

ANALYSIS OF INCOMPLETE CATEGORICAL REPEATED MEASURES

Charles S. Davis, University of Iowa

Abstract

The SAS® CATMOD procedure is a powerful and general tool for the analysis of categorical data. Many widely-known and routinely-used applications are described in the *SAS/STAT User's Guide* and *SUGI Proceedings*. However, the use of PROC CATMOD in analyzing repeated measures categorical data is less well known. In particular, existing documentation has focused on the somewhat unrealistic situation in which the data are complete. In practice, the occurrence of missing data is quite common in repeated measures studies. This paper describes and demonstrates the use of PROC CATMOD in analyzing incomplete categorical repeated measures data. An extension of the methodology for testing whether or not subjects with incomplete data are adequately characterized by the same model as subjects with complete data is also discussed.

1. Introduction

Various methods for the analysis of categorical data have been developed and studied during the past 30 years. Agresti (1990) gives a comprehensive survey of the field. Many of these techniques are available to SAS users through the CATMOD and FREQ procedures. In particular, stratified data analysis using generalized Cochran-Mantel-Haenszel statistics, analysis of log-linear and logit models using maximum likelihood, analysis of linear and nonlinear response functions using weighted least squares, and techniques for the analysis of ordinal categorical data can all be carried out. General discussions and examples of these types of categorical data analyses using SAS software have been given by Davis (1982), Woolson, Clarke, and Amini (1982), Imrey (1985), and Stanish (1986), among others.

More recently, methods for the analysis of repeated measures categorical data have been studied. However, the lack of readily-accessible and reliable software has discouraged the routine use of many of these methods. Fortunately, the CATMOD procedure can be used to fit linear models to correlated functions of response probabilities. This general approach, which was first described for repeated measures by Koch *et al.* (1977), is based on the weighted least squares (WLS) methodology of Grizzle, Starmer, and Koch (1969). Applications of SAS software to repeated measures categorical data have been given by Stanish (1984), Landis *et al.* (1988), Pendergast and Littell (1989), Agresti (1989), and Agresti, Lipsitz, and Lang (1991).

Although repeated measures studies are usually designed to collect data from every subject at each scheduled measurement time, the occurrence of missing data is often unavoidable. Provided that the missing data process is unrelated to the observed or unobserved responses, i.e., that data are missing completely at random (MCAR, using the terminology of Rubin, 1976), the WLS methodology can be applied to incomplete repeated measures data. This application was first discussed by Stanish, Gillings, and Koch (1978), and was later extended by Woolson and Clarke (1984) and Landis *et al.* (1988).

The purpose of this paper is to demonstrate the use of the CATMOD procedure in analyzing incomplete categorical repeated measures data. In order to simplify the exposition, the methodology and SAS statements will be described for a specific example. Extensions to more complicated situations will then be discussed. Since the MCAR assumption is a strong one, a simple modification for testing whether subjects with complete data are in some sense "different" from subjects with incomplete data will also be described.

2. Example

The Iowa 65+ Rural Health Study is sponsored by the U.S. National Institute on Aging as part of the "Established Populations for Epidemiologic Study of the Elderly" project (Cornoni-Huntley *et al.*, 1986). In this study, a cohort of elderly individuals was followed over a six-year period. At each of three

surveys, extensive demographic and social support data were obtained from each respondent.

One of the variables of interest to the investigators was church attendance. Table 1 gives the resulting data from 1935 females and 1150 males included in the study. At each survey, the response was classified as

Yes: subject is a regular church attendee;

No: subject is not a regular church attendee;

Missing: subject did not respond to this question or did not participate in this survey.

There are a substantial number of missing values, most of which occur at the end of a sequence of nonmissing responses. This is almost entirely due to deaths or losses to follow-up. We would like to determine if the church attendance rates differ between females and males, if the attendance rates are changing over time, and if the observed patterns of change are the same for females and males.

Table 1: Church Attendance Data from the Iowa 65+ Rural Health Study

Regular Church Attender at:			Frequency		
Year 0	Year 3	Year 6	Female	Male	Total
Missing	No	Missing	3	2	5
Missing	No	No	1	3	4
Missing	No	Yes	1	1	2
Missing	Yes	Missing	2	2	4
Missing	Yes	Yes	2	0	2
No	Missing	Missing	101	122	223
No	Missing	No	11	5	16
No	Missing	Yes	3	2	5
No	No	Missing	71	86	157
No	No	No	158	143	301
No	No	Yes	30	18	48
No	Yes	Missing	14	5	19
No	Yes	No	22	21	43
No	Yes	Yes	33	15	48
Yes	Missing	Missing	195	125	320
Yes	Missing	No	4	0	4
Yes	Missing	Yes	18	9	27
Yes	No	Missing	28	16	44
Yes	No	No	51	26	77
Yes	No	Yes	25	12	37
Yes	Yes	Missing	170	110	280
Yes	Yes	No	88	36	124
Yes	Yes	Yes	904	391	1295

Thus, we are interested in modelling the marginal probability of regular church attendance as a function of gender and survey year. Let π_{j0} , π_{j3} , and π_{j6} denote the marginal probabilities of regular church attendance in females at survey years 0, 3, and 6, respectively, and let π_{m0} , π_{m3} , and π_{m6} denote the corresponding probabilities for males. If there were no missing data, the standard repeated measures capabilities of PROC CATMOD could be used to fit linear models of the form $\pi = X\beta$, where $\pi' = (\pi_{j0}, \pi_{j3}, \pi_{j6}, \pi_{m0}, \pi_{m3}, \pi_{m6})$, X is a $6 \times v$ model matrix ($v \leq 6$), and β is a $v \times 1$ vector of parameters to be estimated using weighted least squares. Such applications of CATMOD are described in the *SAS/STAT User's Guide* and in the references cited in Section 1.

In this example, one possibility would be to exclude all subjects with a missing response at any one of the three surveys and apply the standard methods for complete repeated measures. Unfortunately, of the 3085 subjects in the study, only 1973 (64%) have complete data. Thus, it seems prefer-

able to use the multivariate ratio methods studied by Stanish, Gillings, and Koch (1978), Woolson and Clarke (1984), and Landis *et al.* (1988).

3. WLS Analysis of Incomplete Categorical Data

The analysis procedure can be described in four steps. Each is illustrated with reference to the example described in Section 2.

a. Structure the Data Set for Repeated Measures Analysis

The SAS data set should be structured to have one observation per subject (rather than one observation per time point per subject). For example, if there are t time points, the data set should contain t response variables in addition to any covariates. With reference to the repeated measures terminology from the GLM documentation in the *SAS/STAT User's Guide*, the structure should be multivariate rather than univariate. In this example, an appropriate structure would consist of 3085 observations and four variables (gender, response at years 0, 3, and 6).

When dealing with categorical data, it is usually more convenient to have one observation for each unique combination of the cross-classification of possible covariate and response variable values. A weight variable is then used to indicate the number of subjects in each possible category of the joint cross-classification. This structuring of the data file is similar to the way categorical data are generally tabulated. Using this approach, the following SAS statements create the data set to be used in the analysis of the example:

```
data church;
* gender: F=female, M=male;
* attend: M=missing, N=no, Y=yes;
input (gender attend0 attend3 attend6 count)
($1. +1 $1. +1 $1. +1 $1. 4.);
cards;
F M M M 3
F M M M 1
...
M Y Y M 36
M Y Y Y 391
;
```

Note that the data are organized in essentially the same fashion as displayed in Table 1.

b. Characterize the Response Profiles

Regardless of the number of covariates and response variables, the CATMOD procedure analyzes data represented in a two-way contingency table. The rows and columns of this underlying table are defined by the population profiles and the response profiles, respectively. In most applications, the default capabilities of CATMOD make it unnecessary for the user to give explicit attention to this underlying table. However, a special understanding of the response profiles is necessary when analyzing incomplete repeated measures data.

Let c denote the number of possible outcomes of a response variable measured at t time points. If there are no missing data, there are c^t potential response profiles. In our example, $c = 2$ (no, yes), $t = 3$, and there are $2^3 = 8$ response profiles for the complete data. However, if "missing" is a potential response, there are now $c + 1$ possible outcomes at each time point and the number of potential response profiles is $(c + 1)^t - 1$ (since a subject who had missing data at all time points would be excluded). In the church attendance example, there are $3^3 - 1 = 26$ possible response profiles.

The difficulty in analyzing repeated measures categorical data is that, in practice, some of the potential response profiles do not occur. Since the response functions must be explicitly defined using the RESPONSE statement, the user must first determine the number and ordering of the response profiles. This can be done most simply using PROC FREQ. In this example, the statements

```
proc freq; weight count;
tables attend0*attend3*attend6 / list;
```

produce the frequency distribution given in the last column of Table 1. It can then be seen that only 23 of the potential 26 response profiles were observed.

c. Define the Response Functions of Interest

The most common response functions of interest in repeated measures analyses are the marginal proportions at each of the t time points. (Other types of response functions will be discussed in Section 4.) For an outcome variable with c categories, the vector p_i of marginal proportions from the i th population contains $(c - 1)t$ linearly independent elements. Although any one of the c categories could be excluded, we will consider the first $c - 1$ values of the outcome variable. Note that in the special case of a dichotomous outcome, there are t marginal proportions from each population.

For complete data, the statement RESPONSE MARGINALS can be used to compute p_i . However, when the data are incomplete, the components of p_i must be calculated as ratios of sums of the multinomial proportions corresponding to the response profiles. In PROC CATMOD, these operations are specified as a series of linear, logarithmic, and exponential transformations of the elements of the vector p_i^* of multinomial proportions corresponding to the response profiles. In general,

$$p_i = \exp(A_2 \log(A_1 p_i^*)). \quad (1)$$

The matrix A_1 has ct rows and as many columns as there are observed response profiles (a maximum of $(c + 1)^t - 1$); A_2 is a $(c - 1)t \times ct$ matrix. It is convenient to define the rows of A_1 as shown in Table 2. With this definition of A_1 ,

$$A_2 = I_t \otimes [I_{(c-1)}, -e_{(c-1)}],$$

where I_k is the $k \times k$ identity matrix, e_k is the $k \times 1$ vector $(1, \dots, 1)'$, and $\otimes$ denotes the Kronecker product operator.

Table 2: Rows of A_1 for Computing Marginal Proportions

Row	Proportion of subjects with:
1	response category 1 at time 1
2	response category 2 at time 1
$\vdots$	$\vdots$
$c - 1$	response category $c - 1$ at time 1
c	non-missing response at time 1
.....	
$c(t - 1) + 1$	response category 1 at time t
$c(t - 1) + 2$	response category 2 at time t
$\vdots$	$\vdots$
$c(t - 1) + (c - 1)$	response category $c - 1$ at time t
$c(t - 1) + c$	non-missing response at time t

In the example from the Iowa 65+ Study, the marginal proportion vectors from the two populations (females, males) are

$$p_1' = (p_{f0}, p_{f3}, p_{f6}), \quad p_2' = (p_{m0}, p_{m3}, p_{m6}).$$

The 6×23 matrix A_1 is given by

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

For each population, rows 1, 3, and 5 compute the proportion of individuals who regularly attended church at years 0, 3, and 6,

respectively. Rows 2, 4, and 6 calculate the corresponding proportions who responded to this question at the three surveys. A_2 is the 3×6 matrix

$$I_3 \otimes (1, -1) = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The SAS statements for invoking PROC CATMOD and defining these response functions are:

```
proc catmod; weight count; population gender;
response exp
    1 -1 0 0 0 0,
    0 0 1 -1 0 0,
    0 0 0 0 1 -1
log 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1,
    0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1,
    0 0 0 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 1,
    1 1 1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 1 1,
    0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1,
    0 1 1 0 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1;
```

d. Fit Appropriate Models

The model-fitting procedure is essentially no different than in any other CATMOD application. The response variables at the t time points are listed on the left-hand side of the MODEL statement (separated by asterisks). In repeated measures studies it is usually not convenient to use named effects to define the columns of the model matrix X . Instead, the model is defined explicitly on the right-hand side of the MODEL statement.

In this example, we will first fit a model with separate intercepts and linear and quadratic time effects for females and males. Since the surveys were equally spaced, the linear and quadratic time effects will be coded using orthogonal polynomial coefficients. The required CATMOD statements are

```
model attend0*attend3*attend6=
    (1 -1 1 0 0 0,
    1 0 -2 0 0 0,
    1 1 1 0 0 0,
    0 0 0 1 -1 1,
    0 0 0 1 0 -2,
    0 0 0 1 1 1)
(1='Females: Intercept',
 2=' Linear',
 3=' Quadratic',
 2 3=' Lin. & Quad.',
 4='Males: Intercept',
 5=' Linear',
 6=' Quadratic',
 5 6=' Lin. & Quad.',
 2 5='Both: Linear',
 3 6=' Quadratic',
 2 3 5 6=' Lin. & Quad.') / noprofile;
```

and the results are displayed in Table 3.

Before interpreting the results of Table 3, it is important to verify that the response functions have been calculated correctly. The three response functions for sample 1 (2) should be equal to the observed proportions of females (males) regularly attending church at years 0, 3, and 6, respectively. For example, it can be seen from Table 1 that 682 males responded to the church attendance question at year 6. Since 448 of these individuals were regular church attenders, response function 3 in sample 2 is $448/682=0.65689$, which agrees with Table 3. It is also important to verify that the columns of the model matrix have been specified and interpreted correctly.

The results of this model indicate that the two degree of freedom (df) time effect is not significantly different from zero in females and in males (as well as jointly). Taking a conservative approach, we will first fit a reduced model with separate intercepts and linear time effects for males and females. Since the reduced model provides a good fit to the observed data (the residual chi-square is 1.83 with 2 df), and since the time effect is

Table 3: Partial Output from Saturated Marginal Probability Model

Sample	Function Number	Response Function	DESIGN MATRIX					
			1	2	3	4	5	6
1	1	0.76999	1	-1	1	0	0	0
	2	0.77043	1	0	-2	0	0	0
	3	0.75204	1	1	1	0	0	0
2	1	0.63485	0	0	0	1	-1	1
	2	0.65389	0	0	0	1	0	-2
	3	0.65689	0	0	0	1	1	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
Females: Intercept	1	7595.57	0.0000
Linear	1	2.44	0.1183
Quadratic	1	1.36	0.2442
Lin. & Quad.	2	3.28	0.1943
Males: Intercept	1	2259.45	0.0000
Linear	1	1.75	0.1860
Quadratic	1	0.47	0.4932
Lin. & Quad.	2	2.73	0.2558
Both: Linear	2	4.19	0.1232
Quadratic	2	1.83	0.4014
Lin. & Quad.	4	6.00	0.1989
RESIDUAL	0	.	.

not significant in females, males, or jointly, we fit a final model with separate intercepts for females and males. The CATMOD statements are

```
model attend0*attend3*attend6=
    (1 0,
    1 0,
    1 0,
    0 1,
    0 1,
    0 1)
(1='Females: Intercept',
 2='Males: Intercept') / noprofile;
contrast 'Int. Equality' all_parms 1 -1;
```

and the results are displayed in Table 4.

This model provides a satisfactory fit to the observed data ($p = 0.20$). The probability of regular church attendance is estimated to be 0.77 for females and 0.64 for males. Although the gender difference is highly significant, there is no statistically significant evidence of a time trend in either of the two sexes.

4. Extensions

It is not difficult to extend this methodology to more complicated response functions such as generalized logits or cumulative logits. The simplest approach is to first calculate the marginal proportions using the transformation defined by equation (1). However, instead of computing only the first $c-1$ linearly independent marginal proportions, it will usually be necessary to calculate all c proportions. Thus, A_1 will have $(c+1)t$ rows and A_2 will be the $ct \times (c+1)t$ matrix $I_t \otimes [I_c, -e_c]$. Additional matrix, exponential, and logarithmic operations can then be applied to obtain the response functions of interest.

For example, consider an outcome variable with three ordered values (none, moderate, severe) measured at three time points. Suppose that the response functions of interest are the cumulative logits

$$l_{j1} = \log \left(\frac{\text{Pr(none) at time } j}{\text{Pr(moderate or severe) at time } j} \right),$$

$$l_{j2} = \log \left(\frac{\text{Pr(none or moderate) at time } j}{\text{Pr(severe) at time } j} \right),$$

Table 4: Output from Final Marginal Probability Model

Sample	Function Number	Response Function	DESIGN MATRIX	
			1	2
1	1	0.76999	1	0
	2	0.77043	1	0
	3	0.75204	1	0
2	1	0.63485	0	1
	2	0.65389	0	1
	3	0.65689	0	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
Females: Intercept	1	7926.90	0.0000
Males: Intercept	1	2355.15	0.0000
RESIDUAL	4	6.00	0.1989

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
MODEL	1	0.7660	0.00860	7926.90	0.0000
	2	0.6434	0.0133	2355.15	0.0000

ANALYSIS OF CONTRASTS

Contrast	DF	Chi-Square	Prob
Int. Equality	1	60.12	0.0000

where l_{jk} is the k th cumulative logit at time j , for $j = 1, 2, 3$, $k = 1, 2$. In the case of a single subpopulation, the vector $l = (l_{11}, l_{12}, l_{21}, l_{22}, l_{31}, l_{32})'$ can be computed from the transformation

$$l = A_4 \log[A_3 \exp(A_2 \log(A_1 p^*))].$$

The matrix A_1 will have 12 rows and at most $4^3 - 1 = 63$ columns, while the 9×12 matrix A_2 , the 12×9 matrix A_3 , and the 6×12 matrix A_4 are given by

$$A_2 = I_3 \otimes [I_3, -e_3],$$

$$A_3 = I_3 \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$A_4 = I_3 \otimes \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

In many cases, the first step of computing the complete set of marginal proportions using the transformation (1) can be eliminated and the RESPONSE statement simplified. With reference to the church attendance example, suppose we preferred to carry out the analysis on the logit scale rather than on the probability scale. In this case, the appropriate vector of response functions is the 6×1 vector of logits $l = (l_1, l_2)'$, where

$$l_1 = \left(\log\left(\frac{p_{f0}}{1 - p_{f0}}\right), \log\left(\frac{p_{f3}}{1 - p_{f3}}\right), \log\left(\frac{p_{f6}}{1 - p_{f6}}\right) \right),$$

$$l_2 = \left(\log\left(\frac{p_{m0}}{1 - p_{m0}}\right), \log\left(\frac{p_{m3}}{1 - p_{m3}}\right), \log\left(\frac{p_{m6}}{1 - p_{m6}}\right) \right).$$

This set of transformations can be defined as $l'_i = B_2 \log(B_1 p'_i)$, where

$$B_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

and $B_2 = I_3 \otimes (1, -1)$. Rows 1, 3, and 5 of B_1 compute the proportion of subjects who regularly attended church at years 0, 3, and 6, respectively, while rows 2, 4, and 6 compute the corresponding proportions who reported that they were not regular attenders. Since the denominator of each marginal proportion cancels when the ratio is computed, it is not necessary to calculate the proportion of subjects with non-missing responses. The resulting RESPONSE statement is:

```
response 1 -1 0 0 0 0,
          0 0 1 -1 0 0,
          0 0 0 0 1 -1
log 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1,
    0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0,
    0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1,
    1 1 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0,
    0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0,
    0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0;
```

5. A Test of the Missing Data Mechanism

Based on the ratio estimation strategy initially proposed by Stanish, Gillings, and Koch (1978), the results from the final model in Section 3 indicate that the marginal probability of regular church attendance differs between males and females, but that there is no evidence of changes over time. If instead we analyze only the 1973 subjects with complete data, the results are much different. First, the initial saturated model with separate intercepts and linear and quadratic time effects for females and males yielded highly significant linear and nonlinear time effects for both genders. Since there was no evidence of differing time trends between males and females, the final model included separate intercepts for females and males, as well as common linear and quadratic time effects. This model provided a very good fit to the complete data, in that the residual chi-square was 0.87 with 2 df. In addition, the four parameters in the model (intercept for females, intercept for males, linear time effect, quadratic time effect) were all significantly different from zero. The conclusion from the analysis of the complete data is that the probability of regular church attendance is decreasing over time and that the decrease from year 0 to year 3 is less than the decrease from year 3 to year 6.

It is possible to fit a single model to the complete and incomplete data and to formally test if the parameter estimates for the complete data are significantly different (individually and jointly) from the parameter estimates for the incomplete data. In this case, the 12×1 vector of response functions is $p = (p_1, p_2)'$, where

$$p_1 = (p_{f0c}, p_{f3c}, p_{f6c}, p_{f0i}, p_{f3i}, p_{f6i}),$$

$$p_2 = (p_{m0c}, p_{m3c}, p_{m6c}, p_{m0i}, p_{m3i}, p_{m6i}),$$

where the third subscript is "c" for subjects with complete data and "i" for subjects with one or more missing responses. The vector p_i can be computed as $\exp(A_2 \log(A_1 p^*_i))$, where

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$A_2 = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The first four rows of A_1 deal with subjects who responded to all three surveys. Rows 1-3 compute the proportion who attended church regularly at years 0, 3, and 6, respectively, while row 4 computes the proportion who responded at all three surveys. The last six rows of A_1 pertain to subjects who had at least one missing response. Rows 5 and 6 compute the numerator (attend church regularly) and denominator (responded to the survey) for year 0, while rows 7-8 and 9-10 calculate the corresponding quantities for year 3 and year 6.

We can now fit a saturated model with separate intercepts and linear and quadratic time effects for four groups of subjects: "complete" females, "incomplete" females, "complete" males, and "incomplete" males. A 6 df contrast tests whether the parameter estimates for the complete cases differ significantly from those for the incomplete cases. Additional contrasts of interest can also be tested. The CATMOD statements are given in Table 5.

The results (not shown) indicate a highly significant difference between complete and incomplete cases (chi-square=82.69 with 6 df). However, the four df test of the linear and quadratic time effects indicates no difference between complete and incomplete cases for these parameters (chi-square=1.58). A second model incorporated common linear and quadratic time effects for complete and incomplete cases and provided a good fit to the data (residual chi-square=1.58, df=4). In addition, the time effects for males and females were not significantly different.

Thus, the final model included separate intercepts for complete and incomplete females and males (four parameters), and a common linear and quadratic time effect (two parameters). The CATMOD statements and output are given in Tables 6 and 7, respectively. This model provided a very good fit to the observed data (residual chi-square=2.58 with 6 df). The time trend for all four groups of subjects was the same (a nonlinear decrease in the probability of regular church attendance from year 0 to year 3 to year 6). Females who responded at all three surveys had the highest initial rate of regular church attendance (.79), followed by complete males (.69), incomplete females (.65), and incomplete males (.52). Note that the conclusions pertaining to time trends are totally different from those obtained from the usual approach to the analysis of incomplete repeated measures.

6. Discussion

This paper has focused on the WLS approach to the analysis of repeated measures categorical data. Although this was the first general method for analyzing repeated categorical outcomes, and many new methods have since been proposed, it offers several advantages. From a practitioner's point of view, the chief strength is undoubtedly its availability and ease of use through the SAS system. In addition, missing data can be accommodated and it is even possible to test for differences between complete and incomplete cases.

In spite of these strengths, some disadvantages should also be noted. First and foremost, it is a large-sample method. It can only be used in studies involving a large sample size (number of subjects), a relatively limited number of time points, and relatively few covariates each with a limited number of possible outcomes. Another shortcoming of this methodology is that all covariates must be categorical.

7. Acknowledgements

This research was supported in part by Grant MH 46011 from the National Institute of Mental Health and by Contract AG 2106 from the National Institute on Aging. SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Table 5: Statements for Saturated Model Comparing Complete and Incomplete Cases

```
proc catmod; weight count; population gender;
response exp
    1 0 0 -1 0 0 0 0 0 0 0,
    0 1 0 -1 0 0 0 0 0 0 0,
    0 0 1 -1 0 0 0 0 0 0 0,
    0 0 0 0 1 -1 0 0 0 0 0,
    0 0 0 0 0 0 1 -1 0 0 0,
    0 0 0 0 0 0 0 0 1 -1 0,
    0 0 0 0 0 0 0 0 0 1 -1
log 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1,
    0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1,
    0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1,
    0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 1 0 1 1,
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0,
    0 0 0 0 0 1 1 1 1 0 0 1 0 0 1 1 1 1 1 0 0 1 0 0,
    0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0,
    1 1 1 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0,
    0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0,
    0 1 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0;
model attend0*attend3*attend6=
    (1 -1 1 0 0 0 0 0 0 0 0 0 0 0,
    1 0 -2 0 0 0 0 0 0 0 0 0 0 0,
    1 1 1 0 0 0 0 0 0 0 0 0 0 0,
    0 0 0 0 0 0 0 1 -1 1 0 0 0 0,
    0 0 0 0 0 0 0 1 0 -2 0 0 0 0,
    0 0 0 0 0 0 0 1 1 1 0 0 0 0,
    0 0 0 1 -1 1 0 0 0 0 0 0 0 0,
    0 0 0 1 0 -2 0 0 0 0 0 0 0 0,
    0 0 0 1 1 1 0 0 0 0 0 0 0 0,
    0 0 0 0 0 0 0 0 0 0 0 1 -1 1,
    0 0 0 0 0 0 0 0 0 0 0 1 0 -2,
    0 0 0 0 0 0 0 0 0 0 0 1 1 1);
(1='C: F-Int.',
2='    Lin.',
3='    Quad',
4='    M-Int.',
5='    Lin.',
6='    Quad',
7='IC: F-Int.',
8='    Lin.',
9='    Quad',
10='    M-Int.',
11='    Lin.',
12='    Quad') / noprofile;
contrast 'C=IC'
    all_parms 1 0 0 0 0 0 0 -1 0 0 0 0 0,
    all_parms 0 1 0 0 0 0 0 0 -1 0 0 0 0,
    all_parms 0 0 1 0 0 0 0 0 0 -1 0 0 0,
    all_parms 0 0 0 1 0 0 0 0 0 0 -1 0 0,
    all_parms 0 0 0 0 1 0 0 0 0 0 0 -1 0,
    all_parms 0 0 0 0 0 1 0 0 0 0 0 0 -1;
contrast 'C=IC: L & Q'
    all_parms 0 1 0 0 0 0 0 0 -1 0 0 0 0,
    all_parms 0 0 1 0 0 0 0 0 0 -1 0 0 0,
    all_parms 0 0 0 1 0 0 0 0 0 0 -1 0,
    all_parms 0 0 0 0 1 0 0 0 0 0 0 -1;
```

8. References

- Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statist. Med.* 8, 1209-1224.
- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.
- Agresti, A., Lipsitz, S., and Lang, J. B. (1991). Analysis of sparse repeated categorical measurement data. *Proceedings of the 16th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.
- Cornoni-Huntley, J., Brock, D. B., Ostfeld, A., Taylor, J. O., and Wallace, R. B. (1986). *Established Populations for Epidemiologic Studies of the Elderly, Resource Data Book*. National Institutes of Health, NIH Pub. No. 86-2443.

Table 6: Statements for Final Model
Comparing Complete and Incomplete Cases

```
proc catmod; weight count; population gender;
response exp
    1 0 0 -1 0 0 0 0 0 0,
    0 1 0 -1 0 0 0 0 0 0,
    0 0 1 -1 0 0 0 0 0 0,
    0 0 0 0 1 -1 0 0 0 0,
    0 0 0 0 0 0 1 -1 0 0,
    0 0 0 0 0 0 0 0 1 -1
log 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1,
    0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1,
    0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1,
    0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 1 0 1 1,
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 1 0 0,
    0 0 0 0 0 1 1 1 1 0 0 1 0 0 1 1 1 1 1 0 0 1 0 0,
    0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0,
    1 1 1 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0,
    0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0,
    0 1 1 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0;
model attend0*attend3*attend6=
    (1 0 0 0 -1 1,
    1 0 0 0 0 -2,
    1 0 0 0 1 1,
    0 1 0 0 -1 1,
    0 1 0 0 0 -2,
    0 1 0 0 1 1,
    0 0 1 0 -1 1,
    0 0 1 0 0 -2,
    0 0 1 0 1 1,
    0 0 0 1 -1 1,
    0 0 0 1 0 -2,
    0 0 0 1 1 1)
(1='F: C Int.',
2=' IC Int.',
3='M: C Int.',
4=' IC Int.',
5='Linear Time',
6='Quadratic Time') / noprofile;
contrast 'Int. Equality' all_parms 1 0 0 -1 0 0,
    all_parms 0 1 0 -1 0 0,
    all_parms 0 0 1 -1 0 0;
```

Table 7: Output from Final Model
Comparing Complete and Incomplete Cases

Sample	Function Number	Response Function	DESIGN MATRIX					
			1	2	3	4	5	6
1	1	0.81465	1	0	0	0	-1	1
	2	0.79863	1	0	0	0	0	-2
	3	0.75667	1	0	0	0	1	1
	4	0.67480	0	1	0	0	-1	1
	5	0.64384	0	1	0	0	0	-2
	6	0.60000	0	1	0	0	1	1
2	1	0.70242	0	0	1	0	-1	1
	2	0.69940	0	0	1	0	0	-2
	3	0.65861	0	0	1	0	1	1
	4	0.54167	0	0	0	1	-1	1
	5	0.52000	0	0	0	1	0	-2
	6	0.60000	0	0	0	1	1	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
F: C Int.	1	6647.83	0.0000
IC Int.	1	1239.44	0.0000
M: C Int.	1	1813.32	0.0000
IC Int.	1	531.98	0.0000
Linear Time	1	40.70	0.0000
Quadratic Time	1	4.92	0.0266
RESIDUAL	6	2.58	0.8590

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
MODEL	1	0.7904	0.0097	6647.83	0.0000
	2	0.6485	0.0184	1239.44	0.0000
	3	0.6864	0.0161	1813.32	0.0000
	4	0.5202	0.0226	531.98	0.0000
	5	-0.0269	0.00421	40.70	0.0000
	6	-0.00445	0.00200	4.92	0.0266

ANALYSIS OF CONTRASTS

Contrast	DF	Chi-Square	Prob
Int. Equality	3	150.89	0.0000

Davis, C. S. (1982). Categorical data analysis using the FUN-CAT procedure: an example. *Proceedings of the 7th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 718-723.

Grizzle, J. E., Stamer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* 25, 489-504.

Imrey, P. B. (1985). SAS software for log-linear models. *Proceedings of the 10th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 1006-1019.

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.

Landis, J. R., Miller, M. E., Davis, C. S., and Koch, G. G. (1988). Some general methods for the analysis of categorical data in longitudinal studies. *Statistics in Medicine* 7, 109-137.

Pendergast, J. F. and Littell, R. C. (1989). Issues related to analyzing repeated measurement data. *Proceedings of the 14th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 169-177.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

Stanish, W. M. (1984). The use of CATMOD for repeated measurement analysis of categorical data. *Proceedings of the 9th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 761-770.

Stanish, W. M. (1986). Categorical data analysis strategies using SAS software. *Computer Science and Statistics: Proceedings of the 17th Symposium on the Interface*. New York: Elsevier Science Publishers, 239-256.

Stanish, W. M., Gillings, D. B., and Koch, G. G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. *Biometrics* 34, 305-317.

Woolson, R. F. and Clarke, W. R. (1984). Analysis of categorical incomplete longitudinal data. *J. R. Statist. Soc. A* 147, 87-99.

Woolson, R. F., Clarke, W. R., and Amini, S. B. (1982). Applications of SAS (FUNCAT) to the analysis of epidemiological/survivorship data. *Proceedings of the 7th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 739-747.

Correspondence: Charles S. Davis, Ph.D.
Division of Biostatistics
University of Iowa College of Medicine
2837 Steindler Building
Iowa City, IA 52242
(319) 335-9625
chuck@davis.pmeh.uiowa.edu