} - \boldsymbol{\mu}_i)\right]$$

- Compute residuals and update $\mathbf{V}_i$.
- Iterate until convergence.

The GEE parameter estimates have many important properties. They are the generalized linear model estimating equations when you have one measurement per cluster. They are the maximum likelihood score equations for multivariate Gaussian data. And, most importantly, the GEE parameter estimates are consistent as the number of clusters becomes large, even if you have misspecified the working correlation matrix, as long as the mean model is correct.

The model-based estimator of Cov$(\hat{\boldsymbol{\beta}})$ is given by

$$\text{Cov}_M(\hat{\boldsymbol{\beta}}) = \mathbf{I}_0^{-1}$$

where

$$\mathbf{I}_0 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}' \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$$

This is a consistent estimator if the model and working correlation matrix are correctly specified.

The empirical, or robust, estimator of Cov($\hat{\boldsymbol{\beta}}$) is given by

$$\mathbf{M} = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$$

where

$$\mathbf{I}_1 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}' \mathbf{V}^{-1}\text{Cov}(\mathbf{Y})\mathbf{V}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$$

This is a consistent estimator of Cov($\hat{\boldsymbol{\beta}}$) as the number of clusters become large, even if the working correlation matrix is not specified correctly.

The GEE approach produces a *marginal model*. It models a known function of the marginal expectation of the dependent variable as a linear function of explanatory variables. The resulting parameter estimates are population-averaged. GEE relies on independence across subjects to consistently estimate the variance. Compare this to a mixed model where you are estimating subject-specific parameter estimates, but you are heavily leveraging the correlation assumption.

Exercise Study

One of the applications of GEE methods is to crossover studies. In a crossover study, subjects have their response measured at different periods under different conditions. In a classic two period, two treatment crossover study, some subjects get a Treatment A during Period 1 and Treatment B during Period 2, while other subjects get the sequence Treatment B during Period 1 and Treatment A during Period 2. Usually, there is some sort of *washout* period so that whatever effects of the treatment during the first period have *washed out* before the second period begins. Also, the nature of the response is such that the subject is able to have the indicated response a few times within a relatively short period of time. In a crossover study, the subject acts as his own control.

More complicated designs, including sequences that can draw from more than two possible treatments (for example, A, B, and Placebo) and additional periods, can provide a better framework for estimating the treatment effect. In this example, GEE methods are used to analyze a three-period crossover study in which patients with a chronic respiratory condition are exposed to different levels of air pollution while exercising and measured for respiratory distress on a four point ordinal scale, ranging from 0 for none to 3 for severe. A dichotomous baseline distress measurement was taken at the beginning of the study. Six sequences were studied: HML, HLM, MHL, MLH, LHM, and LMH, where 'H' means High, 'M' means medium, and 'L' means low.

In this analysis, the subject is the cluster and there may be a maximum of three response corresponding to the three periods. Missing responses occurred at each of the three periods. Interest lies in determining whether there was a pollution effect, baseline effect, period, and carryover effects.

The following DATA step inputs the exercise data. There is one observation per subject per period. The variable Sequence contains the sequence information, for example, observations with the value 'HML' received the sequence High in the first period, Medium in the second period, and Low in the third period. The indicator variables High and Medium take the value '1' is the exposure is High or Medium, respectively, for that period. ID is the subject ID within sequence group, Period1 and Period2 are indicator variables for whether the observation is from Period 1 or Period 2, and CarryHigh and CarryMedium are indicator variables for whether the previous period was High exposure or Medium exposure. The variable Baseline takes the value '1' for respiratory distress at the beginning of the study.

```
data Exercise;
   input Sequence $ ID $ Period1 Period2 High Medium Baseline
   Response CarryHigh CarryMedium @@;
   strata=sequence||id;
   DichotResponse= (Response >0);
datalines;
HML 1 1 0 1 0 0 3 0 0 HML 1 0 1 0 1 0 1 1 0
HML 1 0 0 0 0 0 0 0 1
HML 2 1 0 1 0 0 3 0 0 HML 2 0 1 0 1 0 2 1 0
HML 2 0 0 0 0 0 0 0 1
HML 3 1 0 1 0 1 3 0 0 HML 3 0 1 0 1 0 2 1 0
HML 3 0 0 0 0 0 . 0 1
HML 4 1 0 1 0 0 2 0 0 HML 4 0 1 0 1 0 0 1 0
HML 4 0 0 0 0 0 2 0 1
...
```

The following statements produce a listing of the number of subjects in each of the sequences.

```
proc freq;
   tables Sequence Response;
run;
```

The FREQ Procedure

Sequence	Frequency	Percent	Cumulative Frequency	Cumulative Percent
HLM	72	16.00	72	16.00
HML	78	17.33	150	33.33
LHM	72	16.00	222	49.33
LMH	72	16.00	294	65.33
MHL	60	13.33	354	78.67
MLH	96	21.33	450	100.00

Figure 3. Frequencies of Exercise Sequences

The GEE analysis is performed with the GEE facility in the GENMOD procedure. This has been made much more comprehensive in Version 7 with the inclusion of Type III tests, the CONTRAST, ESTIMATE, and LSMEANS statement, and the capability of handling the ordinal response with the proportional odds model. PROC GENMOD also now provides the alternating logistic regression method for binary data.

The following statements request the analysis. The crossclassification of the variables Sequence and Id uniquely identify each cluster (subject), so that effect is specified with the SUBJECT= option in the REPEATED statement. The model consisting of all the main effects is specified, and the proportional odds model is requested with the DIST=MULTINOMIAL and LINK=CLOGIT specifications.

```
proc genmod;
   class Id Sequence;
   model Response = Period1 Period2 High Medium
                    Baseline CarryHigh CarryMedium
                    / dist=multinomial
                  link=clogit;
   repeated subject= Sequence*Id /type=ind corrw;
   contrast 'Carryover Effect' CarryHigh 1,
                               CarryMedium 1;
   contrast 'Period Effect' period1 1,
                            period2 1 ;
run;
```

In order to assess joint effects for Period and Carryover, two sets of two-row contrasts are specified.

Figure 4 tells you that the link function and distribution have been specified correctly and that there are 406 total period measurements. Missing values for the response occurs 44 times.

Model Information

Data Set	WORK.EXERCISE
Distribution	Multinomial
Link Function	Cumulative Logit
Dependent Variable	Response
Observations Used	406
Missing Values	44

Figure 4. GLM Information

Figure 5 displays the internal ordering of responses values, which is from 0 to 3, for no distress to severe distress.

Response Profile

Ordered Level	Ordered Value	Count
1	0	87
2	1	130
3	2	127
4	3	62

Figure 5. Ordered Values

Figure 6 displays the information for the GEE analysis. The data includes 150 clusters, for which 37 have missing values. Cluster size ranges from 1 (only one period measured) to 3 (all periods represented).

GEE Model Information

Correlation Structure	Independent
Subject Effect	ID*Sequence (150 levels)
Number of Clusters	150
Clusters With Missing Values	37
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	1

Figure 6. GEE Information

Figure 7 contains the parameter estimates. Neither the Carryover nor Period effects appear to be influential. The Medium exposure appears to be marginally influential with a parameter estimate of -0.4693 and a p-value of 0.0756; the High exposure appears to be very significant with a parameter estimate of -3.1225 and a p-value of less than 0.0001.

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits Lower	Upper	Z	Pr > |Z|
Intercept1	-0.8959	0.3074	-1.4983	-0.2934	-2.91	0.0036
Intercept2	0.9478	0.3290	0.3030	1.5926	2.88	0.0040
Intercept3	3.2798	0.3524	2.5891	3.9705	9.31	<.0001
Period1	0.2609	0.2973	-0.3219	0.8436	0.88	0.3803
Period2	-0.0287	0.2380	-0.4953	0.4378	-0.12	0.9040
High	-3.1225	0.3032	-3.7167	-2.5283	-10.30	<.0001
Medium	-0.4693	0.2641	-0.9869	0.0484	-1.78	0.0756
Baseline	0.4932	0.3708	-0.2335	1.2199	1.33	0.1835
CarryHigh	0.3721	0.3041	-0.2240	0.9682	1.22	0.2211
CarryMedium	0.4265	0.2968	-0.1551	1.0081	1.44	0.1507

Figure 7. Parameter Estimates

The contrast results provide the 2 degree of freedom tests for both the Carryover and Period effects.

CONTRAST Statement Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Carryover Effect	2	2.55	0.2799	Score
Period Effect	2	1.03	0.5976	Score

Figure 8. Results of Contrasts

With both p-values greater than 0.25, these joint tests are non-significant. There appears to be neither Carryover nor Period effects for these data. Note that the default test for the CONTRAST statement used for the GEE analysis is a score test; you can also request a Wald statistic. The score statistics are generally more suitable for smaller sample sizes.

The reduced model was fit with the following MODEL statement. Baseline was retained as a covariate.

```
model response = high medium baseline
                 / dist=multinomial
                 link=clogit;
```

Figure 9 contains the parameter estimates for the final model.

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits Lower	Upper	Z	Pr > |Z|
Intercept1	-0.5404	0.1740	-0.8814	-0.1994	-3.11	0.0019
Intercept2	1.2933	0.1949	0.9114	1.6752	6.64	<.0001
Intercept3	3.6171	0.2586	3.1102	4.1239	13.99	<.0001
High	-3.2523	0.3057	-3.8513	-2.6532	-10.64	<.0001
Medium	-0.6204	0.2293	-1.0698	-0.1711	-2.71	0.0068
Baseline	0.5006	0.3381	-0.1620	1.1631	1.48	0.1387

Figure 9. Reduced Model Results

Note that goodness-of-fit statistics are still being researched for GEE analyses. See work by Barnhart and Williamson (1998) and Preisser and Quaqish (1996) for some recent discussions of goodness of fit and deletion diagnostics. Given adequate sample size, it may be advantageous to assess model adequacy by including additional terms in the model such as pairwise and possibly higher interactions and then performing a joint test on those effects. If the test is non-significant, it leads credence to the corresponding reduced model under consideration.

Conditional Logistic Regression

Conditional logistic regression has long been used in epidemiology where a retrospective study matched subjects, or cases, with an event of interest with a similar subject, or control, who didn't have the event. You determine whether the case and control had the risk factors being investigated, and, by using a conditional likelihood, you can predict the event given the explanatory variables. You set up the probabilities for having the exposure given the event and then apply Bayes' theorem to determine a relevant conditional probability.

More recently, conditional logistic regression has also been applied in the situation of highly stratified data and crossover studies. When you have highly stratified data, you may have a small number of subjects pers stratum, and thus you have a small number of subjects relative to the number of parameters you are estimating because you will need to estimate stratification effects. Sample size requirements for the usual maximum likelihood approach to unconditional logistic regression may not be met.

You have a similiar situation with crossover studies, in which subjects are acting as their own controls.

Highly Stratified Data

Stokes et al. (1995) include an example of a clinical trial in which researchers studied the effects of a new treatment for a skin condition. A pair of patients participated from each of 79 clinics. One person received the treatment and another person received the placebo. Age, sex, and an initial score for the skin condition (ranging from 1 to 4 for mild to severe) were recorded. The response was whether the skin condition improved. Note that because there are only two observations per clinic, it would not be possible to estimate properly a clinic effect. Generally speaking, you would want to have at least five observations per clinic in order for that type of estimation.

However, by conditioning away the clinic effects as nuisance parameters, you can perform a logistic regression that results in far fewer parameters. In Stokes et al., this analysis is performed by recognizing that in the case of pairs within strata, you can create a response that is a within-stratum difference and analyze those differences with the LOGISTIC procedure. However, it's more straightforward to use the PHREG procedure to perform this analysis. While designed for proportional hazards regression analysis, through computational equivalences the procedure can also be used for conditional logistic regression.

The data have the following form, where each line consists of two observations. Indicator variables are created for various interactions and to make treatment into a numerical variable. (A CLASS statement is on the list for future PROC PHREG work).

```
data trial;
      input center treat $ sex $ age improve initial @@;
      /* compute model terms for each observation */
      trt=treat='t';
      i_sex=(sex='m');        i_trt=(treat='t');
      trtsex=i_sex*i_trt;     trtinit=i_trt*initial;
      trtage=i_trt*age;       isexage=i_sex*age;
      isexinit=i_sex*initial;iageinit=age*initial;
       cards;
   1  t f 27 0 1  1 p f 32 0 2
   2  t f 41 1 3  2 p f 47 0 1
   3  t m 19 1 4  3 p m 31 0 4
   4  t m 55 1 1  4 p m 24 1 3
   . . .
```

The following statements request the conditional logistic regression. The variable center is the stratification variable. The TIES=DISCRETE option is required. The first four variables in the MODEL statement are automatically included in the model, and then the procedure produces a score statistic for the joint inclusion of the remaining variables. This serves as a goodness-of-fit check.

```
proc phreg data=trial;
       strata center;
       model improve = trt initial age i_sex
          isexage isexinit iageinit
          trtsex trtinit trtage / ties=discrete
          selection=forward include=4 details;
   run;
```

Figure 10 displays the parameter estimates for this analysis.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Hazard Ratio
trt	1	-0.70244	0.36009	3.8052	0.0511	0.495
initial	1	-1.09148	0.33508	10.6104	0.0011	0.336
age	1	-0.02483	0.02243	1.2252	0.2683	0.975
i_sex	1	-0.53115	0.55451	0.9175	0.3381	0.588

Figure 10. Parameter Estimates for Clinical Trial Data

Figure 11 and Figure 12 contain information about entering variables.

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
isexage	0.6593	0.4168
isexinit	0.1775	0.6736
iageinit	2.9194	0.0875
trtsex	0.2681	0.6046
trtinit	0.0121	0.9125
trtage	0.4336	0.5102

Figure 11. Analysis of Entering Variables

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4.7211	6	0.5800

Figure 12. Score Test for Other Variables

First, consider the statistics displayed in Figure 11 and Figure 12. The residual score statistic value of 4.7211 with 6 df (p-value= 0.5800) indicates that the additional terms are of little consequence. Age and sex appear not to be influential, and you may choose to keep them in the model as covariates or further perform model reduction by fitting the model without them.

Note that since the alphanumeric ordering of the response values is (0,1), the model is based on predicting the probability of no improvement. For dichotomous logistic regression, you can simply switch the signs of the parameter estimates to obtain the estimates for the model based on the probability of improvement. Thus, those with treatment have an odds $e^{.70244} = .019$ times higher of improving than those patients receiving the placebo. This is true even after initial grade is adjusted for in the model. And, the odds of improvement are $e^{1.0918} = 2.980$ times higher per unit increase in initial grade. Note that the conditional logistic analysis has also taken into account any effect of clinic.

Crossover Data

Conditional logistic regression also provides a useful analysis strategy for the crossover design. When you apply conditional logistic regression in this setting, you are creating a strata for each subject. You are conditioning out subject to subject variability and focusing on intrasubject information. Thus, you can often perform analyses that would not be possible with population-averaging methods due to small sample size. Note that the odds ratios resulting from the conditioning approach apply to subjects individually instead of to subjects on average. This may be an important consideration depending on your analysis objectives. For example, in a study whose objective is to produce a model that can be used for patient protocol prediction, the conditional logistic model may be more appropriate.

The exercise data above are reanalyzed with the conditional logistic model, using the PHREG procedure. For this analysis, the response is dichotomized into 1, for severe response, and 0, for other responses. In addition, a new variable Strata has been defined, which is a unique identifier for each subject based on a combination of Sequence and Id.

The STRATA statement defines the strata; note that the specification TIES=DISCRETE is required in order to produce the correct estimates. You use the TEST statement to specify tests concerning the parameter estimates: here, joint tests for both the Carryover and Period effects are requested.

```
proc phreg data=Exercise;
   strata Strata;
   model dichotresponse = period1 period2 high
          medium baseline CarryHigh CarryMedium
          / ties=discrete;
   Carryover: test CarryHigh=CarryMedium=0;
   Period:    test Period1=period2=0;
run;
```

Results of the parameter estimation are displayed in Figure 13 and are similar to those obtained in the GEE analysis. High and Medium exposures are influential, and it doesn't appear that there are Carryover or Period effects.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Hazard Ratio
Period1	1	-0.15440	0.44683	0.1194	0.7297	0.857
Period2	1	-0.26202	0.31225	0.7041	0.4014	0.769
High	1	-1.70458	0.35000	23.7198	<.0001	0.182
Medium	1	-0.68624	0.35899	3.6542	0.0559	0.503
Baseline	1	0.65094	0.52766	1.5219	0.2173	1.917
CarryHigh	1	0.40777	0.44493	0.8400	0.3594	1.503
CarryMedum	1	0.28252	0.53024	0.2839	0.5942	1.326

Figure 13. Parameter Estimates for Exercise Data

The results for the joint tests for Carryover and Period are displayed in Figure 14. The Wald statistics indicate that neither effect is important.

A reduced model seemed to fit the data adequately.

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
CARRYOVER	0.8723	2	0.6465
PERIOD	0.7083	2	0.7018

Figure 14. Test Results for Carryover and Period

Exact Logistic Regression

Sometimes, sample sizes are simply not appropriate for the usual logistic regression or conditional logistic regression strategies to be appropriate. In those cases, exact logistic regression provides a means of producing regression estimates and standard error that are statistically valid. You use conditioning principles similar to those used in conditioning on observed margins of contingency tables to obtain exact tests. You eliminate nuisance parameters by conditioning on the observed values of their sufficient statistics and then use the exact permutational distribution of the sufficient statistics for the parameters of interest. You can apply conditional inference to both the unstratified and stratified logistc regression settings. See Mehta and Patel (1995). LogXact software, Cytel Software Corporation (1993), provides this capability currently.

Weighted Least Squares Modeling of Categorical Response Functions

While weighted least squares may not longer be the workhorse of categorical data modeling, it still plays an important role in providing a strategy for modeling the variation among various functions that can be computed for categorical data. Essentially you are applying noniterative generalized least squares to response functions that are of interest and using an observed covariance matrix as the weights. If you have adequate sample sizes, the response functions have a approximate multivariate normal distribution and you can carry out hypothesis tests concerning linear combinations of them. Rank meaures of association are one type of response function that may be of interest; another is incidence densities.

Modeling Rank Measures of Association Statistics

Many studies include outcomes that are ordinal in nature. When the treatment is also ordinal, rank measures of correlation can be modeled using WLS methods to investigate various treatment effects and interactions; such an analysis can often complement statistical models such as the proportional odds model. Refer to Carr et al (1989) for an example of such an analysis applied to Goodman-Kruskal rank correlation coefficients, also known as gamma coefficients.

The Mann-Whitney rank measure of association statistics are also useful statistics for assessing the association between an explanatory variable and an ordinal outcome. Consider the data from a randomized clinical trial of chronic pain (Stokes et al, 1995). Investigators compared a new treatment with a placebo and assessed the response for a particular condition. Patients were obtained from two investigators whose design included stratification relative to four diagnostic classes.

Table 1 displays these data.

Table 1. Chronic Pain Data

Diagnostic Class	Researcher	Treatment	Patient Status P	F	M	G	E
I	A	Active	3	2	2	1	0
I	A	Placebo	7	0	1	1	1
I	B	Active	1	6	1	5	3
I	B	Placebo	5	4	2	3	3
II	A	Active	1	0	1	2	2
II	A	Placebo	1	1	0	1	1
II	B	Active	0	1	1	1	6
II	B	Placebo	3	1	1	5	0
III	A	Active	2	0	3	3	2
III	A	Placebo	5	0	0	8	1
III	B	Active	2	4	1	10	3
III	B	Placebo	2	5	1	4	2
IV	A	Active	8	1	3	4	0
IV	A	Placebo	5	0	3	3	0
IV	B	Active	1	5	2	3	1
IV	B	Placebo	3	4	3	4	2

You may be interested in computing the Mann-Whitney rank measure of association as a way of assessing the extent to which patients with active treatments are more likely to have better response status than those with placebo. You may then be interested in seeing whether diagnostic status and investigator influence this association through model-fitting. You can perform such modeling by first computing the Mann-Whitney statistics and their standard errors and then using these estimates as input to the CATMOD procedure to perform modeling.

You can compute the Mann-Whitney measures as functions of the Somer's D measures, which are produced by the FREQ procedure.

$$U_i = \frac{\{\text{Somer's D C|R} + 1\}}{2} \text{ and } S_i = \frac{SE}{2}$$

S_i is the standard error of U_i, the Mann-Whitney statistic.

The following statements produce measures of association for the eight 2 $\times$ 4 tables formed for the combination of investigator and treatment.

```
data cpain;
      input dstatus $ invest $ treat $
      status $ count @@;
      datalines;
I  A active  poor     3 I A active  fair 2
I  A active  moderate 2 I A active good  1
I  A active  excel    0 I A placebo poor 7
I  A placebo fair     0 I A placebo moderate 1
I  A placebo good     1 I A placebo excel 1
I  B active  poor     1 I B active  fair 6
I  B active  moderate 1 I B active good  5
I  B active  excel    3 I B placebo poor 5
I  B placebo fair     4 I B placebo moderate 2
I  B placebo good     3 I B placebo excel 3
II A active  poor     1 II A active  fair 0
II A active  moderate 1 II A active good  2
II A active  excel 2
...
proc freq;
   weight count;
   tables dstatus*invest*treat*status/ measures;
run;
```

Figure 15 displays the table for Diagnostic Status I and Investigator A. Figure 16 displays the measures of association for that table.

The FREQ Procedure

Table 1 of treat by status
Controlling for dstatus=I invest=A

treat status

Frequency | Row Pct	poor	fair	moderate	good	excel	Total
active	3 | 60.00	2 | 40.00	2 | 66.67	1 | 33.33	0 | 0.00	8
placebo	7 | 70.00	0 | 0.00	1 | 10.00	1 | 10.00	1 | 10.00	10
Total	10	2	3	2	1	18

Figure 15. Frequency Table

Statistics for Table 1 of treat by status
Controlling for dstatus=I invest=A

Statistic	Value	ASE
Gamma	-0.2857	0.3515
Kendall's Tau-b	-0.1763	0.2253
Stuart's Tau-c	-0.1975	0.2485
Somers' D C|R	-0.2000	0.2514
Somers' D R|C	-0.1553	0.2026
Pearson Correlation	-0.0866	0.2331
Spearman Correlation	-0.1900	0.2416

Figure 16. Measures of Association

Table 2 displays the calculated values.

Table 2. Mann Whitney Statistics

Diagnostic Class	Researcher	Somer's	ASE	U_i	S_i
I	A	.2000	.3515	.6000	.1758
I	B	.2002	.1915	.6001	.0958
II	A	.2083	.3622	.6042	.1811
II	B	.6778	.1834	.8389	.0917
III	A	.0260	.2271	.5130	.1136
III	B	.1893	.1923	.5947	.0962
IV	A	.0000	.2007	.5000	.1004
IV	B	−.0156	.2116	.4922	.1058

You compute the covariances and then create a data set that contains the estimates and the covariance matrix. The following DATA step creates the MannWhitney data set.

```
data MannWhitney;
   input b1-b8 _type_ $ _name_ $8.;
   datalines;
    .6000  .6011  .6042 .8389 .5130 .5947 .5000 .4922 parms
    .03091 .0000  .0000 .0000 .0000 .0000 .0000 .0000 cov b1
    .0000  .00918 .0000 .0000 .0000 .0000 .0000 .0000 cov b2
    .0000  .0000  .3280 .0000 .0000 .0000 .0000 .0000 cov b3
    .0000  .0000  .0000 .0084 .0000 .0000 .0000 .0000 cov b4
    .0000  .0000  .0000 .0000 .0129 .0000 .0000 .0000 cov b5
    .0000  .0000  .0000 .0000 .0000 .0093 .0000 .0000 cov b6
    .0000  .0000  .0000 .0000 .0000 .0000 .0101 .0000 cov b7
    .0000  .0000  .0000 .0000 .0000 .0000 .0000 .0112 cov b8
;
```

This data set is then input into the CATMOD procedure. Thus, instead of generating functions from an underlying contingency table, the CATMOD procedure does modeling directly on the input functions using the input covariance matrix as the weights. You define the profiles for each function with the PROFILE option in the FACTORS statement. You also define your factors, along with the number of levels for each, and describe the effects you want to include in your model with the _RESPONSE_ option.

```
proc catmod data=MannWhitney;
        response read b1-b8;
        factors diagnos $ 4 , invest $ 2 /
             _response_ = diagno invest
        profile = (I   A,
                   I   B,
                   II  A,
                   II  B,
                   III A,
                   III B,
                   IV  A,
                   IV  B);
        model _f_ = _response_ / cov;
 run;
```

The ANOVA table results follow. The residual Wald test is a test of the diagnostic class and investigator interaction, which is non-significant with a p-value of 0.78. Neither diagnostic class nor investigator appear to explain significant variation, with diagnostic class appearing to be modestly influential with a p-value of 0.093.

Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	213.14	<.0001
diagnos	3	6.42	0.0930
invest	1	0.58	0.4469
Residual	3	1.06	0.7862

Figure 17. Main Effects Output

By submitting another MODEL statement like the following

```
model _f_ =  / cov; run;
```

You can obtain a test of the hypothesis that the measures have the same value for each diagnostic class and investigator combination. This is the seven degree test that is labeled 'residual.'

Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Model|Mean	7	9.41	0.2247
Residual	0	.	.

Figure 18. Intercept-Only Model

You can't reject this hypothesis with these data.

Modeling Incidence Densities

Incidence densities are defined as the ratio of the number of episodes of a disease or illness to the total person-time at risk. These measures are often of interest in epidemiological work. Lavange et al (1994) studied the effect of passive smoking exposure on the incidence of lower respiratory illness in young children. Measurements were made over time. The investigators used WLS to model the incidence densities; because the study was a complex survey design, the covariance structure was determined with sample survey methods. Interest was in comparing the marginal rates of lower respiratory disease between exposed and unexposed groups. The ratio estimates and their covariances were obtained with SUDAAN software, and then the response functions and covariances wered used as input to the CATMOD procedure for WLS modeling. The exposed group had a significantly higher rate of illness.

Summary

This paper describes recent enhancements in the area of categorical data analysis and discusses several applications of the recent methodology. Exact methods and quasi-likelihood methods provide ways to analyze data that previously had many data analysis limitations. The SAS System includes software for many of these newer methodologies and should contain additional features that implement the recent methodological advances in the next several years.

References

Barnhart, H. and Williamson, J (1998). Goodness-of-Fit Tests for GEE Modeling with Binary Responses, *Biometrics*, 54, 720–729.

Cytel Software Corporation (1993), *LogXact: Software for Exact Logistic Regression*, Cytel Software Corporation, Cambridge, MA.

Carr, G. J., Hafner, K. B., and Koch, G. G., (1989), "Analysis of rank measures of association for ordinal data from longitudinal studies", *Journal of the American Statistical Association*, 84, 797–804.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford: Oxford Science

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, 13–22

LaVange, L. M., Keyes, L. L., Koch, G. G., and Margolis, P. A. (1994). Application of sample survey methods for modelling ratios to incidence densities, *Statistics in Medicine*, 13, 343–355.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics*, 33, 133–158.

Mehta, C. R. and Patel, N. R. (1995), "Exact logistic regression: theory and examples", *Statistics in Medicine*, 14, 2143–1260

Nelder, J.A., and Wedderburn, R.W.M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society A*, 135, 370–384.

Preisser, J. S., and Quaqish, B. F. (1996), Deletion diagnostics for generalised estimating equations, *Biometrika*, 83, 3, 551–562

Preisser, J. S., and Koch, G. G., (1997), "Categorical Data Analysis in Public Health", *Annual Review of Public Health*, 18, 51–82

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

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes, *Biometrics*, 42 121–130

Author

Maura E. Stokes, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. FAX (919) 677-4444 Email sasmzs@wnt.sas.com

SAS is a registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Version 1.0