BIOMETRICS 55, 936–943
September 1999

Modeling a Categorical Variable Allowing Arbitrarily Many Category Choices

Alan Agresti

Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.
e-mail: aa@stat.ufl.edu

and

I-Ming Liu

Department of Statistics, National Chung Hsing University, Taipei, Taiwan, R.O.C.

SUMMARY. This article discusses the modeling of a categorical variable for which subjects can select any number of categories. For c categories, the response variable consists of a cross-classification of c binary components, one pertaining to each category. Using data from a survey (Loughin, T. M. and Scherer, P. N., 1998, *Biometrics*, **54,** 630–637) in which Kansas farmers indicated their primary sources of veterinary information, we discuss simultaneous logit modeling of the binary components of the multivariate response. The use of maximum likelihood or quasi-likelihood fitting provides chi-squared tests with degrees of freedom df $= c(r-1)$ for testing the independence between each of the c response components and an explanatory variable with r categories. These tests are alternatives to the weighted chi-squared test and the bootstrap test proposed by Loughin and Scherer for this hypothesis.

KEY WORDS: Binary data; Chi-squared test; Generalized estimating equations; Logit model; Log-linear model; Marginal model; Repeated categorical responses; Surveys.

1. Introduction

Table 1, from an interesting article by Loughin and Scherer (1998), gives the data from a study that asked a sample of 262 Kansas livestock farmers, "What are your primary sources of veterinary information?" The outcome categories are (A) professional consultant, (B) veterinarian, (C) state or local extension service, (D) magazines, and (E) feed companies and representatives. The farmers were asked to select all categories that were relevant. Table 1 provides response counts when these categories are cross-classified with the farmer's achieved education. This 5 × 5 contingency table contains 453 positive responses from the 262 farmers. Table 1 also lists the sample proportion of farmers who chose each outcome category for each educational level.

Loughin and Scherer (1998) developed a large-sample weighted chi-squared test and a small-sample bootstrap test for the hypothesis that the probability of selecting any given veterinary information source is identical among the five educational levels. They noted that ordinary statistical inference was inappropriate with Table 1, because its 453 entries were not independent. For instance, it is incorrect to apply the chi-squared distribution with the ordinary Pearson statistic to Table 1 to test this hypothesis. In this article, we present valid chi-squared tests for this hypothesis as well as models to describe these data.

At each educational level, there are 2^5 possible response sequences, according to the (yes, no) outcome for the selection of each outcome category. Thus, the response is most fully viewed as a cross-classification of five binary components: variable A indicating whether a farmer said 'yes' to source A, variable B indicating whether a farmer said 'yes' to source B, and so forth. The complete data set is the 5×2^5 contingency table showing the counts of the possible response sequences at each level of X = education. Proper analyses use the data in this form rather than in the form of Table 1. Loughin and Scherer (1998) also exhibited their data in the form of this contingency table, which is referred to as the complete table.

Table 1 is marginal to the complete table, referring to counts of the 'yes' responses in the marginal distributions of the components at each educational level. The hypothesis that Loughin and Scherer tested is that of simultaneous marginal independence between X and A, X and B, X and C, X and D, and X and E in five of the two-way marginal tables of the complete table. As this hypothesis refers to marginal tables, it does not specify the joint distribution for the complete table. Thus, inference about this hypothesis must deal separately with specification of that joint distribution.

Our article discusses the modeling of categorical variables that have multiple potential responses. We demonstrate the use of existing methodology to model directly the marginal distributions of the multivariate response, i.e., the distribu-

Table 1
Veterinary information sources and education groups for 262 farmers

	Information source					Total number of	Total number of
Education	A	B	C	D	E	responses	subjects
High school	19 (0.22)	38 (0.43)	29 (0.33)	47 (0.53)	40 (0.46)	173	88
Vocational	2 (0.12)	6 (0.38)	8 (0.50)	8 (0.50)	4 (0.25)	28	16
2-year college	1 (0.03)	13 (0.42)	10 (0.32)	17 (0.55)	14 (0.45)	55	31
4-year college	19 (0.17)	29 (0.26)	40 (0.35)	53 (0.47)	29 (0.26)	170	113
Others	3 (0.21)	4 (0.29)	8 (0.57)	6 (0.43)	6 (0.43)	27	14
Total	44	90	95	131	93	453	262

Source: Loughin and Scherer (1998); value in parentheses is the proportion selecting that category.

tions for which Table 1 summarizes counts for the 'yes' category. Section 2 discusses logit models of this type. One can use maximum likelihood (ML) or quasi-likelihood (e.g., generalized estimating equations) methods to estimate parameters of the marginal logit models. Section 2 also discusses effects of table sparseness. As a by-product of the modeling, in Section 3 we show how to construct large-sample chi-squared tests for the hypothesis of marginal independence between each component of a c-category response and an r-category predictor. These model-based test statistics have a degree of freedom df $= r(c-1)$. Section 3 also presents and compares other large-sample and small-sample tests for this hypothesis including the tests suggested by Loughin and Scherer. An advantage of the test statistics proposed here is antivariance whether the responses refer to "check all categories that apply" or to "check all categories that do not apply." Section 4 deals with extensions of the models and alternative modeling approaches.

2. The Marginal Logit Model Approach

For concreteness, in formulating models we refer to Table 1, in which there is a categorical explanatory variable and a categorical outcome variable with potentially multiple responses. We denote the explanatory variable, e.g., education, by X, having r levels, and the outcome variable by Y, having c binary components. We refer to the c binary variables that constitute Y as *items.* It is usually natural to assume an independent multinomial distribution for the counts in each of the r subtables of size 2^c of the complete table. For a randomly selected subject at level i of X, let $\pi_{j|i}$ denote the probability of responding with a 'yes' on the jth item (i.e., including category j in the set of chosen categories). Then $\{(\pi_{j|i}, 1-\pi_{j|i}), j = 1, \ldots, c\}$ are the c marginal distributions for the 2^c cross-classification of responses when $X = i$. Also note that $0 \leq \Sigma_j \pi_{j|i} \leq c$.

2.1 *Multiple Marginal Independence Model*

The marginal logit model

$$\log\left(\frac{\pi_{j|i}}{1-\pi_{j|i}}\right) = \beta_j, \quad i = 1, \ldots, r, \; j = 1, \ldots, c \qquad (1)$$

states that, for each item j, the probability of responding with a 'yes' for that item is the same at each level of X. We call this the *multiple marginal independence model* for the effect of X on Y, because it states that each component of Y is marginally independent of X. This model is equivalent to the null hypothesis tested by Loughin and Scherer (1998), whose alternative hypothesis is equivalent to the model

$$\log\left(\frac{\pi_{j|i}}{1-\pi_{j|i}}\right) = \beta_{ij}, \quad i = 1, \ldots, r, \; j = 1, \ldots, c. \qquad (2)$$

These models make no assumption about the association structure for the multinomial distributions specifying the joint distribution. Thus, model (2) is the saturated model for the marginal distributions; the observed counts in the complete $r \times 2^c$ table are its fitted values, and the sample marginal logits at each level of X are the ML estimates of $\{\beta_{ij}\}$.

More generally, marginal models can incorporate a set x of explanatory variables. One might then compare a general model

$$\log\left(\frac{\pi_{j|x}}{1-\pi_{j|x}}\right) = \beta_j' x, \quad j = 1, \ldots, c \qquad (3)$$

permitting different marginal distributions for each combination of item and x to simpler models in which effects are constant across levels of certain explanatory variables.

2.2 *Fitting Marginal Models*

In principle, it is not difficult to fit models of form (3), but ML fitting is not readily amenable to ordinary statistical software. Obtaining ML estimates is awkward because even though the models apply to c marginal tables of the complete table, the likelihood refers to multinomial probabilities within the complete table.

To maximize the product multinomial likelihood subject to the constraint that the marginal distributions satisfy a model, one can iteratively use Lagrange's method of undetermined multipliers together with the Newton–Raphson method (Aitchison and Silvey, 1958; Haber, 1985). We applied an algorithm based on a refined method (Lang and Agresti, 1994) in which the matrix inverted in the Newton–Raphson step has a simpler form. This algorithm applies to generalized log-linear models having the matrix form

$$\mathbf{C} \log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}. \qquad (4)$$

In this context $\boldsymbol{\pi}$ is the vector, with $r \times 2^c$ elements, of the r sets of multinomial probabilities, one set for each level of X, and $\boldsymbol{\beta}$ is the vector of model parameters, c for model (1) and rc for model (2). The matrix $\mathbf{A}$ applied to $\boldsymbol{\pi}$ forms the relevant marginal probabilities, and $\mathbf{C}$ applied to the log

Table 2

Item parameter estimates, with standard errors in parentheses, under multiple marginal independence, and the result of the model-comparison test of independence for a marginal logit model fitted by ML and GEE

Parameter	ML (SE)	GEE (SE)
β_1	−1.565 (0.155)	−1.600 (0.165)
β_2	−0.712 (0.127)	−0.648 (0.130)
β_3	−0.498 (0.123)	−0.564 (0.128)
β_4	−0.022 (0.120)	0.000 (0.124)
β_5	−0.630 (0.126)	−0.597 (0.129)
Test of independence (df = 20)	$G^2 = 30.9$, $X^2 = 26.7$	Wald = 33.2

marginal probabilities forms the marginal logits for the models. An S-plus function ('gllm') for the algorithm may be obtained from Prof J. B. Lang of the Statistics Department, University of Iowa.

Let $\mathbf{p}$ denote the vector of sample proportions in the complete table corresponding to the multinomial probabilities $\boldsymbol{\pi}$. Having fitted the model, one can check the goodness of fit by comparing $\mathbf{p}$ or the cell counts to their fitted values $\widehat{\boldsymbol{\pi}}$ using the usual chi-squared statistics. It can be more informative to compare the sample marginal proportions $\mathbf{Ap}$ (or the marginal counts) with their fitted values $\mathbf{A}\widehat{\boldsymbol{\pi}}$, for instance, forming adjusted residuals by taking the ratio of $\mathbf{A}(\mathbf{p} - \widehat{\boldsymbol{\pi}})$ to their asymptotic standard errors (Lang and Agresti, 1994).

An alternative, quasi-likelihood, approach for obtaining estimates in marginal models is that based on generalized estimating equations (GEE) of Liang and Zeger (1986). The GEE approach is computationally simpler than ML for large complete tables that occur with large c values or multiple explanatory variables. With categorical predictors and an unstructured correlation matrix for the joint distribution of the items, this corresponds to iterating the weighted least squares (WLS) approach of Koch et al. (1977) (Miller, Davis, and Landis, 1993). SAS can implement the WLS approach with PROC CATMOD, which is designed to provide WLS fits for models of form (4), and it can implement the GEE approach for a variety of correlation matrix structures with PROC GENMOD.

Table 2 shows the ML estimates of $\{\beta_j\}$ and their standard errors for the multiple marginal independence model (1) applied to the complete data in Table 1. For comparison, Table 3 shows $\{\widehat{\beta}_{ij}\}$ for the saturated model (2). The goodness-of-fit tests of model (1) yield the likelihood-ratio statistic $G^2 = 30.9$ and the Pearson statistic $X^2 = 26.7$, each with df = 20. Both statistics provide weak evidence of lack of fit ($P = 0.06$ and 0.14). The complete table is sparse, having 262 observations in 160 cells, of which 94 are empty; thus, this conclusion is tentative. The test using G^2 tends to be too liberal when fitted values fall between 0.5 and 5. It also tends to be too conservative when fitted values fall below 0.5, yet the complete table has many cells of each type. For the 25 fitted values corresponding to the marginal counts in Table 1, four adjusted residuals have absolute values greater than 2. These result from fitted counts of 5.4, 29.0, 37.2, 30.6, and 39.3 corresponding to the observed counts of 1, 38, 29, 40, and 29 in Table 1; these cells are identified by asterisks in Table 3. (The differences between observed and fitted marginal totals of 'no' responses necessarily yield exactly the negative of the deviations and of the adjusted residuals for the 'yes' responses.)

For most software for the GEE approach, such as SAS, applied with the multiple marginal independence model (1), the procedure analyzes the complete table collapsed over education because of the lack of an education predictor in the model. In that case, for any choice of correlation structure the process converges in a single iteration and yields the relevant sample marginal logits as estimates. From Table 1 with model (1), for instance, $\widehat{\beta}_1 = \log\{44/(262-44)\} = -1.6$. This is also the WLS estimate for the data collapsed over education. Table 2 also shows these estimates of $\{\beta_j\}$ for model (1), which are similar to those obtained with the ML fitting. Table 4 illustrates the use of SAS for these analyses.

2.3 *Other Unsaturated Marginal Models*

One can construct special cases of the saturated model (2) that have multiple marginal independence as further special cases. For instance, the logit model

$$\log\left(\frac{\pi_{j|i}}{1-\pi_{j|i}}\right) = \alpha_i + \beta_j, \quad i = 1, \ldots, r, \; j = 1, \ldots, c \quad (5)$$

permits the marginal probability to vary across levels of both X and Y. Identifiability requires a constraint such as $\alpha_r = 0$. For the Kansas farmer data, this model has fit statistics G^2

Table 3

Parameter estimates for marginal models for Table 1*: The first sets are for saturated model* (2) *and the second sets are for model* (7)

	Information source				
Education	A	B	C	D	E
High school	−1.29, −1.61	−0.27[a], −0.48	−0.71, −0.48	0.14, −0.01	−0.18[a], −0.48
Vocational	−1.95, −1.61	−0.51, −0.48	0.00, −0.48	0.00, −0.01	−1.10, −0.48
2-year college	−3.40[a], −1.61	−0.33, −0.48	−0.74, −0.48	0.19, −0.01	−0.19, −0.48
4-year college	−1.60, −1.61	−1.06[a], −1.06	−0.60, −0.48	−0.12, −0.01	−1.06[a], −1.06
Others	−1.30, −1.61	−0.92, −0.48	0.29, −0.48	−0.29, −0.01	−0.29, −0.48

[a] These saturated estimates refer to cells for which the multiple marginal independence model has an absolute adjusted residual exceeding 2.

Table 4

Example of SAS code for marginal WLS and GEE analysis

```
data wls;
  input educ  a b c d e count ;
  datalines;  * 1 line for each of 160 counts in complete table, as illustrated;
1  1 1 1 1 1  7
1  1 1 1 0 1  0
... ...
5  0 0 0 1 0  1
5  0 0 0 0 0  0
;
proc catmod  order=data; weight count;
  population educ; response logit;
  model a*b*c*d*e = (
  1 0 0 0 0, 0 1 0 0 0, 0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1,
  1 0 0 0 0, 0 1 0 0 0, 0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1,
  1 0 0 0 0, 0 1 0 0 0, 0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1,
  1 0 0 0 0, 0 1 0 0 0, 0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1,
  1 0 0 0 0, 0 1 0 0 0, 0 0 1 0 0, 0 0 0 1 0, 0 0 0 0 1)
  (1=‘a’, 2=‘b’, 3=‘c’, 4=‘d’, 5=‘e’) ;
run;
data gee;
  input case educ $ item $ response @@ ;
  datalines; * 262 lines, 1 for each subject, as illustrated below;
1 1 a 1    1  1 b 1    1  1 c 1    1  1 d 1    1  1 e 1
2 1 a 1    2  1 b 1    2  1 c 1    2  1 d 1    2  1 e 1
... ...
261 5 a 0   261 5 b 0   261 5 c 1   261 5 d 0   261 5 e 0
262 5 a 0   262 5 b 0   262 5 c 0   262 5 d 1   262 5 e 0
;
proc genmod data=gee;  class case item;
  model response = item   / dist=bin noint;
  repeated subject=case  /  type = exch  corrw ;
run;
```

= 22.8 and $X^2 = 19.3$ with df = 16. It also provides only weak evidence of improvement over the multiple marginal independence model (1) with change in $G^2 = 8.1$ when df = 4 and $P = 0.09$.

This model makes the rather strong assumption that the difference between logits at two levels of X is identical for all items. If it holds with much variability in $\{\alpha_i\}$, then considerable variability would occur among those levels in the expected number of 'yes' responses. In practice, however, the overall 'yes' response rate would often be relatively stable among levels of X. In Table 1, for instance, the values of ratio (number of responses)/(number of subjects) are 1.97, 1.75, 1.77, 1.50, and 1.93 for the five education groups, and model (5) has $\widehat{\alpha}_1 = 0.134, \widehat{\alpha}_2 = -0.086, \widehat{\alpha}_3 = -0.109, \widehat{\alpha}_4 = -0.313$, and $\widehat{\alpha}_5 = 0$. When multiple marginal independence is violated, it may be more common for any particular group to have relatively low probabilities for some items and relatively high ones for others.

When X has ordered levels, a way to obtain an unsaturated model and yet obtain patterns such as those just described is by permitting linear trends for the effects of X. For fixed scores $\{x_i\}$ for levels of X, the model

$$\log\left(\frac{\pi_{j|i}}{1-\pi_{j|i}}\right) = \beta_j + \gamma_j x_i, \quad i = 1, \ldots, r, \;\; j = 1, \ldots, c \quad (6)$$

permits a separate slope for each item. The model of interest may be nested between this model and the saturated model (2), whereby only some items have a linear trend and the others have arbitrary effects, or it may be nested between this model and the multiple marginal independence model, whereby some subset of items are marginally independent of the group variable. In Table 1, the educational levels, high school (HS), 2-year college, and 4-year college, have a clear ordering, but their relation to the other two educational levels is unclear. In any case, an inspection of the marginal proportions in Table 1 or of $\{\widehat{\beta}_{ij}\}$ for model (2) in Table 3 does not suggest any monotonic trends.

Most of the data in Table 1 come from groups, HS and 4-year college, and a noticeably lower sample-response rate occurs for items B and E with the 4-year college group. Otherwise, taking into account the small sample sizes at the other educational levels, it seems plausible that the marginal probabilities are roughly equal for items B, C, and E regardless of the educational level and are roughly equal among educational levels for item A and roughly equal among educational levels for item D. That is, the sample seems relatively well described by the marginal model

$$\log\left(\frac{\pi_{j|i}}{1-\pi_{j|i}}\right) = \beta_A I(j=1) + \beta_D I(j=4) + \beta_{BCE} I(j=2,3,5) + \beta_{BE4} I(i=4; j=2,5), \tag{7}$$

where $I(\cdot)$ is the indicator function. Table 3 also shows the estimates for this model. For all i, this model yields the fitted marginal probability estimates $\widehat{p}_{1|i} = 0.167$, $\widehat{p}_{4|i} = 0.497$, and $\widehat{p}_{j|i} = 0.383$ for $j = 2,3,5$, except that $\widehat{p}_{2|4} = \widehat{p}_{5|4} = 0.257$. Here, $G^2 = 21.5$ and $X^2 = 17.5$, for df $= 21$, summarize the goodness of this description. Since an inspection of the marginal proportions in Table 1 suggested this model, one should not use the above in any formal sense. Having additional explanatory variables besides education would provide a greater scope for building unsaturated marginal models.

2.4 *Marginal Models with Simpler Joint Distributions*

The main questions of interest pertaining to Table 1 refer to marginal distributions of the joint distribution of responses on the five items. The actual form of this joint distribution may be regarded as a nuisance or at best of secondary interest. Thus, the models considered so far deal directly with the marginal distributions and make no attempt to describe the joint distribution of the responses. Because of this, model (2), which does not specify any pattern for the marginal probabilities, is saturated.

Alternatively, one can model the marginal distributions while simultaneously modeling the joint distribution of the responses. With this approach, one can consider an unsaturated model, such as an ordinary log-linear model, for the joint distribution and can then also add structure for the marginal distributions. An advantage is that the resulting fitted values are smoother, lacking certain higher-order interactions. One can fit log-linear models simultaneously with compatible marginal models using methods described in Fitzmaurice and Laird (1993) and in Lang and Agresti (1994). Lang's S-plus function 'gllm' mentioned earlier can fit such models. (However, 'gllm' sometimes requires the addition of a small constant to empty cells when a sufficient statistic for the log-linear model falls on the boundary, in which case finite estimates of some parameters of the model do not exist for the joint distribution.)

We now illustrate marginal models having simpler joint distributions. Model (1) deals with the saturated log-linear model for the joint distribution and imposes the constraint of multiple marginal independence. However, standard log-linear analyses reveal that simpler models adequately describe the joint distribution of the group variable X and the five items. Although models permitting only two-factor associations fit poorly, models having three-factor interactions in addition to two-factor associations fit well. In the common log-linear model notation for the sufficient marginal configurations, the model (ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, XAB, XAC, XAD, XAE, XBC, XBD, XBE, XCD, XCE, XDE) is adequate. Goodness-of-fit statistics with large values of df for such sparse data no longer have approximate chi-squared null distributions, but $X^2 = 21.3$ and $G^2 = 19.1$ with df $= 70$ summarize a fit that is very close to the cell counts. (The ordinary df value loses some validity here, as certain sufficient marginal totals assumed the boundary value of zero, which implies that many of the cell-fitted values must also equal zero and provide perfect fits.) We also fitted this log-linear model of three-factor interactions with the addition of the constraints deleting the marginal education effects. For a variety of added small constants and for a variety of models including simpler ones that deleted some three-factor interaction terms, results were consistent with those obtained using model (1); specifically, the deviance increased on the order of about 30 for df $= 20$, with the addition of the constraint of multiple marginal independence.

2.5 *Sparseness Issues*

One of the limitations of the ML modeling approach that occurs commonly for repeated measurement data is that of sparseness of the data. With r levels of predictors and c outcome categories, sparseness can occur in the $r \times 2^c$ contingency table as a result of having many possible choices (i.e., large c) or as a result of having multiple or continuous predictors (large r). Chi-squared tests of model goodness of fit are best suited for modest c, with at most a few categorical predictors.

For the parameters of interest in comparing marginal probabilities across levels of a predictor, the sparseness is relevant in terms of the marginal totals at combinations of the predictor levels and the outcome categories. The marginal models do not have reduced sufficient statistics, but ideally for each j most of the counts in the $r \times 2$ marginal tables relating X to selection of outcome category j should exceed five. For Table 1, for instance, the models refer to five separate 5×2 marginal tables, and 45 of those 50 cells have a cell count exceeding five. When the number of cells in the table is very large, model goodness-of-fit statistics may not have null distributions close to chi-squared, but ordinary inference applies reasonably well to estimate $\{\beta_{ij}\}$ and to compare nested models as long as these marginal counts are not too small.

One way of dealing with sparseness is to study separately parts of the data set containing most of the data and in which the sparseness is not severe. In Table 1, for instance, 201 of the 262 subjects fall into just two of the educational levels, HS and 4-year college, with which one could conduct analyses. By doing so, one obtains somewhat stronger conclusions. To illustrate, the likelihood-ratio statistic, which compares fits of the marginal models with and without education effects, equals 15.9 with df $= 5$ and a P value of 0.007. The $\{\widehat{\beta}_j\}$ for the multiple marginal independence model (1) are $-1.42, -0.72, -0.61, -0.02$, and -0.67, compared to $-1.56, -0.71, -0.50, -0.02$, and -0.63 for the complete data with all five educational levels.

3. Testing Multiple Marginal Independence

To test the null hypothesis of multiple marginal independence with Table 1, Loughin and Scherer (1998) constructed a modified Pearson statistic that compares the cell counts in Table 1 with their proper expected values under the null hypothesis. They showed that its null asymptotic distribution is that of a linear combination of chi-squared random variables each having df $= 1$. This is reminiscent of Rotnitzky and

Jewell (1990), who showed that in generalized linear models with correlated observations, a naive likelihood-ratio statistic based on assuming independent observations behaves asymptotically like a weighted combination of independent χ_1^2 random variables. Loughin and Scherer noted that the coefficients for the linear combination in their statistic depend on the unknown cell probabilities for the complete table. Another disadvantage of their statistic is that it is not invariant to switching the 'yes' and 'no' labels for all the items. For example, their modified Pearson statistic equals 20.85 for Table 1; if instead we form the table with outcome categories referring to the wording "Indicate which are *not* your primary sources of veterinary information," then their statistic would equal 9.98 instead.

3.1 *Model-Based Tests*

Alternatively, to test the hypothesis of multiple marginal independence with large samples, one could use a likelihood-ratio or Pearson statistic for testing the goodness of fit of the logit model (1) specifying this hypothesis. This corresponds to a comparison of that model with the saturated model (2). The null hypothesis of multiple marginal independence corresponds to $H_0 : \beta_{1j} = \cdots = \beta_{rj}, j = 1,\ldots,c$, in the saturated model. The likelihood-ratio statistic G^2 equals -2 times the log of the ratio of the maximized likelihoods for models (1) and (2). This statistic and the corresponding Pearson statistic X^2, which compares the $r \times 2^c$ observed and fitted counts for model (1), have large-sample chi-squared distributions with df $= c(r-1)$, the difference in parameter dimensionality of the two models. For these statistics, the resulting null distribution does not assume any particular structure for the joint distribution.

These tests of multiple marginal independence, i.e., goodness-of-fit tests of model (1), yield $G^2 = 30.9$ and $X^2 = 26.7$, each with df = 20, providing weak evidence against the hypothesis with $P = 0.06$ and 0.14. Similar test statistic values occur while conducting the test using an unsaturated structure for the joint distribution; for instance, one such test compares the three-factor interaction models with and without the marginal constraint. Similar statistics also occur by comparing the models using the GEE methods. The GEE methods are not likelihood based and require Wald tests to check the lack of fit; the test statistic is a quadratic form that uses the estimates of the extra terms that are in model (2) but not in model (1) and their inverse covariance matrix. Here, the Wald statistic equals 33.2 with df = 20 and $P = 0.03$.

Because the complete 5×2^5 table corresponding to Table 1 is sparse, we cautiously make conclusions based on tests having df = 20. More reliable and informative tests use a model-based comparison of the multiple marginal independence model with a model that provides some structure for the nature of the marginal inhomogeneity. Using a more specific alternative provides the potential for increased power and also focuses attention on estimating the effects that may exist. To illustrate, one can compare model (1) with the additive-effects model (5). That is, assuming model (5), one tests multiple marginal independence by testing $\alpha_1 = \cdots = \alpha_r$, yielding chi-squared statistics with df $= r-1$. This model also provides only a weak evidence of improvement over the multiple marginal independence model, the change in $G^2 = 8.1$, df = 4, and $P = 0.09$.

3.2 *Testing based on the Bonferroni Approach*

Loughin and Scherer (1998) noted the inappropriateness of testing multiple marginal independence by applying the usual chi-squared statistics directly to tables that are similar to Table 1 in form. Another naive approach calculates chi-squared for each of the c marginal $r \times 2$ tables relating X to each component of Y and treats them as being independent by adding the values and their df values to obtain a summary 'chi-squared' statistic with df $= c(r-1)$. For Table 1, the five marginal Pearson statistics are 5.96, 7.89, 4.62, 1.42, and 10.95, each having df = 4. The summary naive 'chi-squared' statistic equals 30.84 with df = 20, having $P = 0.06$. The corresponding naive likelihood-ratio statistic is 32.44, with $P = 0.04$. Perhaps surprisingly, these statistics are only slightly different from the legitimate df = 20 chi-squared values reported earlier and obtained while comparing the ML fits of the marginal models. However, an inspection of the asymptotic covariance matrix for estimates $\{\widehat{\beta}_{ij}\}$ for model (2) reveals weak correlations between terms $\widehat{\beta}_{ij}$ and $\widehat{\beta}_{ik}$ describing the association in different marginal tables.

A valid albeit somewhat conservative way of simultaneously using the c marginal Pearson statistics to test multiple marginal independence is via the Bonferroni approach. For instance, to ensure asymptotic overall size of at most α, one conducts each separate df $= r-1$ chi-squared test in the ordinary way by rejecting the overall hypothesis if the minimum P value is $\leq \alpha/c$; i.e., if P_j is the P value for the jth test, an adjusted overall P value is $\min(cP_j, 1)$. When this overall test gives evidence against the null hypothesis, the separate chi-squared components provide information about the marginal tables that are responsible. With this approach, none of the five Pearson statistics just reported is sufficiently large to cast serious doubts on the hypothesis of multiple marginal independence, as the overall adjusted P value equals 0.136.

3.3 *Small-Sample Tests*

Because the chi-squared weights for the Loughin and Scherer (1998) large-sample modified Pearson test depend on unknown probabilities, these authors also proposed an alternative bootstrap-based test that would be more appropriate for small samples or for highly sparse data. Their resampling mechanism uses the data in the complete table and generates bootstrap resamples of the same sample size under the assumption that X is independent of the joint distribution of the five items. That is, the log-linear model for the resampling mechanism is $(ABCDE, X)$. This is actually a special case of multiple marginal independence that is equivalent to two-way independence in the 5×2^5 table in which the rows are the five educational groups and the columns are the 32 possible response combinations. Their bootstrap-test P value is the proportion of generated resamples for which their modified Pearson test statistic is at least as large as the sample value.

One could, of course, use the bootstrap approach with an alternative test statistic for the hypothesis of multiple marginal independence. For instance, one could obtain a bootstrap distribution for the likelihood ratio or Pearson statistic for testing the fit of the marginal model (1). This is computationally more complex, however, because of the necessity of fitting the model with each bootstrap resample.

A simpler statistic to use is the naive chi-squared statistic just mentioned, which is based on summing Pearson statistics from the separate marginal tables. Again, the P value is the proportion of generated resamples for which this test statistic is at least as large as the sample value.

The naive test statistic that sums marginal Pearson statistics has great appeal, as it expands the Loughin and Scherer modified Pearson statistic to consider the 'no' outcome cells as well as the 'yes' cells in the marginal tables. Hence, the sum of marginal Pearson statistics is invariant to the switching of the 'yes' and 'no' labels for all the items. For Table 1, the sum of marginal Pearson statistics equals 30.8 regardless of whether the table is expressed in terms of "check all that apply" or "check all that do not apply." We obtained a bootstrap P value of 0.062 for 10,000 bootstrap resamples. For Table 1, with the Loughin and Scherer modified Pearson statistic, for which the sample value is 20.85, we obtained a bootstrap P value of 0.049, whereas they reported a value of 0.047, with 5000 resamples. By contrast, when we analyzed the data in terms of the wording "Indicate which are *not* your primary sources of veterinary information," with the same bootstrap resamples, their modified Pearson statistic of 9.98 had a bootstrap P value of 0.105.

3.4 *Comparison of Approaches*

An appealing aspect of the Loughin and Scherer bootstrap approach is that it does not have to rely large-sample asymptotics with small samples or on highly sparse data. As previously discussed, the tests of marginal independence based on the comparison of model (1) with more complex models are valid only asymptotically, with the actual size being close to the nominal size when the sample size is large relative to the number of parameters. A disadvantage of their bootstrap approach is that the sampling distribution of the test statistic employed is generated under the log-linear model $(ABCDE, X)$, which is narrower than that of the null hypothesis. Although this need not adversely affect the performance of the resulting tests in terms of power or size, it would be more satisfying to use a null sampling distribution that applies to *all* cases in which multiple marginal independence holds.

One way to come closer to this ideal is to apply the bootstrap method by resampling from the multinomial distribution corresponding to the fit of the null model (1) to the complete table. When we did this using 10,000 resamples, we obtained (1) $P = 0.114$ with the sum of marginal Pearson statistics, (2) $P = 0.108$ with the Loughin and Scherer modified Pearson statistics, and (3) $P = 0.131$ with their statistic applied with 'no' responses. Even this approach, however, is not ideal, as it is better to use conditional rather than unconditional distributions for studying sparse asymptotic behavior. However, marginal models such as (1) have no reduced sufficient statistics; the usual small-sample exact conditional testing methods are inapplicable here, because marginal models are not canonical-link generalized linear models for the complete table. It is comforting to know that for linear exponential family models, the Pearson statistic and the model parameter estimates are asymptotically independent (McCullagh, 1985). One way to reduce slightly the error resulting from the fact that the bootstrap does not resample with the true probabilities is to use the double bootstrap. However, this also involves refitting the model for all resamples at the first stage, and with the same total computational effort, this reduced error is likely to be offset by an increased Monte Carlo error, even with recycling methods (Newton and Geyer, 1994). The development of improved small-sample analyses for marginal models is by itself a useful topic for future research.

It does not seem possible to make general statements about the relative performance (e.g., efficiency comparisons) of the bootstrap and model-based tests. As the sample size increases, we feel it is more advantageous to use the marginal model-comparison tests. The sampling distribution then applies for all cases under which multiple marginal independence holds, and attention is directed toward the more important issue of describing departures from the null hypothesis. When using bootstrap or model-based tests, we recommend the use of test statistics that are invariant to the choice of outcome category for each item.

4. Extensions and Alternative Approaches

In summary, the data summarized in Table 1 provide only weak evidence against multiple marginal independence of education and the five items. Advantages of employing marginal modeling include the by-products of probability estimates of selecting various outcome categories, estimates of parameters that describe possible departures from multiple marginal independence, and the capacity to employ residuals and other measures of lack of fit.

With the marginal modeling approach, one can extend the analyses discussed in this paper. For instance, suppose that each of two response variables can have multiple responses, where Y_1 has c_1 categories and Y_2 has c_2 categories. Then the models apply to a $2^{c_1} \times 2^{c_2}$ contingency table at each setting of explanatory variables. To illustrate, for a sample of farmers, Y_1 might refer to primary sources of veterinary information and Y_2 to different swine management practices employed. One might test for independence between Y_1 and Y_2 in the sense of simultaneous pairwise marginal independence between each component of Y_1 and each component of Y_2. The marginal modeling approach specifies independence simultaneously for c_1c_2 separate 2×2 marginal tables, one for each such pair of components. One can fit this model using the methodology of Lang and Agresti (1994) alluded to in Section 2, and ordinary goodness-of-fit statistics have $\text{df} = c_1c_2$. Also, one could study the effects of explanatory variables on components of each response variable. Of course, sparseness becomes even more of an issue for such extensions.

This article focused on marginal models because they directly address response probabilities for the various outcome categories. Other modeling approaches could be used for which multiple marginal independence occurs as a special case. Agresti and Liu (1998) surveyed various ways of modeling a categorical variable allowing arbitrarily many category choices. These include item response models (Baker 1992), random-effects models, quasi-symmetric log-linear models that have connections with item response models, and models that describe the actual subset of the responses chosen. In our experience, random-effects models usually fit data of this sort poorly. For instance, for the Kansas farmer data set, none of the subjects responded with a 'no' for all five of the sources, showing a likely violation of the assumption of independence of successive responses, given the random effect. Given that someone does not choose a particular four of the sources, they seem much more likely to choose the fifth source than if independence held. In applications of this type, respondents may psychologically feel obligated to select at least one category. In addition, unconditionally random-effects models imply nonnegative associations between pairs of items, whereas these data show rather marked negative associations between responses on items A and D and between responses on items A and E.

Finally, a challenge for future research is to develop tests of multiple marginal independence that work well for small samples or for highly sparse data and that have a distribution that applies under the entire null hypothesis. For sparse data with a very large number of cells, the usual goodness-of-fit statistics have approximate normal distributions (Morris, 1975). There is some evidence that the jackknife can work well in estimating asymptotic variances of such statistics (Simonoff, 1986), and investigating this and various bootstrap methods, including finding suitable pivotal statistics, for marginal models with large tables is a useful area for future research.

Acknowledgements

The research of the first author was supported by grants from the National Institute of Health and the National Science Foundation. The authors thank Dr James Booth and Dr Thomas Loughin for helpful discussions, Dr Joseph Lang for his S-plus function 'gllm' for fitting generalized log-linear models, and the two referees for their helpful comments.

Résumé

Cet article discute la modélisation d'une variable catégorielle pour laquelle les sujets peuvent sélectionner n'importe quel nombre de catégories. Pour c catégories, la variable de réponse consiste alors en une classification croisée de c composantes binaires, chacune représentant une catégorie. EN utilisant les données d'un sondage (Loughin et Scherer, 1998) dans lequel des fermiers du Kansas indiquaient leurs sources primaires d'information vétérinaire, nous discutons la modélisation simultanée du logit des composantes binaires d'une réponse multivariée. L'usage d'un ajustement du maximum de vraisemblance ou de quasi vraisemblance fournit des tests du Chi-2 avec $c(r-1)$ ddl pour tester l'indépendance entre chacune des c composantes de réponse et une variable explicative à r catégories. Ces tests sont des alternatives au test du Chi-2 pondéré et au test du bootstrap que Loughin et Scherer ont proposés dans cette hypothèse.

References

Agresti, A. and Liu, I.-M. (1998). *Strategies for modeling responses to a categorical variable allowing arbitrarily many category choices.* Technical Report 575, University of Florida, Department of Statistics, Gainsville, FL.

Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics* **29,** 813–828.

Baker, F. B. (1992). *Item Response Theory. Parameter Estimation Techniques.* New York: Marcel Dekker.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80,** 141–151.

Haber, M. (1985). Maximum likelihood methods for linear and log-linear models in categorical data. *Computational Statistics and Data Analysis* **3,** 1–10.

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.

Lang, J. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* **89,** 625–632.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Loughin, T. M. and Scherer, P. N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics* **54,** 630–637.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *International Statistical Review* **53,** 61–67.

Miller, M. E., Davis, C. S., and Landis, J. R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49,** 1033–1044.

Morris, C. (1975). Central limit theorems for multinomial sums. *Annals of Statistics* **3,** 165–188.

Newton, M. A. and Geyer, C. J. (1994). Bootstrap recycling: A Monte Carlo alternative to the nested bootstrap. *Journal of the American Statistical Association* **89,** 905–912.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77,** 485–497.

Simonoff, J. S. (1986). Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *Journal of the American Statistical Association* **81,** 1005–1011.

Received July 1998. *Revised February* 1999.
Accepted February 1999.