Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses

Author(s): Joseph B. Lang and Alan Agresti

Source: *Journal of the American Statistical Association*, Vol. 89, No. 426 (Jun., 1994), pp. 625-632

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2290865

Accessed: 13-04-2020 13:10 UTC

Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses

Joseph B. Lang and Alan Agresti*

We discuss model-fitting methods for analyzing simultaneously the joint and marginal distributions of multivariate categorical responses. The models are members of a broad class of generalized logit and loglinear models. We fit them by improving a maximum likelihood algorithm that uses Lagrange's method of undetermined multipliers and a Newton–Raphson iterative scheme. We also discuss goodness-of-fit tests and adjusted residuals, and give asymptotic distributions of model parameter estimators. For this class of models, inferences are equivalent for Poisson and multinomial sampling assumptions. Simultaneous models for joint and marginal distributions may be useful in a variety of applications, including studies dealing with longitudinal data, multiple indicators in opinion research, cross-over designs, social mobility, and inter-rater agreement. The models are illustrated for one such application, using data from a recent General Social Survey regarding opinions about various types of government spending.

KEY WORDS: Adjusted residuals; Constrained maximum likelihood; Lagrange multiplier; Marginal models; Ordinal data; Repeated measurement.

1. INTRODUCTION

Consider Table 1, taken from the 1989 General Social Survey conducted by the National Opinion Research Center at the University of Chicago. Subjects in the sample were asked their opinion regarding government spending on (1) the environment, (2) health, (3) assistance to big cities, and (4) law enforcement. The common response scale was (too little, about right, too much).

Two types of questions about Table 1 lead to distinct types of models. One question relates to how the response distributions differ for the four items. For instance, one might ask whether subjects regarded spending as relatively higher on one item than on the others. Answers to this question refer to modeling the four one-way marginal distributions of Table 1. A second question pertains to the dependence structure of the responses. For instance, one might study the strength of association between responses on various pairs of items, analyzing whether some pairs are more strongly associated than others or whether the pairwise association varies according to responses on the other indicators. Consideration of these matters relates to modeling the joint distribution in Table 1.

Standard models for joint distributions of categorical responses do not imply simple relationships among marginal distributions. Hence modeling marginal distributions of categorical responses is normally conducted separately from modeling of joint distributions. In this article we show how models of the two types can be fitted simultaneously. This process provides improved model parsimony. One also obtains a single test that simultaneously summarizes goodness of fit and a single set of fitted values and residuals. Estimators of model parameters and cell expected frequencies are also potentially more efficient than with separate fitting procedures.

We consider the modeling of multivariate categorical responses in which a different response scale is allowed for each response. The response scales may be nominal or ordinal. Section 2 specifies a generalized class of log-linear and logit models that one can apply simultaneously to the joint and marginal distributions. The marginal distributions modeled may be of any order. For instance, the modeling of pairwise associations without controlling for other variables refers to second-order marginal tables of the joint distribution. Section 3 expresses the models using constraint specifications. We then propose a model-fitting approach that improves on approaches discussed by Aitchison and Silvey (1958) and Haber (1985a), using the method of Lagrange's undetermined multipliers. The improved algorithm applies a Newton–Raphson iterative scheme in which the matrix to be inverted has much simpler form than in previously proposed algorithms.

Our analyses assume a multinomial sampling scheme for the cell counts. Section 4 provides asymptotic distributions of the parameter estimators and generalizes results of Birch (1963) and Palmgren (1981) relating these distributions to those obtained assuming Poisson sampling. Section 5 considers model goodness of fit and shows that it can be partitioned into goodness of fit for the marginal and joint components. This section also generalizes Haberman's (1973) adjusted residuals for inspecting the fit both in cells of the table and in marginal totals. Section 6 illustrates the simultaneous fitting approach using Table 1, and Section 7 discusses potential compatibility problems of simultaneous joint and marginal models. Finally, Section 8 discusses other types of data sets for which simultaneous joint and marginal modeling is relevant.

2. SIMULTANEOUS JOINT AND MARGINAL MODELS

Let T denote the number of component variables in the multivariate response, let I_t denote the number of categories in the response scale for response variable t, and let $r = \prod_{t=1}^{T} I_t$ denote the number of possible response profiles. Suppose that observations on the responses are obtained at s fixed levels of a set of explanatory variables. The observed data can be displayed in an $s \times r$ contingency table. For simplicity, most of our notation refers to a single population ($s = 1$).

* Joseph B. Lang is Assistant Professor, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242. Alan Agresti is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611.

Journal of the American Statistical Association
June 1994, Vol. 89, No. 426, Theory and Methods

Table 1. Opinions About Government Spending

Cities		1			2			3		
Law Enforcement		1	2	3	1	2	3	1	2	3
Environment	Health									
1	1	62	17	5	90	42	3	74	31	11
	2	11	7	0	22	18	1	19	14	3
	3	2	3	1	2	0	1	1	3	1
2	1	11	3	0	21	13	2	20	8	3
	2	1	4	0	6	9	0	6	5	2
	3	1	0	1	2	1	1	4	3	1
3	1	3	0	0	2	1	0	9	2	1
	2	1	0	0	2	1	0	4	2	0
	3	1	0	0	0	0	0	1	2	3

NOTE: These data are from the 1989 General Social Survey, with categories 1 = too little, 2 = about right, and 3 = too much.

Let Y_i denote the number of subjects having response profile i, where $i = (i_1, \ldots, i_T)$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_r)'$, where π_i denotes the probability that a randomly selected subject has response profile i. We assume that $\mathbf{Y} = (Y_1, \ldots, Y_r)'$ has a multinomial distribution with probabilities $\boldsymbol{\pi}$. We denote corresponding expected frequencies in the cells of the contingency table by $\boldsymbol{\mu}$. For marginal distribution t, let $\phi_k(t)$ denote the probability of response k. When response variable t is ordinal, let $\gamma_j(t) = \sum_{k=1}^{j} \phi_k(t)$ denote the jth cumulative marginal probability.

Let $J(\cdot)$ denote a model for the joint distribution, and let $M(\cdot)$ denote a model for first-order marginal distributions. Let $J(A) \cap M(B)$ denote the model that specifies simultaneously model A for the joint distribution and model B for the marginal distributions. Let S denote the saturated model. For instance, the model $J(S) \cap M(B)$ assigns structure only to the marginal distributions and permits arbitrary higher-order interactions among the responses. Fitting a marginal model $M(B)$ alone is equivalent to fitting the simultaneous model $J(S) \cap M(B)$. The simplest model of interest for $M(\cdot)$ is first-order marginal homogeneity, denoted by $M(H)$.

Fitting a joint model $J(A)$ alone is equivalent to fitting the simultaneous model $J(A) \cap M(S)$, which focuses on the interaction structure without making assumptions about the marginal distributions. This class includes ordinary log-linear models for expected cell counts in contingency tables. These models are not designed to estimate effects in marginal distributions, because their parameters have a conditional interpretation. The simplest model of interest for $J(\cdot)$ is mutual independence of the responses, denoted by $J(I)$.

This article focuses on a more parsimonious class of models that use unsaturated components for both the joint and marginal distributions. Specifically, we consider models for each distribution that have form $\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. We denote the model for the joint distribution by $\mathbf{C}_1 \log \mathbf{A}_1\boldsymbol{\mu} = \mathbf{X}_1\boldsymbol{\beta}_1$ and the model for the marginal distributions by $\mathbf{C}_2 \log \mathbf{A}_2\boldsymbol{\mu} = \mathbf{X}_2\boldsymbol{\beta}_2$. For each component, we take $\mathbf{C}$ to be either an identity, contrast (rows sum to 0), or zero matrix and we assume that the model matrix $\mathbf{X}$ has full column rank. The matrices in the specification of the joint model need not have the same form as the matrices in the specification of the marginal model.

Permissible models for the joint distribution include not only simple log-linear and logit models but also models for log odds ratios using groupings of cells, such as models for global odds ratios (Dale 1986). The marginal model can also be a log-linear or a corresponding logit model, or some other form of multinomial response model, such as a cumulative logit model. Haber (1985a,b) gave examples of nonstandard models that fall in this class of marginal models.

3. MAXIMUM LIKELIHOOD FITTING OF SIMULTANEOUS MODELS

The standard approach to maximum likelihood (ML) fitting of marginal or simultaneous models involves solving the score equations using the Newton–Raphson method, Fisher scoring, or some other iterative reweighted least squares algorithm. The most common approach—used, for instance, by Dale (1986), McCullagh and Nelder (1989, p. 219), Lipsitz, Laird, and Harrington (1990), and Becker and Balagtas (1991)—reparameterizes the cell probabilities in the multinomial log-likelihood in terms of the joint and marginal distribution model parameters. A severe limitation of this approach is that the reparameterization is typically difficult and very awkward for the general T-variate case.

An alternative method, discussed by Aitchison and Silvey (1958), Haber (1985a), and Haber and Brown (1986), is Lagrange's method of undetermined multipliers. One views the models as inducing constraints on the cell probabilities and maximizes the likelihood subject to these constraints. Because the algorithm that we use is a modification of Haber's, we first briefly describe his algorithm.

The simultaneous models can be specified as

$$\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \qquad \text{ident}(\boldsymbol{\mu}) = \mathbf{0}, \tag{1}$$

where $\mathbf{C} = \mathbf{C}_1 \oplus \mathbf{C}_2$, $\mathbf{A}' = (\mathbf{A}_1', \mathbf{A}_2')$, $\mathbf{X} = \mathbf{X}_1 \oplus \mathbf{X}_2$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, and $\text{ident}(\boldsymbol{\mu}) = 0$ denotes multinomial identifiability constraints. The symbol $\oplus$ denotes a direct sum. (For example, $\mathbf{C}_1 \oplus \mathbf{C}_2 = \oplus_{i=1}^{2} \mathbf{C}_i$ is the block diagonal matrix with $\mathbf{C}_1$ and $\mathbf{C}_2$ as the blocks.) For the case of s independent multinomial samples of sizes $\mathbf{n} = (n_1, n_2, \ldots, n_s)'$, the multinomial identifiability constraints have the form $(\oplus_1^s \mathbf{1}_r')\boldsymbol{\mu} - \mathbf{n} = \mathbf{0}$. Let $\mathbf{U}$ denote a full column rank matrix such that the space spanned by the columns of $\mathbf{U}$ is the orthogonal complement of the space spanned by the columns of $\mathbf{X}$, so that $\mathbf{U}'\mathbf{X} = \mathbf{0}$. Model (1) is equivalently expressed as the constraint model

$$\mathbf{U}'\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{0}, \qquad \text{ident}(\boldsymbol{\mu}) = \mathbf{0}. \tag{2}$$

Haber (1985a) outlined a method for computing $\mathbf{U}$.

The objective is to maximize the kernel of the multinomial log-likelihood, $l(\boldsymbol{\mu};\mathbf{y}) = \mathbf{y}' \log \boldsymbol{\mu}$, subject to model (1) or, equivalently, (2), holding. We express the model parameter space as

$$\{\boldsymbol{\mu} : \mathbf{U}'\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{0}, \quad \text{ident}(\boldsymbol{\mu}) = \mathbf{0}\} = \{\boldsymbol{\mu} : \mathbf{f}(\boldsymbol{\mu}) = \mathbf{0}, \quad \text{ident}(\boldsymbol{\mu}) = \mathbf{0}\}.$$

Haber (1985a) used a Newton–Raphson algorithm to solve the Lagrangian likelihood equations

$$\mathbf{g}(\hat{\theta}^+) = \begin{bmatrix} \dfrac{\partial l(\hat{\mu};\mathbf{y})}{\partial \mu} + \dfrac{\partial\, \text{ident}(\hat{\mu})'}{\partial \mu}\hat{\tau} + \dfrac{\partial \mathbf{f}(\hat{\mu})'}{\partial \mu}\hat{\lambda} \\ \mathbf{f}(\hat{\mu}) \\ \text{ident}(\hat{\mu}) \end{bmatrix} = \mathbf{0},$$

where $\theta^+ = \text{vec}(\mu, \lambda, \tau)$, with λ and τ being vectors of undetermined multipliers. For the models he considered, $\hat{\tau}$ could be solved for explicitly and was nonstochastic. Hence he considered the solution $\hat{\theta} = \text{vec}(\hat{\mu}, \hat{\lambda})$ to the simpler set of equations

$$\mathbf{g}(\hat{\theta}) = \begin{bmatrix} \dfrac{\partial l(\hat{\mu};\mathbf{y})}{\partial \mu} - \mathbf{1}_u + \dfrac{\partial \mathbf{f}(\hat{\mu})'}{\partial \mu}\hat{\lambda} \\ \mathbf{f}(\hat{\mu}) \end{bmatrix} = \mathbf{0},$$

where $u = rs$ denotes the length of the vectors $\mathbf{y}$ and μ. The Newton–Raphson iterative scheme is

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{\partial \mathbf{g}(\theta^{(t)})}{\partial \theta'}\right)^{-1} \mathbf{g}(\theta^{(t)}), \qquad t = 0, 1, 2, \ldots,$$

where the derivative matrix can be calculated as

$$\frac{\partial \mathbf{g}(\theta)}{\partial \theta'} = \begin{bmatrix} \dfrac{\partial^2 l(\mu;\mathbf{y})}{\partial \mu' \partial \mu} + \left(\dfrac{\partial^2 \mathbf{f}(\mu)'}{\partial \mu' \partial \mu}\right)(\lambda \otimes \mathbf{I}_u) & \dfrac{\partial \mathbf{f}(\mu)'}{\partial \mu} \\ \dfrac{\partial \mathbf{f}(\mu)}{\partial \mu'} & \mathbf{0} \end{bmatrix}.$$

A drawback to this algorithm is that the matrix $\partial \mathbf{g}(\theta)/\partial \theta'$ is typically very large (with the numbers of rows and columns exceeding the number of cells in the contingency table) and does not have a simple form, making inversion difficult. Our proposed modifications of this algorithm uses an iterative scheme involving a matrix that is much simpler to invert. Also, we use a reparameterization from μ to $\xi = \log \mu$ for both the log-likelihood and model parameter space, to avoid interim out-of-range values during the iterative process.

For the reparameterized kernel of the log-likelihood, $l(\xi;\mathbf{y}) = \mathbf{y}'\xi$, we express the model parameter space as

$$\{\xi : \mathbf{U}'\mathbf{C} \log \mathbf{A}\mathbf{e}^{\xi} = \mathbf{0}, \quad \text{ident}(\xi) = \mathbf{0}\} = \{\xi : \mathbf{h}(\xi) = \mathbf{0}, \quad \text{ident}(\xi) = \mathbf{0}\}.$$

We solve for $\hat{\xi}$ by solving for $\hat{\theta}$ in the likelihood equations

$$\mathbf{g}(\hat{\theta}) = \begin{bmatrix} \dfrac{\partial l(\hat{\xi};\mathbf{y})}{\partial \xi} - \mathbf{e}^{\hat{\xi}} + \dfrac{\partial \mathbf{h}(\hat{\xi})'}{\partial \xi}\hat{\lambda} \\ \mathbf{h}(\hat{\xi}) \end{bmatrix} = \begin{bmatrix} \mathbf{y} - \mathbf{e}^{\hat{\xi}} + \mathbf{H}(\hat{\xi})\hat{\lambda} \\ \mathbf{h}(\hat{\xi}) \end{bmatrix} = \mathbf{0}, \tag{3}$$

where $\mathbf{H}(\xi) = \partial \mathbf{h}(\xi)'/\partial \xi$ and $\theta = \text{vec}(\xi, \lambda)$. To do this, we use the modified Newton–Raphson iterative scheme

$$\theta^{(t+1)} = \theta^{(t)} - (\mathbf{G}(\theta^{(t)}))^{-1}\mathbf{g}(\theta^{(t)}), \qquad t = 0, 1, 2, \ldots, \tag{4}$$

with

$$\mathbf{G}(\theta) = \begin{bmatrix} -\mathbf{D}(\mathbf{e}^{\xi}) & \mathbf{H}(\xi) \\ \mathbf{H}(\xi)' & \mathbf{0} \end{bmatrix},$$

where $\mathbf{D}(\mathbf{e}^{\xi})$ denotes a diagonal matrix with the elements of $\mathbf{e}^{\xi}$ on the main diagonal. We set $\xi^{(0)} = \log y$ and $\lambda^{(0)} = \mathbf{0}$, making the slight adjustment $\xi^{(0)} = \log(\mathbf{y} + \epsilon)$ for some small $\epsilon > 0$ when there are sampling zeros.

Lang (1992) proved that $\mathbf{G}(\theta)$ is the dominant part of $\partial \mathbf{g}(\theta)/\partial \theta'$, in the sense that

$$n_*^{-1}\frac{\partial \mathbf{g}(\theta)}{\partial \theta'} = n_*^{-1}\mathbf{G}(\theta) + \begin{bmatrix} o(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $o(1)$ represents a sequence that converges to 0 as $n_* = \min\{n_1, \ldots, n_s\} \to \infty$. The utility of using $\mathbf{G}(\theta)$ is due to the simplicity of its inverse, which is

$$\mathbf{G}^{-1}(\theta) = \begin{bmatrix} -\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1} & \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1} \\ (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1} & (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1} \end{bmatrix},$$

where $\mathbf{D} = \mathbf{D}(\mathbf{e}^{\xi})$ and $\mathbf{H} = \mathbf{H}(\xi)$. To invert $\mathbf{G}$, we need only invert a diagonal matrix, $\mathbf{D}$, and a symmetric positive definite matrix, $\mathbf{H}'\mathbf{D}^{-1}\mathbf{H}$. After obtaining the fitted values $\hat{\mu} = \mathbf{e}^{\hat{\xi}}$ on convergence of the algorithm, we calculate model parameter estimates using $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} \log \mathbf{A}\hat{\mu}$.

4. ASYMPTOTIC BEHAVIOR OF ESTIMATORS

One can use the delta method to describe the asymptotic behavior of $\hat{\theta} = \text{vec}(\hat{\xi}, \hat{\lambda})$ and several continuous functions of $\hat{\theta}$, such as $\hat{\mu} = \mathbf{e}^{\hat{\xi}}$ and $\hat{\beta}$. In this section, ξ, μ, and β denote the true (unknown) parameter values. Assuming that model (1) holds, Lang (1993) showed that under certain nonrestrictive assumptions, the asymptotic normal distributions

$$\hat{\lambda} \sim \mathbf{A}N(0, \Sigma_{\hat{\lambda}}^{(M)}),$$
$$\hat{\xi} \sim \mathbf{A}N(\xi, \Sigma_{\hat{\xi}}^{(M)}),$$
$$\hat{\mu} \sim \mathbf{A}N(\mu, \Sigma_{\hat{\mu}}^{(M)}),$$

and

$$\hat{\beta} \sim \mathbf{A}N(\beta, \Sigma_{\hat{\beta}}^{(M)}). \tag{5}$$

hold, where

$$\Sigma_{\hat{\lambda}}^{(M)} = (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1},$$

$$\Sigma_{\hat{\xi}}^{(M)} = \mathbf{D}^{-1} - \oplus_{k=1}^{s}\frac{\mathbf{1}_r\mathbf{1}_r'}{n_k} - \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1},$$

$$\Sigma_{\hat{\mu}}^{(M)} = \mathbf{D} - \oplus_{k=1}^{s}\frac{\mu_k\mu_k'}{n_k} - \mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}',$$

and

$$\Sigma_{\hat{\beta}}^{(M)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}(\mathbf{A}\mu)^{-1}\mathbf{A}\Sigma_{\hat{\mu}}^{(M)}\mathbf{A}'\mathbf{D}(\mathbf{A}\mu)^{-1}\mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The assumptions are analogous to requiring that a standard log-linear model contain all "fixed-by-design" parameters. Moreover, $\hat{\lambda}$ and $\hat{\xi}$ are asymptotically independent. The limiting distributions in (5) hold as $n_* = \min\{n_1, \ldots, n_s\}$ converges to infinity such that each multinomial index n_k grows at the same rate.

Under these assumptions, it follows from Birch (1963) that the same point estimates occur when we treat cell counts

as independent Poisson random variables. The asymptotic covariance matrices for Poisson sampling are

$$\boldsymbol{\Sigma}_{\hat{\lambda}}^{(P)} = (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1},$$
$$\boldsymbol{\Sigma}_{\hat{\xi}}^{(P)} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1},$$
$$\boldsymbol{\Sigma}_{\hat{\mu}}^{(P)} = \mathbf{D} - \mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}',$$

and

$$\boldsymbol{\Sigma}_{\hat{\beta}}^{(P)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}(\mathbf{A}\boldsymbol{\mu})^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\hat{\mu}}^{(P)}\mathbf{A}'\mathbf{D}(\mathbf{A}\boldsymbol{\mu})^{-1}\mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The limiting distributions hold as $\mu_* = \min\{\mu_i\}$ converges to infinity, such that each μ_i grows at the same rate. The estimated asymptotic covariance matrices for either sampling scheme are easily computed on convergence of iterative scheme (4), because they involve only matrices that are input or computed during the inversion of $\mathbf{G}(\boldsymbol{\theta})$.

The asymptotic covariance matrices of model parameter estimators under multinomial and Poisson sampling are related by $\boldsymbol{\Sigma}_{\hat{\beta}}^{(M)} = \boldsymbol{\Sigma}_{\hat{\beta}}^{(P)} - \boldsymbol{\Lambda}$, where

$$\boldsymbol{\Lambda} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}(\mathbf{A}\boldsymbol{\mu})^{-1} \times \mathbf{A}\left(\oplus_{k=1}^{s}\frac{\boldsymbol{\mu}_k\boldsymbol{\mu}_k'}{n_k}\right)\mathbf{A}'\mathbf{D}(\mathbf{A}\boldsymbol{\mu})^{-1}\mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

is a nonnegative definite matrix. Lang (1992, 1993) generalized a result of Palmgren (1981) for simple log-linear models regarding standard errors of elements in $\hat{\boldsymbol{\beta}}$ being the same for the two sampling schemes (i.e., the corresponding components of $\boldsymbol{\Lambda}$ are 0). For the generalized log-linear models (1), inferences for all parameters except the two intercept parameters (one for the joint and one for the marginal model) are the same whether one assumes full multinomial sampling or Poisson sampling. When categorical covariates are present and one assumes product multinomial sampling, the inferences are the same as with Poisson sampling for all the parameters except those fixed by the sampling design.

5. GOODNESS OF FIT AND RESIDUALS

To assess model goodness of fit, one can compare observed and fitted cell counts using the likelihood-ratio statistic G^2 or the Pearson statistic X^2. For nonsparse tables, assuming that the model holds, these statistics have approximate chi-squared distributions with degrees of freedom equal to the number of constraints implied by $\mathbf{C}\log\mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. From (3), X^2 can be rewritten as $X^2 = \hat{\boldsymbol{\lambda}}'\hat{\mathbf{H}}'\hat{\mathbf{D}}^{-1}\hat{\mathbf{H}}\hat{\boldsymbol{\lambda}}$. This statistic is identical to the Lagrange multiplier statistic (Aitchison and Silvey 1958, 1960; Silvey 1959), $L^2 = \hat{\boldsymbol{\lambda}}'(\hat{\boldsymbol{\Sigma}}_{\hat{\lambda}})^{-1}\hat{\boldsymbol{\lambda}}$.

To analyze lack of fit for a simultaneous model in a localized manner, we generalized Haberman's (1973) adjusted residuals for log-linear models. Cell i has adjusted residual defined by $\mathbf{e}_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)/ASE(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$.

From (3), $(\mathbf{y} - \hat{\boldsymbol{\mu}}) = -\mathbf{H}(\tilde{\boldsymbol{\xi}})(\hat{\boldsymbol{\lambda}} - \mathbf{0}) \approx -\mathbf{H}(\boldsymbol{\xi})(\hat{\boldsymbol{\lambda}} - \mathbf{0})$. By the delta method, using the asymptotic distribution of $\hat{\boldsymbol{\lambda}}$, the estimated asymptotic covariance matrix of $(\mathbf{y} - \hat{\boldsymbol{\mu}})$ is

$$\hat{\boldsymbol{\Sigma}}_{(\mathbf{y}-\hat{\mu})} = \mathbf{H}(\tilde{\boldsymbol{\xi}})(\mathbf{H}(\tilde{\boldsymbol{\xi}})'\mathbf{D}(\hat{\boldsymbol{\mu}})^{-1}\mathbf{H}(\tilde{\boldsymbol{\xi}}))^{-1}\mathbf{H}(\tilde{\boldsymbol{\xi}})'.$$

For the generalized log-linear models (1), this matrix is available as a byproduct of iterative scheme (4). It yields the estimated asymptotic standard errors needed to form the adjusted residuals. From this matrix, one can also easily obtain the estimated covariance matrix for differences between observed and fitted marginal counts, which are sums of corresponding differences for the cells. Using standard errors of such marginal differences, one can construct adjusted residuals for marginal counts or marginal cumulative counts.

At the start of the model-fitting process, it is often unclear which simultaneous models should be considered. To help determine which models may fit well, one can first investigate joint and marginal models separately. If a separate model of each type fits well, then generally so will the simultaneous model consisting of those two components. In fact, for models considered in this article, the simultaneous model $J(A) \cap M(B)$ has a likelihood-ratio statistic asymptotically equivalent to the sum of the statistical values for the separate models $J(A) \cap M(S)$ and $J(S) \cap M(B)$. We now outline why this happens.

Consider two hypotheses, $[\omega_1]$ and $[\omega_2]$, such that both are special cases of a hypothesis $[\omega_0]$ but neither is a special case of the other. Following Aitchison (1962), $[\omega_1]$ and $[\omega_2]$ are called *asymptotically separable* if the tests of $[\omega_1] \cap [\omega_2]$ against $[\omega_2] - ([\omega_1] \cap [\omega_2])$ and of $[\omega_1] \cap [\omega_2]$ against $[\omega_1] - ([\omega_1] \cap [\omega_2])$ use the same large-sample critical regions as do the unrestricted tests of $[\omega_1]$ against $[\omega_0] - [\omega_1]$ and of $[\omega_2]$ against $[\omega_0] - [\omega_2]$. In our context, let $[\omega_1]$ refer to $J(A) \cap M(S)$ and let $[\omega_2]$ refer to $J(S) \cap M(B)$. From standard results for G^2 with nested models, G^2 for testing the fit of $J(A) \cap M(B)$ can be partitioned into (1) G^2 for testing this model against the alternative of $J(S) \cap M(B)$, plus (2) G^2 for testing $J(S) \cap M(B)$ against $J(S) \cap M(S)$. If $[\omega_1]$ and $[\omega_2]$ are separable, then the first statistic is asymptotically equivalent to G^2 for testing the fit of the separate joint model, $J(A)$. Thus, assuming separability, G^2 for the simultaneous model is the sum of the G^2 components for the separate joint and marginal models.

Aitchison provided a sufficient condition for $[\omega_1]$ and $[\omega_2]$ to be asymptotically separable. In our context, let $\mathbf{h}_1 = \mathbf{0}$ denote the constraints from the set $\mathbf{h}(\boldsymbol{\xi}) = \mathbf{0}$ that pertain to the joint model and let $\mathbf{h}_2 = \mathbf{0}$ denote the constraints from this set that pertain to the marginal model. Let $\mathbf{H}_i = \partial\mathbf{h}_i(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$, $i = 1, 2$, and let $\mathbf{B}$ denote the information matrix under $[\omega_1] \cap [\omega_2]$. Aitchison's condition states that $\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_2 = \mathbf{0}$ for all $\boldsymbol{\xi}$ satisfying $[\omega_1] \cap [\omega_2]$. He used this condition to show asymptotic equivalence of the Wald statistic for $[\omega_1] \cap [\omega_2]$ and the sum of the Wald statistics for testing $[\omega_1]$ and $[\omega_2]$ separately. The asymptotic equivalence of Wald and likelihood-ratio statistics implies the corresponding result for G^2 statistics.

When $J(A)$ is an ordinary log-linear model that includes all the fixed-by-design parameters, and $M(B)$ constrains only the first-order marginal expected counts, Aitchison's condition holds. Specifically, if the log-linear model has constraints $\mathbf{h}_1 = \mathbf{0}$ of form $\mathbf{U}_1'\boldsymbol{\xi} = \mathbf{0}$, and if the marginal model has constraints $\mathbf{h}_2 = \mathbf{0}$ of form $\mathbf{U}_2'\mathbf{C}_2\log\mathbf{A}_2\mathbf{e}^{\boldsymbol{\xi}} = \mathbf{0}$, then $\mathbf{H}_1 = \mathbf{U}_1$ and $\mathbf{H}_2 = \mathbf{D}(\boldsymbol{\mu})\mathbf{A}_2'\mathbf{D}(\mathbf{A}_2\mathbf{e}^{\boldsymbol{\xi}})^{-1}\mathbf{C}_2'\mathbf{U}_2$. If the range space of $\mathbf{X}_1$ (for the joint model) contains the range space of $\mathbf{A}_2'$ (for the marginal model), straightforward calculation shows

that Aitchison's condition is satisfied. This holds for all simultaneous models considered in the next section.

Aitchison's results also imply that for a large class of simultaneous models, (1) the goodness-of-fit test for $J(S) \cap M(B)$ and the test for $J(A) \cap M(B)$ versus $J(A) \cap M(S)$ are asymptotically equivalent (provided $J(A) \cap M(B)$ holds), and (2) the goodness-of-fit test for $J(A) \cap M(S)$ and the test for $J(A) \cap M(B)$ versus $J(S) \cap M(B)$ are asymptotically equivalent (provided that $J(A) \cap M(B)$ holds). This result has practical implications for cases that are computationally complex. For instance, suppose that the joint distribution structure is of secondary interest and that one is primarily interested in testing the goodness of fit of a marginal model, $M(B)$. The joint table may contain many sampling zeros, and saturated (or nearly saturated) joint distribution models may be difficult to fit. Thus the model $J(S) \cap M(B)$ may be difficult to fit. Provided that a simpler joint distribution model holds, one has an asymptotically equivalent way to test the goodness of fit of marginal model $M(B)$ without having to fit the saturated joint distribution model.

6. EXAMPLE OF SIMULTANEOUS MODELING

We now consider simultaneous joint and marginal models for Table 1. We denote the responses by E for environment, H for health, C for assistance to big cities, and L for law enforcement. Let μ_{hijk} denote the expected frequency for the cell in level h of E, i of H, j of C, and k of L.

Simple models for the joint distribution are those that assume no three-factor interaction. These include:

$$J(I): \quad \log \mu_{hijk} = \alpha + \alpha_h^E + \alpha_i^H + \alpha_j^C + \alpha_k^L,$$

$$J(\text{2-factor}): \quad \log \mu_{hijk} = \alpha + \alpha_h^E + \alpha_i^H + \alpha_j^C + \alpha_k^L + \alpha_{hi}^{EH} + \alpha_{hj}^{EC} + \alpha_{hk}^{EL} + \alpha_{ij}^{HC} + \alpha_{ik}^{HL} + \alpha_{jk}^{CL},$$

and

$$J(L \times L): \quad \log \mu_{hijk} = \alpha + \alpha_h^E + \alpha_i^H + \alpha_j^C + \alpha_k^L + \beta^{EH} u_h v_i + \beta^{EC} u_h w_j + \beta^{EL} u_h x_k + \beta^{HC} v_i w_j + \beta^{HL} v_i x_k + \beta^{CL} w_j x_k.$$

The second model permits all two-factor associations, whereas the third model is a simpler one assuming linear-by-linear forms for those associations, for fixed monotone scores $\{u_h\}$, $\{v_i\}$, $\{w_j\}$, and $\{x_k\}$ assigned to levels of the responses. The latter type of model fits well when underlying continuous variables have joint normal distributions (Becker 1989; Goodman 1981; Holland and Wang 1987). Possible models for the marginal distributions include the cumulative logit models

$$M(H): \quad \text{logit } \gamma_h(t) = \omega_h,$$

$$M(\text{PO}): \quad \text{logit } \gamma_h(t) = \omega_h + \delta_t,$$

and

$$M(S): \quad \text{logit } \gamma_h(t) = \omega_{ht},$$

where PO denotes proportional odds. The linear-by-linear joint model and the proportional odds marginal model are parsimonious forms that reflect the ordering of the response categories. These particular models are reasonable, because all four responses are measured on the same ordinal response scale.

Table 2 contains goodness-of-fit statistics for several models. Linear-by-linear terms used equally spaced scores for the three categories, which seems reasonable for the response scale (too little, about right, too much). The majority of cells contain small counts, and the fit statistics are mainly useful for comparative purposes. Separate fitting suggested that the linear-by-linear model provides a reasonable fit for the joint distribution and that the cumulative logit model provides a reasonable fit for the marginal distributions. Thus we considered a simultaneous model containing both these forms. Table 2 shows that this model provides a substantially improved fit over models that assume independence of responses or homogeneous margins.

Because the data are sparse, we also computed adjusted residuals for a cell-by-cell comparison of observed and fitted counts. The fit and the residuals are shown in Table 3. We cannot take these residuals too literally, because the sampling design was somewhat more complicated than simple random sampling. Nevertheless, the model seems to fit reasonably well, with only 5 of 81 adjusted residuals having absolute values in the neighborhood of 2 or greater. The more complex model containing two-factor association terms of general form provides some improvement but with the loss of model parsimony, requiring four parameters to describe each association.

Because lack of fit may also result from inadequacies of the marginal model, we studied adjusted residuals for marginal proportions. These are displayed in Table 4. The lack of fit in the margins seems to reflect slightly greater observed dispersion than was predicted for the cities response and slightly less dispersion than was predicted for the law enforcement response. Comparing the observed to estimated marginal proportions, we see that the lack of fit is not severe in substantive terms.

Table 5 shows the association and marginal parameter estimates for model $J(L \times L) \cap M(PO)$. The association parameter estimates reveal quite strong positive partial associations between responses on health and responses on the environment and on law enforcement. For instance, given responses on C and L, the estimated odds that the response on E is "too much" rather than "too little" is $\exp(4 \times .499) = 7.4$ times as great when H is "too much" than when it is "too little." The marginal parameter estimates indicate substantially less support for spending on cities, particularly in relation to health and the environment. For instance, the estimated odds that the response is above any fixed level is $\exp(2.34) = 10.4$ times as high for cities as for environment.

Table 2. Goodness of Fit for Models Fitted to Table 1

Model	df	G^2	X^2
$J(S) \cap M(PO)$	3	6.2	6.0
$J(L \times L) \cap M(S)$	66	65.9	61.5
$J(L \times L) \cap M(PO)$	69	71.5	64.3
$J(L \times L) \cap M(H)$	72	519.2	455.1
$J(I) \cap M(PO)$	75	129.9	260.1

Table 3. Fit and Adjusted Cell Residuals for Simultaneous Model

Cities		1			2			3		
Law Enforcement		1	2	3	1	2	3	1	2	3
Environment	Health									
1	1	62	17	5	90	42	3	74	31	11
		(58.3)	(18.0)	(3.2)	(99.1)	(37.3)	(8.1)	(70.6)	(32.4)	(8.6)
		.79	−.28	1.13	−1.26	.94	−2.00	.74	−.31	1.00
	2	11	7	0	22	18	1	19	14	3
		(11.9)	(5.8)	(1.6)	(21.3)	(12.6)	(4.3)	(16.0)	(11.5)	(4.8)
		−.30	.55	−1.35	.18	1.69	−1.68	.93	.81	−.92
	3	2	3	1	2	0	1	1	3	1
		(1.6)	(1.2)	(.5)	(3.0)	(2.8)	(1.5)	(2.4)	(2.7)	(1.8)
		.35	1.70	.65	−.62	−1.74	−.44	−.99	.20	−.64
2	1	11	3	0	21	13	2	20	8	3
		(9.9)	(3.0)	(.5)	(22.9)	(8.6)	(1.9)	(22.4)	(10.2)	(2.7)
		.43	.02	−.76	−.47	1.64	.10	.63	−.77	.19
	2	1	4	0	6	9	0	6	5	2
		(3.3)	(1.6)	(0.4)	(8.1)	(4.8)	(1.6)	(8.3)	(6.0)	(2.5)
		−1.35	1.96	−.68	−.80	2.04	−1.31	−.89	−.43	−.33
	3	1	0	1	2	1	1	4	3	1
		(.7)	(.6)	(.2)	(1.9)	(1.8)	(.9)	(2.0)	(2.3)	(1.5)
		.33	−.76	1.55	.09	−.58	.06	1.47	.47	−.45
3	1	3	0	0	2	1	0	9	2	1
		(1.2)	(.4)	(.1)	(3.9)	(1.4)	(.3)	(5.2)	(2.3)	(.6)
		1.75	−.63	−.26	−1.03	−.38	−.57	1.99	−.24	.51
	2	1	0	0	2	1	0	4	2	0
		(.7)	(.3)	(.1)	(2.2)	(1.3)	(.5)	(3.2)	(2.3)	(.9)
		.42	−.58	−.30	−.17	−.29	−.69	.51	−.18	−1.02
	3	1	0	0	0	0	0	1	2	3
		(.2)	(.2)	(.1)	(.9)	(.8)	(.4)	(1.3)	(1.4)	(.9)
		1.58	−.44	−.29	−.97	−.92	−.68	−.26	.49	2.28

7. MINIMALLY SPECIFIED MODELS

When simultaneously modeling marginal and joint distributions, one must be concerned with possible dependence between constraints in the two components of the model and consequent poorly specified simultaneous models. This section introduces concepts that are helpful for ensuring that simultaneous models are "properly defined."

Table 4. Marginal Adjusted Residuals for Simultaneous Model

		Observed Marginal Proportions	Estimated Marginal Proportions	Adjusted Residuals
Environment	1	.732	.730	.58
	2	.211	.215	−.48
	3	.058	.055	.43
Health	1	.715	.714	.42
	2	.227	.227	.001
	3	.058	.059	−.17
Cities	1	.221	.207	1.62
	2	.395	.419	−1.65
	3	.386	.374	1.68
Law Enforcement	1	.623	.630	−2.00
	2	.311	.286	2.20
	3	.066	.084	−2.26

We represent a model with parameter space ω by $[\omega]$. Consider a joint or marginal model $[\omega]$ of form $\mathbf{C}\log\mathbf{A}\mu = \mathbf{X}\beta$ (or, equivalently, $\mathbf{U}'\mathbf{C}\log\mathbf{A}\mu = \mathbf{0}$) and an alternative model $[\omega^*]$ specified by $\mathbf{C}^*\log\mathbf{A}^*\mu = \mathbf{X}^*\mathbf{B}^*$ (or $\mathbf{U}^{*\prime}\mathbf{C}^*\log\mathbf{A}^*\mu = 0$). We say that $[\omega^*]$ is *at least as simple as* $[\omega]$, denoted by $[\omega^*] \le [\omega]$, if $\mathbf{A}^* = \mathbf{A}$ and if the range space of $\mathbf{C}'\mathbf{U}$ is a subset of (possibly equal to) the range space of $\mathbf{C}^{*\prime}\mathbf{U}^*$. For the special case $\mathbf{C} = \mathbf{C}^*$, the range space of $\mathbf{C}'\mathbf{U}$ is a subset of the range space of $\mathbf{C}^{*\prime}\mathbf{U}^*$ if and only if the range space of $\mathbf{X}^*$ is a subset of the range space of $\mathbf{X}$.

A set of constraints $\mathcal{F}: \mathbf{f}(\mu) = (f_1(\mu), \ldots, f_v(\mu))' = \mathbf{0}$ is redundant if there exists a subset of the constraints $\mathcal{F}^*: \mathbf{f}^*(\mu) = \mathbf{0}$ and also another constraint $f_c(\mu) = 0$ in $\mathcal{F} - \mathcal{F}^*$ satisfying for all $\mu \ni \mathbf{f}^*(\mu) = \mathbf{0}, f_c(\mu) = 0$.

Table 5. Parameter Estimates for Simultaneous Model

Association Parameters			Marginal Parameters		
Parameter	Estimate	Std. Error	Parameter	Estimate	Std. Error
EH	.499	.112	E	.0	—
EC	.314	.104	H	−.081	.115
EL	−.003	.112	C	−2.337	.117
HC	.052	.100	L	−.462	.120
HL	.455	.103			
CL	.199	.090			

We call a model $[\omega]$ *minimally specified* if the constraints specified in $[\omega]$ are functionally independent, in the sense that those constraints are not redundant. For instance, the simultaneous model $J(SYM) \cap M(H)$ specifying complete symmetry for the joint distribution and homogeneity for the marginal distributions is not minimally specified, because the constraints that specify $J(SYM)$ imply the constraints used to specify $M(H)$. Minimally specified models can be fitted using the algorithm described in Section 3, have residual degrees of freedom equal to the number of independent constraints, not including the model identifiability constraint (i.e., number of rows minus the number of columns in $\mathbf{X}$), and have asymptotic behavior for ML estimators as derived in Section 4.

There are broad classes of simultaneous models that are necessarily minimally specified, as shown by the following result from Lang (1992), which is easily generalizable to multiresponse variables and more than one covariate:

> Let $J(AP, BP)$ denote the log-linear joint model whereby the responses A and B are conditionally independent, given a covariate factor P. Let $M(R, OP)$ denote the log-linear marginal model whereby the response (R) is jointly independent of the occasion and the covariates (for no covariate, this is simply $M(H)$). Then, if $J \geq J(AP, BP)$, $M \geq M(R, OP)$ and both J and M are minimally specified, the simultaneous model $J \cap M$ is minimally specified.

We outline the argument used to show this result. The model $J(AP, BP)$ only constrains the marginal expected counts to satisfy the multinomial constraints. Therefore, any model $J \geq J(AP, BP)$ will not imply any marginal model constraints. On the other hand, corresponding to any set of marginal counts and any association and interaction pattern as measured using odds ratios, there exists a set of cell counts with those margins and that pattern (see, for example, Bishop, Fienberg, and Holland 1975, p. 375). Because log-linear model constraints are functions of the expected counts only through odds ratios, it follows that the marginal model constraints will not imply any joint model constraints. Also, the marginal model $M(R, OP)$ is minimally specified (assuming parameter identifiability), because the OP terms allow a perfect fit to the marginal totals that are fixed by design.

In the bivariate response case with responses A and B and no covariate, the simultaneous model $J(A, B) \cap M(H)$ is minimally specified, because the constraints that specify the joint distribution model $J(A, B)$ leave the marginal expected counts arbitrary up to the model identifiability constraints. All models discussed in the previous section were minimally specified.

8. DISCUSSION

There are two main points that we wish to make with this article. First, we have provided a more efficient way of fitting and checking the fit of a generalized class of log-linear models, $\mathbf{C} \log \mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. In particular, this methodology provides improved ways of fitting nonstandard models, such as joint models for global odds ratios and models for first- or second-order marginal distributions. Second, we have shown the utility of members of this generalized class that correspond to modeling simultaneously both joint and marginal distributions.

As a by-product of these two main points, we also want to emphasize that maximum likelihood methods have greater feasibility than is commonly recognized. In particular, for marginal modeling of multivariate categorical responses, maximum likelihood is often regarded as having limited use because of its complexity. Currently, these marginal models are routinely fitted using weighted least squares methods (e.g., Koch et al., 1977, Landis et al., 1988). However, sparse data, which are commonplace when there are multiple responses, create problems for this method. This, along with the perceived complexity of maximum likelihood methods, has partly led to alternative ways of fitting such models, such as methods using generalized estimating equations (see, for example, Liang, Zeger, and Qaqish 1992). Of course, such methodology can be used with large tables for which our algorithm is impractical, or when one prefers to specify a model without assuming a distribution for the multivariate response.

There are many situations in which simultaneous models for joint and marginal distributions may be useful. One broad application type is longitudinal studies, in which one's interest focuses on how the distribution of the response changes over time. Modeling the marginal distributions gives a "population-averaged" description of such change. In longitudinal studies, modeling dependence is often of secondary importance to modeling marginal changes, but sometimes both are relevant. An example is the modeling of social mobility. We might consider how the distribution across (upper, middle, lower) social classes changes for successive generations. We might also consider the potential for change, in terms of the degree to which social class in generation $t + 1$ depends on social class in generation t.

Models for joint and marginal distributions are also relevant in studies of inter-observer reliability. Suppose that a set of observers rate the same sample of subjects using a categorical scale. Each margin of the table refers to a different observer, and the cells present the possible combinations of responses by the observers. Good agreement between ratings by different observers requires strong association between their ratings, and similar marginal distributions. Both components are needed. For instance, the marginal distributions (summarizing relative frequencies of the possible ratings for each observer) could be identical, yet the joint ratings could reveal that the observers' judgments are statistically independent. Or there could be strong association, but one observer might systematically rate subjects a category higher than another observer. Thus to describe agreement, it is relevant to model both the joint and the marginal distributions.

Simultaneous models may also be useful for analyzing data from crossover designs. Consider a two-period, two-treatment crossover design with treatments A and B. Each subject in the study is randomly assigned to one of two sequence groups, either receiving treatment A followed by treatment B or receiving B followed by A. We then have a bivariate response with the number of covariate levels at two, the number of sequence groups. Often, interest refers primarily to comparing the marginal distributions of the two treatment responses to determine which treatment is most beneficial. But it may also be important to describe the as-

sociation between the two responses. A difference in the strength of association for the two sequence groups could indicate an important difference between the two treatments, such as a carry-over effect for one of the treatments.

For Table 1, we simultaneously modeled the joint and first-order marginal distributions. One can use the same methodology to model any order of marginal distributions. For instance, in analyzing agreement among several observers, one could simultaneously model the first- and second-order marginal distributions. In modeling the second-order distributions rather than the joint distribution, one would be considering association between pairs of observers without conditioning on ratings by other observers. More generally, one could use our methodology to model simultaneously the joint distribution and the marginal distributions of every order. For most applications, we believe that the first- and second-order marginal distributions and the joint distribution would have greatest relevance.

To some readers, the marginality principle may suggest that it is more natural to construct a single model relating to the joint distribution and then consider the model for the marginal distributions implied by that model. Unfortunately, standard forms of models (e.g., log-linear models) for categorical data do not imply similar forms for the marginal distributions (Laird 1991). Hence modeling is more complex than the direct modeling of means for normally distributed responses. In some cases, the simultaneous model for the joint and marginal distributions may be equivalent to a single model for the joint distribution. For instance, the simultaneous model that specifies quasi-symmetry for the joint distribution and marginal homogeneity for the marginal distributions is simply the symmetry model for the joint distribution.

The applicability of our methodology will largely depend on the size of contingency tables to which it can be applied. One can use it for much larger tables than the standard approach (see, for example, McCullagh and Nelder 1989, p. 219), because it is not necessary to express cell probabilities in terms of model parameters. Our algorithm is also applicable to considerably larger tables than those of Haber (1985a), because of the relative simplicity of the matrices inverted. Due to the additional structure induced by simultaneous models, it may also be possible to fit them to sparse tables for which ordinary fitting procedures would be unstable for finding estimates in models having a saturated component (i.e., standard models for only the joint or marginal distribution). This is because the added structure may sufficiently "smooth" the data, thereby mitigating problems with sampling zeros. Finally, in future research it is important to generalize the methodology of this article to handle missing data.

[Received January 1993. Revised April 1993.]

REFERENCES

Aitchison, J. (1962), "Large-Sample Restricted Parametric Tests," *Journal of the Royal Statistical Society,* Ser. B, 1, 234–250.

Aitchison, J., and Silvey, S. D. (1958), "Maximum-Likelihood Estimation of Parameters Subject to Restraints," *Annals of Mathematical Statistics,* 29, 813–828.

——— (1960), "Maximum-Likelihood Estimation Procedures and Associated Tests of Significance," *Journal of the Royal Statistical Society,* Ser. B, 1, 154–171.

Becker, M. (1989), "On the Bivariate Normal Distribution and Association Models for Ordinal Categorical Data," *Statistics and Probability Letters,* 8, 435–440.

Becker, M. P., and Balagtas, C. C. (1991), "A Log-Nonlinear Model for Binary Crossover Data," unpublished technical report.

Birch, M. W. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *Journal of the Royal Statistical Society,* Ser. B, 25, 220–233.

Bishop, Y., Fienberg, S. E., and Holland, P. (1975), *Discrete Multivariate Analysis,* Cambridge, MA: MIT Press.

Dale, J. R. (1986), "Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses," *Biometrics,* 42, 909–917.

Goodman, L. A. (1981), "Association Models and the Bivariate Normal," *Biometrika,* 68, 347–355.

Haber, M. (1985a), "Maximum Likelihood Methods for Linear and Log-Linear Models in Categorical Data," *Computational Statistics & Data Analysis,* 3, 1–10.

——— (1985b), "Log-Linear Models for Correlated Marginal Totals of a Contingency Table," *Communications in Statistics, Part A—Theory and Methods,* 14, 2845–2856.

Haber, M., and Brown, M. (1986), "Maximum Likelihood Methods for Log-Linear Models When Expected Frequencies are Subject to Linear Constraints," *Journal of the American Statistical Association,* 81, 477–482.

Haberman, S. J. (1973), "The Analysis of Residuals in Cross-Classification Tables," *Biometrics,* 29, 205–220.

Holland, P. W., and Wang, Y. J. (1987), "Dependence Functions for Continuous Bivariate Densities," *Communications in Statistics, Part A—Theory and Methods,* 16, 863–876.

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.

Laird, N. M. (1991), "Topics in Likelihood-Based Methods for Longitudinal Data Analysis," *Statistica Sinica,* 1, 33–50.

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.

Lang, J. B. (1992), "On Model Fitting for Multivariate Polytomous Response Data," unpublished dissertation, University of Florida.

——— (1993), "Large Sample Behavior For a Broad Class of Multivariate Categorical Data Models," Unpublished technical report, University of Iowa.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analyses for Categorical Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 3–40.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1990), "Maximum Likelihood Regression Methods for Paired Binary Data," *Statistics in Medicine,* 9, 1517–1525.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models,* London: Chapman and Hall.

Palmgren, J. (1981), "The Fisher Information Matrix for Log-Linear Models Arguing Conditionally on Observed Explanatory Variables," *Biometrika,* 68, 563–566.

Silvey, S. D. (1959), "The Lagrange-Multiplier Test," *Annals of Mathematical Statistics,* 30, 389–407.