The Annals of Statistics
1996, Vol. 24, No. 2, 726–752

MAXIMUM LIKELIHOOD METHODS FOR A GENERALIZED CLASS OF LOG-LINEAR MODELS

BY JOSEPH B. LANG

University of Iowa

We discuss maximum likelihood methods for fitting a broad class of multivariate categorical response data models. In particular, we derive the large-sample distributions for maximum likelihood estimators of parameters of product-multinomial generalized log-linear models. The large-sample behavior of other relevant likelihood-based statistics such as goodness-of-fit statistics and adjusted residuals is also described. The asymptotic results are derived within the framework of the constraint specification, rather than the more common freedom specification, of the model. We also outline an improved fitting algorithm for computing parameter maximum likelihood estimates and other relevant statistics. The broad class of multivariate categorical response data models, which are referred to as generalized log-linear models, can imply structure on several response configuration distributions (e.g., joint and marginal distributions). These models, which include as special cases log-linear, logit and cumulative-logit models, enjoy a wide breadth of application including longitudinal, rater-agreement and crossover data analyses.

1. Introduction. In this paper, we consider maximum likelihood methods for a broad class of models useful for describing multivariate categorical response data. These models, which are referred to as generalized log-linear models (GLLM's), can be specified (in terms of the vector of cell probabilities π) as

$$\mathbf{C}\log\mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}, \qquad \text{samp}(\boldsymbol{\pi}) = \mathbf{0}, \tag{1.1}$$

where the matrices $\mathbf{C}$, $\mathbf{A}$ and $\mathbf{X}$ are of a certain nonrestrictive structure and $\text{samp}(\boldsymbol{\pi}) = \mathbf{0}$ is a multinomial sampling constraint (e.g., $\boldsymbol{\pi}'\mathbf{1} - 1 = 0$ for full-multinomial sampling). Standard log- and logit-linear models [Bishop, Fienberg and Holland (1975); Agresti (1990)], cumulative and adjacent-category logit models for marginal distributions [Lang and Agresti (1994)] and global cross-ratios models [Dale (1986)] are all special cases of these generalized log-linear models.

When given an opportunity to analyze multivariate categorical response data, it is often desirable to have at one's disposal a broad class of models

Received September 1994; revised May 1995.
AMS 1991 *subject classifications*. 62H17, 62E20.
Key words and phrases. Asymptotics, constraint equation, freedom equation, marginal model, multinomial distribution, multivariate categorical data, simultaneous model.

that can be used to simultaneously answer several questions about the multivariate distributions. For instance, we may wish to describe both the first-order marginal distributions and the joint distributions [cf. Lang and Agresti (1994)]. More generally, we may wish to simultaneously describe several different response configuration distributions; a response configuration is simply a collection of response variables. Generalized log-linear models are well suited for this simultaneous modeling. They can imply structure on several response configuration distributions and, hence, enjoy a wide breadth of application including longitudinal, rater-agreement and crossover data analyses.

Several papers written in the late 1950's and early 1960's by Aitchison and Silvey [e.g., Aitchison and Silvey (1958, 1960); Silvey (1959); Aitchison (1962)] laid out much of the foundation for the results produced in this paper. In those seminal papers, Aitchison and Silvey discussed likelihood methods useful for testing nonstandard hypotheses. They introduced a terminology useful for describing two different ways of specifying a model—freedom equation and constraint equation specifications. As a simple example, a log-linear model can be used to easily test the hypothesis that two binary variables (A and B) are stochastically independent. Let $\pi_{ij} = \Pr(A = i, B = j)$, $i, j = 1, 2$. Setting $\boldsymbol{\eta} = \log(\boldsymbol{\pi})$, the parameter space for the multinomial log-linear model of independence can be written as

$$\omega_I = \left\{ \boldsymbol{\pi} : \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \sum_{i=1}^{2} \sum_{j=1}^{2} \pi_{ij} = 1, \boldsymbol{\beta} \in R^3 \right\},$$

where the design matrix $\mathbf{X}$ can be specified as

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Aitchison and Silvey call this specification the *freedom equation* specification. The parameters $\{\pi_{ij}\}$ and $\boldsymbol{\beta}$ are called *model* and *freedom* parameters, respectively. Alternatively, the independence model could be specified in terms of *constraint equations*, namely,

$$\omega_I = \left\{ \boldsymbol{\pi} : \eta_{11} + \eta_{22} - \eta_{12} - \eta_{21} = 0, \sum_{i=1}^{2} \sum_{j=1}^{2} \pi_{ij} = 1 \right\}.$$

Several authors have discussed maximum likelihood methods for fitting certain models in the class of generalized log-linear models [e.g., McCullagh and Nelder (1989); Dale (1986); Becker and Balagtas (1993)]. In particular, they considered models for bivariate categorical data that implied structure on both the joint distributions and the marginal distributions. Their methods do not extend easily to the general multivariate response case. This is due in part because they use the freedom equation specification of the model and their method requires writing the joint probabilities as explicit functions of

the freedom parameters. This is generally not an easy task since, for example, marginal models utilize marginal probabilities, rather than joint probabilities to which the likelihood refers [Laird (1991)].

Haber (1985a, b) and Haber and Brown (1986) considered maximum likelihood fitting of models of the form (1.1) using the constraint specification of the model. They were thereby able to avoid the difficulty inherent in the reparameterization in terms of freedom parameters. However, their fitting algorithm was only practical for relatively small tables of counts and they did not describe the large-sample behavior of the maximum likelihood estimators.

In this paper, we explore maximum likelihood (ML) methods for fitting the broad class of generalized log-linear models. In particular, we derive the large-sample distributions of model parameter maximum likelihood estimators (MLE's). We explore this asymptotic behavior and the asymptotic behavior of other relevant statistics such as goodness-of-fit statistics and adjusted residuals within the framework of constraint, rather than freedom, specifications of the models. In an appendix of Gilula and Haberman (1986), a general expression, first introduced by Aitchison and Silvey (1958), for the asymptotic behavior of restricted multinomial estimators is given. In this paper, we give an important modification of Gilula and Haberman's expression: We parameterize the model in such a way so that the multinomial sampling constraints can be accounted for explicitly. More specifically, we prove that for this class of models the Jacobian matrix of the multinominal sampling constraints is orthogonal (with respect to a simple inner product) to the Jacobian matrix of the remaining model constraints. The resulting expressions we give for the asymptotic distributions of the restricted MLEs are convenient for several reasons, including: (1) it simplifies the expressions for the asymptotic distributions of certain statistics, (2) a comparison of asymptotic behavior under different sampling schemes (e.g., Poisson, multinomial, product-multinomial) is straightforward and (3) expressions for goodness-of-fit statistics such as the Wald statistic can be simplified and easily shown to be invariant to the sampling scheme.

In Section 2, we describe in detail the sampling schemes and generalized log-linear models that will be used throughout this paper. The (restricted) likelihood equations involving Lagrange multipliers are investigated in Section 3. We show that for a broad class of models the equations can be simplified due to the orthogonality of the Jacobian matrices for the multinomial sampling constraints and the remaining model constraints. In Section 4, the asymptotic distribution of the cell probability estimators and Lagrange multipliers is derived within the framework of a constraint model. The technique used is similar to that of Aitchison and Silvey (1958). For practical and mathematical reasons, it is often better to reparameterize the likelihood in terms of log(expected counts) rather than joint probabilities. In Section 5, we show that the ML estimators for the joint probabilities can be obtained by solving likelihood equations reparameterized in terms of log(expected counts). The asymptotic behavior of these reparameterized estimators is also investigated. The form and large-sample behavior of many relevant model assess-

ment statistics, including goodness-of-fit statistics and adjusted residuals, is investigated in Section 6. Section 7 briefly outlines an improved fitting algorithm for finding the parameter ML estimates and other relevant statistics. The algorithm, which is a modification of Haber's (1985b) algorithm, has several positive features. For example, it uses a parameterization that enables us to avoid out-of-range iterate estimates. We also argue that this algorithm can be used for relatively large problems since the matrix that is to be inverted in the Newton–Raphson iterative scheme is of a very special form; a form that is simply inverted using well-known numerical techniques. A summary and discussion is given in Section 8. As a matter of style, most of the longer and/or less germane proofs are left out of the body of the paper; they can be found in the Appendix.

2. Setup and notation. We assume that $\mathbf{Y} = \text{vec}(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K)$, where

$$\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kr})' \sim \text{independent Mult}(N_k, \boldsymbol{\pi}_k), \qquad k = 1, \ldots, K,$$

and the probability vector $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kr})'$, $k = 1, \ldots, K$. That is, the random vector $\mathbf{Y}$ is a product-multinomial random vector, each component (e.g., $\mathbf{Y}_k$) being an r-dimensional multinomial vector. The symbol $\boldsymbol{\pi}$ will represent the concatenation of each of the $\{\boldsymbol{\pi}_k\}$, namely,

$$\boldsymbol{\pi} = \text{vec}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_k).$$

The $rK \times 1$ probability vector $\boldsymbol{\pi}$ satisfies the multinomial sampling constraint

$$\text{samp}(\boldsymbol{\pi}) = \left(\bigoplus_{k=1}^{K} \mathbf{1}_r' \right) \boldsymbol{\pi} - \mathbf{1}_K = \mathbf{0},$$

where $\oplus$ is the direct-sum operator. For convenience, we will set $rK \equiv s$, so that the length of $\boldsymbol{\pi}$ is s.

The diagonal matrix $\mathbf{N}$ is defined to be $\oplus_1^K N_k \mathbf{I}_r$, where the matrix $\mathbf{I}_r$ is the $r \times r$ identity matrix. The total sample size is $n = \sum_1^K N_k$. In the following text we make use of the parameters $\boldsymbol{\mu} = \mathbf{N}\boldsymbol{\pi}$ and $\boldsymbol{\xi} = \log \mathbf{N}\boldsymbol{\pi}$. The symbol $\mathbf{D}_{\mathbf{x}}$ denotes the diagonal matrix with the components in $\mathbf{x}$ on the diagonal.

Our objective is to make model-based inferences about the probability vector $\boldsymbol{\pi}$ (or $\boldsymbol{\mu}$ or $\boldsymbol{\xi}$) based on a realization $\mathbf{y}$ of the product-multinomial vector $\mathbf{Y}$. We begin by specifying a generalized log-linear model.

2.1. *Model specification.* We will assume that a model $[\omega^{(M)}]$ with model parameter space $\omega^{(M)}$ can be specified as

$$\omega^{(M)} = \{\boldsymbol{\pi} : \mathbf{C} \log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}, \text{samp}(\boldsymbol{\pi}) = \mathbf{0}\}, \tag{2.1}$$

where $\mathbf{X}$ is some full-rank design matrix and the vector $\boldsymbol{\beta}$ is an unconstrained regression (or freedom) parameter. The model space

$$\Omega^{(M)} = \{\boldsymbol{\pi} : \text{samp}(\boldsymbol{\pi}) = \mathbf{0}\}$$

will be referred to as the saturated model space.

There is an equivalent model specification that uses so-called constraint equations [Aitchison and Silvey (1958, 1960)]. It is

$$\begin{aligned}\omega^{(M)} &= \{\boldsymbol{\pi}: \mathbf{U}'\mathbf{C}\log\mathbf{A}\boldsymbol{\pi} = \mathbf{0}, \operatorname{samp}(\boldsymbol{\pi}) = \mathbf{0}\} \\ &= \{\boldsymbol{\pi}: \mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}, \operatorname{samp}(\boldsymbol{\pi}) = \mathbf{0}\}.\end{aligned} \tag{2.2}$$

The matrix $\mathbf{U}$ is assumed to be of full column rank u and has range space that spans the space orthogonal to the range space of $\mathbf{X}$; in symbols, $R(\mathbf{U}) = R(\mathbf{X})^{\perp}$.

2.2. *Generalized log-linear model assumptions*. To emphasize the simultaneous modeling aspect of generalized log-linear models, we will assume that z different groups of response configurations are to be explicitly modeled. For instance, suppose V_1 and V_2 are two categorical responses. We may want to simultaneously model the joint and marginal distributions corresponding to the following two groups of response configurations: $\{(V_1, V_2)\}$ and $\{(V_1), (V_2)\}$. In general, we will assume that the model matrices satisfy the following five nonrestrictive assumptions:

ASSUMPTION A1. $\mathbf{C} = \oplus_{i=1}^{z}\mathbf{C}_i$, $\mathbf{C}_i = \oplus_{k=1}^{K}\mathbf{C}_{ik}$ and $\mathbf{C}_{ik} \equiv \mathbf{C}_{i1}$ is a $q_i/K \times m_i/K$ contrast, zero or identity matrix ($z \geq 1$).

ASSUMPTION A2. $\mathbf{A}' = (\mathbf{A}'_1, \ldots, \mathbf{A}'_z)$, $\mathbf{A}_i = \oplus_{k=1}^{K}\mathbf{A}_{ik}$ and $\mathbf{A}_{ik} \equiv \mathbf{A}_{i1}$ is $m_i/K \times r$.

ASSUMPTION A3. $\mathbf{X} = \oplus_{i=1}^{z}\mathbf{X}_i$, where $\mathbf{X}_i$ is a full column rank design matrix ($q_i \times p_i$).

ASSUMPTION A4. If $\mathbf{C}_i = \mathbf{I}_{q_i}$, then $R(\mathbf{X}_i) \supseteq R(\oplus_{k=1}^{K}\mathbf{1}_{q_i/K})$.

ASSUMPTION A5. The $s \times u$ matrix $\mathbf{F}(\boldsymbol{\pi}) \equiv \partial\mathbf{f}(\boldsymbol{\pi})'/\partial\boldsymbol{\pi} = \mathbf{A}'\mathbf{D}_{\mathbf{A}\boldsymbol{\pi}}^{-1}\mathbf{C}'\mathbf{U}$ is of full column rank u for all $\boldsymbol{\pi} \in \Omega^{(M)}$.

Assumptions A1–A3 imply that the models for the z groups of response configuration distributions may have different forms, but that for a particular group of response configurations the same model is used across the K levels of the covariate. For the special case when $\mathbf{C}_i$ is an identity matrix, assumption A4 implies that there must exist a set of columns in $\mathbf{X}_i$ that spans a space containing the range space of $\oplus_{k=1}^{K}\mathbf{1}_{q_i/K}$. For standard log-linear models, this condition is met whenever the model includes a parameter for each of the K multinomials. In general, this assumption will allow us to equivalently fit the model assuming Poisson, rather than multinomial, sampling. The last assumption, A5, implies that the constraints $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{U}'\mathbf{C}\log\mathbf{A}\boldsymbol{\pi} = \mathbf{0}$ are nonredundant. A model satisfying all five assumptions (A1–A5) is said more simply to satisfy Assumptions A.

3. Solution to the likelihood equations. In this section, we find the restricted likelihood equations and show that for models that satisfy Assumptions A, the likelihood equations can be simplified.

Let the (kernel of the) multinomial log likelihood be denoted by

$$l^{(M)}(\boldsymbol{\pi};\mathbf{y}) = \mathbf{y}' \log \boldsymbol{\pi}, \qquad \boldsymbol{\pi} \in \Omega^{(M)}. \tag{3.1}$$

Our objective is to find the estimate $\hat{\boldsymbol{\pi}}$ in $\omega^{(M)} \subseteq \Omega^{(M)}$ that maximizes the multinomial log likelihood in (3.1). That is, we must find

$$\hat{\boldsymbol{\pi}} \in \omega^{(M)} \ni \sup_{\boldsymbol{\pi} \in \omega^{(M)}} l^{(M)}(\boldsymbol{\pi};\mathbf{y}) = \sup_{\boldsymbol{\pi} \in \omega^{(M)}} \mathbf{y}' \log \boldsymbol{\pi} = \mathbf{y}' \log \hat{\boldsymbol{\pi}}. \tag{3.2}$$

Assuming that $\hat{\boldsymbol{\pi}}$ is unique, Aitchison and Silvey (1958) show that it is consistent and that it is the solution to the restricted likelihood equations

$$\begin{bmatrix} \dfrac{\partial}{\partial \boldsymbol{\pi}} l^{(M)}(\boldsymbol{\pi};\mathbf{y}) - \dfrac{\partial}{\partial \boldsymbol{\pi}} \left\{ \boldsymbol{\tau}' \left[\left(\bigoplus_{k=1}^{K} \mathbf{1}_r' \right) \boldsymbol{\pi} - \mathbf{1}_K \right] \right\} + \dfrac{\partial}{\partial \boldsymbol{\pi}} (\boldsymbol{\lambda}' \mathbf{f}(\boldsymbol{\pi})) \\ \mathbf{f}(\boldsymbol{\pi}) \\ \left(\bigoplus_{k=1}^{K} \mathbf{1}_r' \right) \boldsymbol{\pi} - \mathbf{1}_K \end{bmatrix} = \mathbf{0}$$

or, after differentiating,

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1} \mathbf{y} - \left(\bigoplus_{k=1}^{K} \mathbf{1}_r \right) \boldsymbol{\tau} + \mathbf{F}(\boldsymbol{\pi}) \boldsymbol{\lambda} \\ \mathbf{f}(\boldsymbol{\pi}) \\ \left(\bigoplus_{k=1}^{K} \mathbf{1}_r' \right) \boldsymbol{\pi} - \mathbf{1}_K \end{bmatrix} = \mathbf{0}, \tag{3.3}$$

where $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ are $K \times 1$ and $u \times 1$ vectors of undetermined multipliers corresponding to $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}$ and $\text{samp}(\boldsymbol{\pi}) = \mathbf{0}$, respectively; $\mathbf{F}(\boldsymbol{\pi}) = \partial \mathbf{f}(\boldsymbol{\pi})' / \partial \boldsymbol{\pi}$ is the $s \times u$ matrix of derivatives (see Assumption A5).

The following lemma will be used to show that for models that satisfy Assumptions A, the solution to (3.3) can be found using a simpler set of equations. The proof can be found in the Appendix.

LEMMA 3.1. *If $F(\boldsymbol{\pi})$ is the derivative matrix corresponding to a model that satisfies Assumptions A, then*

$$\left(\bigoplus_{k=1}^{K} \boldsymbol{\pi}_k' \right) \mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}, \qquad \forall \boldsymbol{\pi} \in \Omega^{(M)}.$$

This lemma states that the Jacobian matrices $\mathbf{F}(\boldsymbol{\pi})$ and $\oplus_{k=1}^{K} \mathbf{1}_r$ are orthogonal with respect to $\mathbf{D}_{\boldsymbol{\pi}}$—that is, $(\oplus_{k=1}^{K} \mathbf{1}_r') \mathbf{D}_{\boldsymbol{\pi}} \mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$—since $\oplus_{k=1}^{K} \mathbf{1}_r'$

can be written as $(\oplus_{k=1}^{K}\boldsymbol{\pi}_k')\mathbf{D}_{\boldsymbol{\pi}}^{-1}$. This orthogonality proves to be important for several reasons. For example, it follows that the $s \times (u + K)$ matrix

$$[\mathbf{F}(\boldsymbol{\pi}), \partial\, \text{samp}(\boldsymbol{\pi})'/\partial\boldsymbol{\pi}] = \left[\mathbf{F}(\boldsymbol{\pi}), \bigoplus_{k=1}^{K}\mathbf{1}_r\right] \text{ is of full column rank } u + K. \tag{3.4}$$

Notice that this means Assumptions A imply that the constraints in $[\mathbf{f}(\boldsymbol{\pi}), \text{samp}(\boldsymbol{\pi})] = \mathbf{0}$ are nonredundant. Another consequence of this orthogonality is that the following theorem holds. The proof can be found in the Appendix.

THEOREM 3.1. *Suppose that the model can be specified as in* (2.1) *or equivalently* (2.2) *and that it satisfies Assumptions A. Let* $\text{vec}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\tau}})$ *be the solution to the* $s + u + K$ *equations in* (3.3). *It follows that the subvector* $\text{vec}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ *is the solution to the following reduced set of* $s + u$ *equations*:

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \mathbf{N}\mathbf{1}_s + \mathbf{F}(\boldsymbol{\pi})\boldsymbol{\lambda} \\ \mathbf{f}(\boldsymbol{\pi}) \end{bmatrix} = \mathbf{0}. \tag{3.5}$$

Theorem 3.1 shows that, for the generalized log-linear models that satisfy Assumptions A, the multinomial sampling constraints can be accounted for explicitly. Not only does this simplify our search for the MLE, but also results in important simplications of several statistics and the specification of their asymptotic distributions.

4. Asymptotic behavior of multinomial estimators. All of the asymptotics in this section hold as $n \to \infty$ in such a way so that

$$n^{-1}\mathbf{N} \to \mathbf{W} = \bigoplus_{k=1}^{K} w_k\mathbf{I}_r, \quad \text{as } n \to \infty, \tag{4.1}$$

where $0 < w_k \leq 1$, $k = 1, \ldots, K$. That is, the relative sample size N_k/n converges to some positive constant w_k as n, the total sample size, gets large. The number of independent multinomials (or covariate levels) K is considered fixed. For notational convenience, we will let the symbol $\boldsymbol{\pi}$ represent both an arbitrary element of some set $\omega^{(M)} \subseteq \Omega^{(M)}$ and the true unknown parameter value. Which of the two the symbol stands for should be clear from the context. As an example, we might say that the difference between the sample proportions $\mathbf{N}^{-1}\mathbf{Y} = \mathbf{p}$ and the true parameter $\boldsymbol{\pi}$ converges in probability to zero.

Define the multinomial "score" vector to be

$$\mathbf{s}(\boldsymbol{\pi}; \mathbf{y}) = \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \mathbf{N}\mathbf{1}_s$$

and notice that

$$n^{-1}\mathbf{s}(\boldsymbol{\pi}; \mathbf{Y}) = n^{-1}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{Y} - n^{-1}\mathbf{N}\mathbf{1}_s = \mathbf{D}_{\boldsymbol{\pi}}^{-1}n^{-1}\mathbf{N}(\mathbf{p} - \boldsymbol{\pi}) \to_P 0 \tag{4.2}$$

and

$$
\begin{aligned}
n^{-1}\frac{\partial \mathbf{s}(\boldsymbol{\pi};\mathbf{Y})}{\partial \boldsymbol{\pi}'} &= -n^{-1}\operatorname{diag}\left(\frac{\mathbf{Y}}{\boldsymbol{\pi}^2}\right) = -n^{-1}\mathbf{N}\operatorname{diag}\left(\frac{\mathbf{N}^{-1}\mathbf{Y}}{\boldsymbol{\pi}^2}\right) \\
&= -n^{-1}\mathbf{N}\left[\mathbf{D}_{\boldsymbol{\pi}}^{-1} + O_P(n^{-1/2})\right] \\
&= -\mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1} + O_P(n^{-1/2}),
\end{aligned} \tag{4.3}
$$

since $\mathbf{p} - \boldsymbol{\pi} = O_P(n^{-1/2})$ and $n^{-1}\mathbf{N} = \mathbf{W} + o(1)$. Also,

$$
n^{-1}\frac{\partial^2 \mathbf{s}(\boldsymbol{\pi};\mathbf{Y})}{\partial \boldsymbol{\pi}\,\partial \boldsymbol{\pi}'} = -n^{-1}\frac{\partial \operatorname{diag}(\mathbf{Y}/\boldsymbol{\pi}^2)}{\partial \boldsymbol{\pi}} = O_P(1), \tag{4.4}
$$

since $n^{-1}\mathbf{Y} = O_P(1)$.

The constraint function $\mathbf{f}(\boldsymbol{\pi})$ is continuous with derivative $\mathbf{F}(\boldsymbol{\pi})$ which, by A5, is of full column rank u. Finally, the matrix $\partial \mathbf{F}(\boldsymbol{\pi})/\partial \boldsymbol{\pi}' = O(1)$. These properties of $\mathbf{f}$, along with properties (4.2), (4.3), and (4.4), imply that the regularity conditions of Aitchison and Silvey (1958) hold. In that paper, they showed that under these regularity conditions, the ML solution $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ to (3.5) exists with probability going to 1. Moreover, assuming that model $[\omega^{(M)}]$ holds, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ are consistent estimators of $\boldsymbol{\pi}$ and $\mathbf{0}$ in the following sense:

$$
\hat{\boldsymbol{\pi}} - \boldsymbol{\pi} = O_P(n^{-1/2}), \tag{4.5}
$$

$$
\hat{\boldsymbol{\lambda}} = O_P(n^{1/2}). \tag{4.6}
$$

More specifically, the joint limiting distribution of these estimators is described in the next theorem. The proof, which uses the technique of Aitchison and Silvey (1958), allows us to avoid reparameterizing the probability vector $\boldsymbol{\pi}$ in terms of the freedom parameters $\boldsymbol{\beta}$. This is an important modification to standard asymptotic arguments, as this reparameterization is not usually possible for this general class of models [cf. Laird (1991); Becker and Balagtas (1993)].

THEOREM 4.1. *Suppose the model satisfies Assumptions A and that the vector* $\operatorname{vec}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ *is the solution to the multinomial likelihood equations* (3.5). *It follows that the limiting distribution of* $\operatorname{vec}[n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}), n^{-1/2}\hat{\boldsymbol{\lambda}}]$ *is multivariate normal with mean vector zero and variance–covariance matrix*

$$
\begin{bmatrix}
\left[\mathbf{D}_{\boldsymbol{\pi}} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\left(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}} - \bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\right]\mathbf{W}^{-1} & \mathbf{0} \\
\mathbf{0} & \left(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}
\end{bmatrix},
$$

where $\mathbf{F} = \mathbf{F}(\boldsymbol{\pi})$.

PROOF. The limiting distributions of (properly standardized) estimators $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ can be found using the technique of Aitchison and Silvey (1958). Our brief outline of the technique makes use of the following relationships:

$$\begin{bmatrix} n^{-1}\mathbf{s}(\hat{\boldsymbol{\pi}};\mathbf{Y}) + \mathbf{F}(\hat{\boldsymbol{\pi}})(n^{-1}\hat{\boldsymbol{\lambda}}) \\ \mathbf{f}(\hat{\boldsymbol{\pi}}) \end{bmatrix} = \mathbf{0}, \tag{4.7}$$

$$\begin{aligned} n^{-1}\mathbf{s}(\hat{\boldsymbol{\pi}},\mathbf{Y}) &= n^{-1}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) + n^{-1}\frac{\partial \mathbf{s}(\boldsymbol{\pi};\mathbf{Y})}{\partial \boldsymbol{\pi}'}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + O_P(n^{-1}) \\ &= n^{-1}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) - \mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + O_P(n^{-1}), \end{aligned} \tag{4.8}$$

$$\mathbf{f}(\hat{\boldsymbol{\pi}}) = \mathbf{f}(\boldsymbol{\pi}) + \mathbf{F}(\boldsymbol{\pi})'(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + O_P(n^{-1}), \tag{4.9}$$

$$\mathbf{F}(\hat{\boldsymbol{\pi}}) = \mathbf{F}(\boldsymbol{\pi}) + O_P(n^{-1/2}) \tag{4.10}$$

and

$$\mathbf{F}(\hat{\boldsymbol{\pi}})\frac{\hat{\boldsymbol{\lambda}}}{n} = \mathbf{F}(\boldsymbol{\pi})\frac{\hat{\boldsymbol{\lambda}}}{n} + O_P(n^{-1/2})O_P(n^{-1/2}) = \mathbf{F}(\boldsymbol{\pi})\frac{\hat{\boldsymbol{\lambda}}}{n} + O_P(n^{-1}). \tag{4.11}$$

By (4.7)–(4.11), the likelihood equations evaluated at the MLE can be written as

$$\begin{aligned} \mathbf{0} &= \begin{bmatrix} n^{-1}\mathbf{s}(\hat{\boldsymbol{\pi}};\mathbf{Y}) + \mathbf{F}(\hat{\boldsymbol{\pi}})n^{-1}\hat{\boldsymbol{\lambda}} \\ \mathbf{f}(\hat{\boldsymbol{\pi}}) \end{bmatrix} \\ &= \begin{bmatrix} n^{-1}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) - \mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + \mathbf{F}(\boldsymbol{\pi})n^{-1}\hat{\boldsymbol{\lambda}} \\ \mathbf{F}(\boldsymbol{\pi})'(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \end{bmatrix} + O_P(n^{-1}). \end{aligned}$$

That is,

$$\begin{bmatrix} n^{-1}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1} & -\mathbf{F}(\boldsymbol{\pi}) \\ -\mathbf{F}(\boldsymbol{\pi})' & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\pi}} - \boldsymbol{\pi} \\ n^{-1}\hat{\boldsymbol{\lambda}} \end{bmatrix} + O_P(n^{-1})$$

or, multiplying both sides of this equation by $n^{1/2}$,

$$\begin{aligned} &\begin{bmatrix} n^{-1/2}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) \\ \mathbf{0} \end{bmatrix} \\ &\quad = \begin{bmatrix} \mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1} & -\mathbf{F}(\boldsymbol{\pi}) \\ -\mathbf{F}(\boldsymbol{\pi})' & \mathbf{0} \end{bmatrix}\begin{bmatrix} n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ n^{-1/2}\hat{\boldsymbol{\lambda}} \end{bmatrix} + O_P(n^{-1/2}). \end{aligned} \tag{4.12}$$

However,

$$\begin{aligned} n^{-1/2}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y}) &= n^{1/2}\big(n^{-1}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y})\big) = n^{1/2}\mathbf{D}_{\boldsymbol{\pi}}^{-1}n^{-1}\mathbf{N}(\mathbf{p} - \boldsymbol{\pi}) \\ &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}n^{-1/2}\mathbf{N}^{1/2}\big[\mathbf{N}^{1/2}(\mathbf{p} - \boldsymbol{\pi})\big] \\ &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}^{1/2}\mathbf{N}^{1/2}(\mathbf{p} - \boldsymbol{\pi}) + o_P(1) \end{aligned}$$

and so, by the multivariate central limit theorem and Slutsky's theorem, it follows that $n^{-1/2}\mathbf{s}(\boldsymbol{\pi};\mathbf{Y})$ has a multivariate normal limiting distribution with mean vector zero and variance–covariance matrix equal to the asymp-

totic variance–covariance matrix of $\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}^{1/2}\mathbf{N}^{1/2}(\mathbf{p}-\boldsymbol{\pi})$. Specifically, the asymptotic variance–covariance matrix is

$$\begin{aligned}\operatorname{avar}\left[\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}^{1/2}\mathbf{N}^{1/2}(\mathbf{p}-\boldsymbol{\pi})\right] &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}^{1/2}\operatorname{avar}\left[\mathbf{N}^{1/2}(\mathbf{p}-\boldsymbol{\pi})\right]\mathbf{W}^{1/2}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\\ &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}^{1/2}\left[\mathbf{D}_{\boldsymbol{\pi}}-\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\right]\mathbf{W}^{1/2}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\\ &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}-\mathbf{D}_{\boldsymbol{\pi}}^{-1}\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}\\ &= \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{W}-\left(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\right)\mathbf{W}.\end{aligned}$$

Therefore, in view of (4.12), the joint limiting distribution of the estimator $\operatorname{vec}[\,n^{1/2}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}),\, n^{-1/2}\hat{\boldsymbol{\lambda}}\,]$ is multivariate normal with mean vector zero and variance–covariance matrix

$$\begin{aligned}&\begin{bmatrix}\mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1} & -\mathbf{F}(\boldsymbol{\pi})\\ -\mathbf{F}(\boldsymbol{\pi})' & \mathbf{0}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1}-\left(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\right)\mathbf{W} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\\ &\quad\times\begin{bmatrix}\mathbf{W}\mathbf{D}_{\boldsymbol{\pi}}^{-1} & -\mathbf{F}(\boldsymbol{\pi})\\ -\mathbf{F}(\boldsymbol{\pi})' & \mathbf{0}\end{bmatrix}^{-1},\end{aligned}\tag{4.13}$$

which can be shown to equal

$$\begin{bmatrix}\left[\mathbf{D}_{\boldsymbol{\pi}}-\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\left(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}-\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\right]\mathbf{W}^{-1} & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}\end{bmatrix},\tag{4.14}$$

where $\mathbf{F}=\mathbf{F}(\boldsymbol{\pi})$. □

The block diagonal form of the asymptotic variance (4.14) implies that the two statistics $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\lambda}}$ are asymptotically independent. This independence is important since the Lagrange-multiplier statistic [Silvey (1959)], which is a quadratic form in $\hat{\boldsymbol{\lambda}}$, is used to assess the goodness-of-fit of the model.

5. Solution to the reparameterized likelihood equations. In this section, we show that when the model matrices used to specify $\omega^{(M)}$ of (2.1) satisfy Assumptions A, the estimator $\hat{\boldsymbol{\xi}}$ is the restricted MLE under the reparameterized model $[\,\omega_{\xi}^{(M)}]$, where

$$\begin{aligned}\omega_{\xi}^{(M)} &= \left\{\boldsymbol{\xi}\colon \mathbf{C}\log\mathbf{A}e^{\boldsymbol{\xi}}=\mathbf{X}\boldsymbol{\beta},\ \operatorname{samp}(\mathbf{N}^{-1}e^{\boldsymbol{\xi}})=\mathbf{0}\right\}\\ &= \left\{\boldsymbol{\xi}\colon \mathbf{U}'\mathbf{C}\log\mathbf{A}e^{\boldsymbol{\xi}}=\mathbf{0},\ \operatorname{samp}(\mathbf{N}^{-1}e^{\boldsymbol{\xi}})=\mathbf{0}\right\}\end{aligned}\tag{5.1}$$

$$= \left\{\boldsymbol{\xi}\colon \mathbf{h}(\boldsymbol{\xi})=\mathbf{0},\ \operatorname{samp}(\mathbf{N}^{-1}e^{\boldsymbol{\xi}})=\mathbf{0}\right\}.\tag{5.2}$$

This means that in practice we can also find the restricted (under $[\omega^{(M)}]$) MLE of $\boldsymbol{\pi}$ by first finding the restricted (under $[\omega_{\xi}^{(M)}]$) MLE of $\boldsymbol{\xi}$ and then using the relationship $\mathbf{N}^{-1}e^{\hat{\boldsymbol{\xi}}} = \hat{\boldsymbol{\pi}}$.

The restricted MLE of $\boldsymbol{\xi}$ under $[\omega_{\xi}^{(M)}]$ may be, for practical reasons, easier to find. For example, using the $\boldsymbol{\xi}$-parameterization, we are able to (i) avoid out-of-range iterate estimates and (ii) simplify the modified Newton–Raphson iterative root-finding scheme (see Section 7). There are other reasons for using the $\boldsymbol{\xi}$-parameterization. One in particular is that the parameter $\boldsymbol{\xi}$, which is the vector of natural logs of the expected counts, is well defined for both product-multinomial and product-Poisson sampling—$\boldsymbol{\pi}$ is not. The $\boldsymbol{\xi}$-parameterization facilitates a direct comparison of multinomial and Poisson maximum likelihood estimator behavior.

Before stating and proving Theorem 5.1, it will be convenient to state several useful lemmas; their proofs are in the Appendix. The first lemma shows the relationship between the constraint functions $\mathbf{f}(\boldsymbol{\pi})$ and $\mathbf{h}(\boldsymbol{\xi})$.

LEMMA 5.1. *Assume that the model matrices satisfy Assumptions A. Then, for* $\boldsymbol{\xi} = \log \mathbf{N}\boldsymbol{\pi}$, *we have that*

$$\mathbf{h}(\boldsymbol{\xi}) = \mathbf{f}(\boldsymbol{\pi}),$$

where the constraint functions $\mathbf{f}$ *and* $\mathbf{h}$ *are those used to specify* (2.2) *and* (5.2), *respectively.*

A consequence of Lemma 5.1 is that the following set equivalence holds:

$$\begin{aligned} \omega^{(M)} &\equiv \{\boldsymbol{\pi}: \mathbf{C}\log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}, \operatorname{samp}(\boldsymbol{\pi}) = \mathbf{0}\} \\ &= \{\boldsymbol{\pi}: \mathbf{C}\log \mathbf{AN}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}, \operatorname{samp}(\boldsymbol{\pi}) = \mathbf{0}\} \equiv \omega_N^{(M)}. \end{aligned} \tag{5.3}$$

The next lemma relates the two derivative matrices $\mathbf{F}(\boldsymbol{\pi})$ and $\mathbf{H}(\boldsymbol{\xi})$, where

$$\mathbf{H}(\boldsymbol{\xi}) = \frac{\partial \mathbf{h}(\boldsymbol{\xi})'}{\partial \boldsymbol{\xi}} = \mathbf{D}_{\boldsymbol{\mu}}\mathbf{A}'\mathbf{D}_{\mathbf{A}\boldsymbol{\mu}}^{-1}\mathbf{C}'\mathbf{U}. \tag{5.4}$$

LEMMA 5.2. *For models satisfying Assumptions A, the derivative matrices* $\mathbf{H}(\boldsymbol{\xi}) = \partial \mathbf{h}(\boldsymbol{\xi})'/\partial \boldsymbol{\xi}$ *and* $\mathbf{F}(\boldsymbol{\pi}) = \partial \mathbf{f}(\boldsymbol{\pi})'/\partial \boldsymbol{\pi}$ *are related according to*

$$\mathbf{H}(\boldsymbol{\xi}) = \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\boldsymbol{\pi}),$$

where $\boldsymbol{\xi} = \log \mathbf{N}\boldsymbol{\pi}$.

Lemma 5.2 shows that the derivative matrix $\mathbf{H}(\boldsymbol{\xi})$ is free of $\mathbf{N}$ and, hence, bounded as the total sample size n goes to infinity.

Consider the reparameterized model space

$$\begin{aligned} \omega_{\xi}^{(M)} &= \{\boldsymbol{\xi}: \mathbf{C}\log \mathbf{A}e^{\boldsymbol{\xi}} = \mathbf{X}\boldsymbol{\beta}, \operatorname{samp}(\mathbf{N}^{-1}e^{\boldsymbol{\xi}}) = \mathbf{0}\} \\ &= \{\boldsymbol{\xi}: \mathbf{h}(\boldsymbol{\xi}) = \mathbf{0}, \operatorname{samp}(\mathbf{N}^{-1}e^{\boldsymbol{\xi}}) = \mathbf{0}\} \end{aligned}$$

and notice, by (5.3), that $\log(\mathbf{N}\omega_N^{(M)}) = \omega_{\xi}^{(M)}$. This, along with the fact that $\omega^{(M)} = \omega_N^{(M)}$, is used to prove (see Appendix) the following lemma.

LEMMA 5.3. *Let* $\boldsymbol{\xi} = \log \mathbf{N}\boldsymbol{\pi}$ *and assume that the model matrices used to specify* (2.1) *satisfy Assumptions A. Then*,

$$\sup_{\boldsymbol{\pi} \in \omega^{(M)}} \mathbf{y}' \log \boldsymbol{\pi} = \mathbf{y}' \log \hat{\boldsymbol{\pi}} \quad \textit{if and only if} \quad \sup_{\boldsymbol{\xi} \in \omega_{\xi}^{(M)}} \mathbf{y}'\boldsymbol{\xi} = \mathbf{y}'\hat{\boldsymbol{\xi}},$$

where $\hat{\boldsymbol{\xi}} = \log \mathbf{N}\hat{\boldsymbol{\pi}}$.

THEOREM 5.1. *Suppose the model Assumptions A holds. Then the solution* $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ *to the likelihood equations* (3.5) *is* $(\mathbf{N}^{-1}e^{\hat{\boldsymbol{\xi}}}, \hat{\boldsymbol{\lambda}})$, *where* $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}})$ *is the solution to*

$$\begin{bmatrix} \mathbf{y} - e^{\boldsymbol{\xi}} + \mathbf{H}(\boldsymbol{\xi})\boldsymbol{\lambda} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0}. \tag{5.5}$$

The implication of Theorem 5.1 is that we can find the restricted MLE of $\boldsymbol{\pi}$ by finding the solution $\hat{\boldsymbol{\xi}}$ to the reparameterized likelihood equations (5.5) and then setting $\hat{\boldsymbol{\pi}} = \mathbf{N}^{-1}e^{\hat{\boldsymbol{\xi}}}$. An iterative procedure for solving this reparameterized set of likelihood equations is outlined in Section 7.

The following theorem gives the limiting distributions for several relevant statistics. These limiting distributions are derived under the assumption that the data are product-multinomial.

THEOREM 5.2. *Suppose that the model matrices satisfy Assumptions A and that the counts are product-multinomial. Then the following results hold*:

(i) $n^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) \to_d \text{MVN}(\mathbf{0}, \mathbf{W}^{-1}\mathbf{D}_{\boldsymbol{\pi}}^{-1} - \mathbf{W}^{-1}\mathbf{F}(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{W}^{-1} - (\oplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r')\mathbf{W}^{-1})$.

(ii) $n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to_d \text{MVN}(\mathbf{0}, \mathbf{D}_{\boldsymbol{\pi}}\mathbf{W} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}} - \mathbf{W} \oplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k')$.

(iii) $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \text{MVN}(\mathbf{0}, \mathbf{Z}_X[\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}} - \mathbf{W} \oplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k']\mathbf{Z}_X')$,

where $\mathbf{Z}_X = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}_{\mathbf{A}\mathbf{W}\boldsymbol{\pi}}^{-1}\mathbf{A}$. *Here* $\mathbf{F} = \mathbf{F}(\boldsymbol{\pi})$ *and the freedom parameter* $\boldsymbol{\beta}$ *is from model* (5.1). *Moreover*, *each of these random variables is asymptotically independent of* $\hat{\boldsymbol{\lambda}}$, *the estimator of the Lagrange multipliers*.

Using the notation of Serfling (1980) and writing the asymptotic variances in terms of the parameters $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ for easy comparability with the asymptotic variances for the Poisson models, we have the following corollary:

COROLLARY 5.1.

(i) $\hat{\boldsymbol{\xi}} - \boldsymbol{\xi} \sim \text{AMVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}})$,
(ii) $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \sim \text{AMVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}})$,
(iii) $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \text{AMVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})$,
(iv) $\hat{\boldsymbol{\pi}} - \boldsymbol{\pi} \sim \text{AMVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}})$,
(v) $\hat{\boldsymbol{\lambda}} \sim \text{AMVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\lambda}}})$,

where

$$\Sigma_{\hat{\xi}} = \mathbf{D}_{\mu}^{-1} - \mathbf{D}_{\mu}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{D}_{\mu}^{-1} - \bigoplus_{k=1}^{K}\frac{\mathbf{1}_r\mathbf{1}_r'}{N_k}, \tag{5.6}$$

$$\Sigma_{\hat{\mu}} = \mathbf{D}_{\mu} - \mathbf{H}\left(\mathbf{H}'\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}' - \bigoplus_{k=1}^{K}\frac{\boldsymbol{\mu}_k\boldsymbol{\mu}'_k}{N_k}, \tag{5.7}$$

$$\Sigma_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}_{\mathbf{A}\mu}^{-1}\mathbf{A}\Sigma_{\hat{\mu}}\mathbf{A}'\mathbf{D}_{\mathbf{A}\mu}^{-1}\mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{5.8}$$

$$\Sigma_{\hat{\pi}} = \mathbf{N}^{-1}\left[\mathbf{D}_{\mu} - \mathbf{H}\left(\mathbf{H}'\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}' - \bigoplus_{k=1}^{K}\frac{\boldsymbol{\mu}_k\boldsymbol{\mu}'_k}{N_k}\right]\mathbf{N}^{-1}, \tag{5.9}$$

$$\Sigma_{\hat{\lambda}} = \left(\mathbf{H}'\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1}, \tag{5.10}$$

and $\mathbf{H} = \mathbf{H}(\boldsymbol{\xi})$ *of* (5.4).

PROOF. The result is an immediate consequence of Theorem 4.1, Theorem 5.2, the fact that $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{D}_{\pi}^{-1}\mathbf{H}(\boldsymbol{\xi}) = \mathbf{N}\mathbf{D}_{\mu}^{-1}\mathbf{H}(\boldsymbol{\xi})$ and that $n^{-1}\mathbf{N} = \mathbf{W} + o(1)$. For example, consider the asymptotic variance of $\hat{\boldsymbol{\xi}}$:

$$\begin{aligned}
&n^{-1}\left[\mathbf{W}^{-1}\mathbf{D}_{\pi}^{-1} - \mathbf{W}^{-1}\mathbf{F}\left(\mathbf{F}'\mathbf{D}_{\pi}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{W}^{-1} - \left(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\right)\mathbf{W}^{-1}\right] \\
&\quad = n^{-1}\Bigg[\mathbf{W}^{-1}\mathbf{D}_{\pi}^{-1} - \mathbf{W}^{-1}\mathbf{D}_{\pi}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}^{-1}\mathbf{D}_{\pi}^{-1}\mathbf{H}\right)^{-1} \\
&\qquad\qquad \times\mathbf{H}'\mathbf{D}_{\pi}^{-1}\mathbf{W}^{-1} - \left(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\right)\mathbf{W}^{-1}\Bigg] \\
&\quad = \mathbf{W}^{-1}n^{-1}\mathbf{N}\mathbf{D}_{\mu}^{-1} - \mathbf{W}^{-1}n^{-1}\mathbf{N}\mathbf{D}_{\mu}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}^{-1}n^{-1}\mathbf{N}\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1} \\
&\qquad \times\mathbf{H}'\mathbf{D}_{\mu}^{-1}n^{-1}\mathbf{N}\mathbf{W}^{-1} - \left(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\right)\mathbf{N}^{-1}\mathbf{N}n^{-1}\mathbf{W}^{-1} \\
&\quad = \mathbf{D}_{\mu}^{-1} - \mathbf{D}_{\mu}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{D}_{\mu}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{D}_{\mu}^{-1} - \bigoplus_{k=1}^{K}\frac{\mathbf{1}_r\mathbf{1}_r'}{N_k} + o(n^{-1}).
\end{aligned}$$

The other results (2)–(5) follow in the same way. □

These expressions (5.6)–(5.10), allow for simple comparison with the behavior under Poisson sampling. For example, the estimated asymptotic variance of the Poisson estimator of $\boldsymbol{\mu}$ can be shown to be $\hat{\boldsymbol{\Sigma}}_{\hat{\mu}} + \oplus_{k=1}^{K}(\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}'_k/N_k)$.

6. Assessing model goodness-of-fit. In this section, we derive simplified forms of certain goodness-of-fit statistics—those statistics that can be

used to test the appropriateness of the model $[\omega_\xi^{(M)}]$ of (5.1)—and describe the large-sample behavior of generalized adjusted residuals. Aitchison and Silvey (1958, 1960), Silvey (1959) and Aitchison (1962) give the form for the three asymptotically equivalent goodness-of-fit statistics—the likelihood ratio (G^2), the Wald (W^2) and the Lagrange-multiplier (L^2) statistics. All three of these statistics will have null limiting distributions that are central chi-square with degrees of freedom equal to u, the length of the constraint function $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{h}(\boldsymbol{\xi})$.

The statistics have the form

$$G^2 = 2\mathbf{Y}' \log(\bar{\boldsymbol{\pi}}/\hat{\boldsymbol{\pi}}),$$

$$W^2 = \mathbf{h}(\bar{\boldsymbol{\xi}})' \left[\overline{\operatorname{avar}}(\mathbf{h}(\bar{\boldsymbol{\xi}}))\right]^{-1} \mathbf{h}(\bar{\boldsymbol{\xi}}),$$

$$L^2 = \hat{\boldsymbol{\lambda}}' \left[\widehat{\operatorname{avar}}(\hat{\boldsymbol{\lambda}})\right]^{-1} \hat{\boldsymbol{\lambda}},$$

where the careted symbols represent the MLE's under the model which is being tested, $[\omega_\xi^{(M)}]$, and the barred symbols represent the MLE's under the saturated model.

Now, assuming that the model $[\omega_\xi^{(M)}]$ holds so that the constraint function evaluated at the true parameter value is zero [i.e., $\mathbf{h}(\boldsymbol{\xi}) = \mathbf{0}$],

$$\mathbf{h}(\bar{\boldsymbol{\xi}}) = \mathbf{h}(\boldsymbol{\xi}) + \mathbf{H}(\boldsymbol{\xi})'(\bar{\boldsymbol{\xi}} - \boldsymbol{\xi}) + O_P(n^{-1}) = \mathbf{H}(\boldsymbol{\xi})'(\bar{\boldsymbol{\xi}} - \boldsymbol{\xi}) + O_P(n^{-1}).$$

We can find the asymptotic variance of $\mathbf{h}(\bar{\boldsymbol{\xi}})$ using the delta method. It is

$$\begin{aligned}
\operatorname{avar}(\mathbf{h}(\bar{\boldsymbol{\xi}})) &= \mathbf{H}(\boldsymbol{\xi})' \operatorname{avar}(\bar{\boldsymbol{\xi}}) \mathbf{H}(\boldsymbol{\xi}) \\
&= \mathbf{H}(\boldsymbol{\xi})' \left(\mathbf{D}_{\boldsymbol{\mu}}^{-1} - \bigoplus_{k=1}^{K} \frac{\mathbf{1}_r \mathbf{1}_r'}{N_k} \right) \mathbf{H}(\boldsymbol{\xi}) \\
&= \mathbf{H}(\boldsymbol{\xi})' \mathbf{D}_{\boldsymbol{\mu}}^{-1} \mathbf{H}(\boldsymbol{\xi}),
\end{aligned}$$

since $(\oplus_{k=1}^{K} \mathbf{1}_r \mathbf{1}_r'/N_k)\mathbf{H}(\boldsymbol{\xi}) = (\oplus_{k=1}^{K} \mathbf{1}_r/N_k)(\oplus_{k=1}^{K} \boldsymbol{\pi}_k')\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$ by Lemmas 3.1 and 5.2. Also, $\widehat{\operatorname{avar}}(\hat{\boldsymbol{\lambda}}) = [\mathbf{H}(\hat{\boldsymbol{\xi}})' \mathbf{D}_{\hat{\boldsymbol{\mu}}}^{-1} \mathbf{H}(\hat{\boldsymbol{\xi}})]^{-1}$ by Corollary 5.1. Hence, the goodness-of-fit statistics can be written more explicitly as

$$G^2 = 2\mathbf{Y}' \log(\mathbf{Y}/\hat{\boldsymbol{\mu}}), \tag{6.1}$$

$$W^2 = \mathbf{h}(\bar{\boldsymbol{\xi}})' \left[\mathbf{H}(\bar{\boldsymbol{\xi}})' \mathbf{D}_{\bar{\boldsymbol{\mu}}}^{-1} \mathbf{H}(\bar{\boldsymbol{\xi}})\right]^{-1} \mathbf{h}(\bar{\boldsymbol{\xi}}), \tag{6.2}$$

$$L^2 = \hat{\boldsymbol{\lambda}}' \mathbf{H}(\hat{\boldsymbol{\xi}})' \mathbf{D}_{\hat{\boldsymbol{\mu}}}^{-1} \mathbf{H}(\hat{\boldsymbol{\xi}}) \hat{\boldsymbol{\lambda}}. \tag{6.3}$$

REMARK 1. The Poisson goodness-of-fit statistics are numerically equal to the corresponding multinomial test statistics (6.1), (6.2) and (6.3).

REMARK 2. By likelihood equations (5.5), it is evident that

$$X^2 = (\mathbf{y} - \hat{\boldsymbol{\mu}})' \mathbf{D}_{\hat{\boldsymbol{\mu}}}^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}) = \hat{\boldsymbol{\lambda}}' \mathbf{H}(\hat{\boldsymbol{\xi}})' \mathbf{D}_{\hat{\boldsymbol{\mu}}}^{-1} \mathbf{H}(\hat{\boldsymbol{\xi}}) \hat{\boldsymbol{\lambda}} = L^2.$$

That is, the Pearson chi-squared and Lagrange-multiplier statistics are numerically equal for these models.

It is usually desirable to investigate more closely the fit of a model using cell residuals. A nice feature of our fitting and descriptive method is that adjusted cell residuals [Haberman (1973)] for these generalized log-linear models are simple to compute; they are calculated using matrices that are a by-product of the fitting algorithm. Specifically, Haberman (1973) defined *adjusted residuals* as

$$r_i = \frac{(\mathbf{y} - \hat{\boldsymbol{\mu}})_i}{\widehat{\operatorname{ase}}[(\mathbf{Y} - \hat{\boldsymbol{\mu}})_i]},$$

where $(\mathbf{x})_i$ is the ith component of the vector $\mathbf{x}$. These residuals have many nice properties. For example, they more closely resemble standard normal deviates than their competitors (e.g., Pearson residuals, $e_i = (\mathbf{y} - \hat{\boldsymbol{\mu}})_i / \widehat{\operatorname{ase}}[(\mathbf{Y})_i]$). They are not often computed, however, since they are thought to be difficult to determine. Within this constraint setting, we show that not only are they simple to compute for standard log-linear models, but also for the broader class of generalized log-linear models.

Notice that by the likelihood equations (5.5),

$$n^{-1/2}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = -\mathbf{H}(\hat{\boldsymbol{\xi}})(n^{-1/2}\hat{\boldsymbol{\lambda}}) = -\mathbf{H}(\boldsymbol{\xi})(n^{-1/2}\hat{\boldsymbol{\lambda}}) + o_P(1),$$

since $\mathbf{H}(\hat{\boldsymbol{\xi}}) = \mathbf{H}(\boldsymbol{\xi}) + O_P(n^{-1/2})$ and $n^{-1/2}\hat{\boldsymbol{\lambda}} = O_P(1)$. By the delta method, the asymptotic variance of $(\mathbf{Y} - \hat{\boldsymbol{\mu}})$ is

$$\begin{aligned} \operatorname{avar}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) &= \mathbf{H}(\boldsymbol{\xi})\operatorname{avar}(\hat{\boldsymbol{\lambda}})\mathbf{H}(\boldsymbol{\xi})' \\ &= \mathbf{H}(\boldsymbol{\xi})\left[\mathbf{H}(\boldsymbol{\xi})'\mathbf{D}_{\boldsymbol{\mu}}^{-1}\mathbf{H}(\boldsymbol{\xi})\right]^{-1}\mathbf{H}(\boldsymbol{\xi})'. \end{aligned} \tag{6.4}$$

Therefore, the adjusted residuals have the form

$$r_i = \frac{(\mathbf{y} - \hat{\boldsymbol{\mu}})_i}{s_i}, \tag{6.5}$$

where s_i is the square root of the ith diagonal element of the estimated version of (6.4). It can be shown that for models satisfying Assumptions A, these adjusted residuals are invariant to sampling scheme—multinomial or Poisson.

For models that imply structure on marginal distributions, it may be appropriate to compute adjusted marginal-cell residuals [cf. Lang and Agresti (1994)]. These residuals are defined as

$$r_i^L = \frac{(\mathbf{L}\mathbf{y})_i - (\mathbf{L}\hat{\boldsymbol{\mu}})_i}{s_i^L},$$

where the components $(\mathbf{L}\boldsymbol{\mu})_i$ of the vector $\mathbf{L}\boldsymbol{\mu}$ are linear combinations of the cell means $\boldsymbol{\mu}$, and $s_i^L = \widehat{\operatorname{ase}}((\mathbf{L}\mathbf{Y})_i - (\mathbf{L}\hat{\boldsymbol{\mu}})_i)$ is the square root of the ith diagonal element of

$$\mathbf{L}\mathbf{H}(\hat{\boldsymbol{\xi}})\left[\mathbf{H}(\hat{\boldsymbol{\xi}})'\mathbf{D}_{\hat{\boldsymbol{\mu}}}^{-1}\mathbf{H}(\hat{\boldsymbol{\xi}})\right]^{-1}\mathbf{H}(\hat{\boldsymbol{\xi}})'\mathbf{L}'.$$

7. A maximum likelihood fitting algorithm. We outline a simple iterative technique for finding parameter MLE's and other relevant statistics for the broad class of generalized log-linear models. The algorithm

is a modification of Haber's (1985a, b) and Aitchison and Silvey's (1958) algorithms.

One modification is to reparameterize the model in terms of $\boldsymbol{\xi} = \log \boldsymbol{\mu}$ to avoid out-of-range iterate estimates. This reparameterization also results in a simplified Newton–Raphson (NR) algorithm. For models that satisfy Assumptions A, the parameter MLEs can be found by solving for $\hat{\boldsymbol{\gamma}} = \operatorname{vec}(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}})$ in

$$\mathbf{g}(\hat{\boldsymbol{\gamma}}) = \begin{bmatrix} \mathbf{y} - e^{\hat{\boldsymbol{\xi}}} + \mathbf{H}(\hat{\boldsymbol{\xi}})\hat{\boldsymbol{\lambda}} \\ \mathbf{h}(\hat{\boldsymbol{\xi}}) \end{bmatrix} = \mathbf{0}.$$

An unmodified NR iterative scheme is

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \mathbf{G}(\boldsymbol{\gamma}^{(t)})^{-1}\mathbf{g}(\boldsymbol{\gamma}^{(t)}), \qquad t = 0, 1, \ldots, \tag{7.1}$$

where the derivative matrix $\mathbf{G}(\boldsymbol{\gamma}) = \partial \mathbf{g}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}'$ has the form

$$\mathbf{G}(\gamma) = \begin{bmatrix} -\operatorname{diag}(e^{\boldsymbol{\xi}}) + \dfrac{\partial \mathbf{H}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}'}(\boldsymbol{\lambda} \otimes \mathbf{I}_s) & \mathbf{H}(\boldsymbol{\xi}) \\ \mathbf{H}(\boldsymbol{\xi})' & \mathbf{0} \end{bmatrix}.$$

Haber (1985a, b) used an analogous unmodified NR algorithm, but under the $\boldsymbol{\pi}$-parameterization. Two drawbacks to that iterative scheme were that (1) out-of-range iterate estimates (e.g., negative $\boldsymbol{\pi}$ estimates) could occur and (2) the matrix analogous to $\mathbf{G}$, which is potentially very large, does not have a simple inverse.

Aitchison and Silvey (1958) advocated a simplified algorithm that required a single inversion. It was similar to the iterative scheme

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \mathbf{G}(\boldsymbol{\gamma}^{(0)})^{-1}\mathbf{g}(\boldsymbol{\gamma}^{(t)}), \qquad t = 0, 1, \ldots . \tag{7.2}$$

Reasonable starting estimates are obtained by setting $\boldsymbol{\xi}^{(0)} = \log \mathbf{y}$ and $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. We make the slight adjustment $\boldsymbol{\xi}^{(0)} = \log(\mathbf{y} + \varepsilon)$ for some small ε when some of the observed counts y_{kj} are zero. For these starting estimates,

$$\mathbf{G}(\boldsymbol{\gamma}^{(0)}) = \begin{bmatrix} -\operatorname{diag}(e^{\boldsymbol{\xi}^{(0)}}) & \mathbf{H}(\boldsymbol{\xi}^{(0)}) \\ \mathbf{H}(\boldsymbol{\xi}^{(0)})' & \mathbf{0} \end{bmatrix}$$

and the inverse $\mathbf{G}^{-1}$ is simple to compute. Specifically,

$$\mathbf{G}(\boldsymbol{\gamma})^{-1} = \begin{bmatrix} -\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1} & \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1} \\ (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}^{-1} & (\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})^{-1} \end{bmatrix}, \tag{7.3}$$

where $\mathbf{D} = \operatorname{diag}(e^{\boldsymbol{\xi}})$ and $\mathbf{H} = \mathbf{H}(\boldsymbol{\xi})$.

The iterative scheme we advocate is a modification of both (7.1) and (7.2). In view of (7.3), to determine the inverse of $\mathbf{G}$, we need only invert the diagonal matrix $\mathbf{D} = \operatorname{diag}(e^{\boldsymbol{\xi}})$ and the symmetric positive-definite (by Assumption A5) matrix $(\mathbf{H}'\mathbf{D}^{-1}\mathbf{H})$. This simplification is a result of our choice of

parameterization and the fact that the Jacobian matrices of the multinomial sampling and remaining model constraints are orthogonal. Since there are many good numerical techniques for inverting large positive-definite matrices, it seems reasonable to invert an updated matrix $\mathbf{G}$ at each iteration. We advocate the following iterative scheme:

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \mathbf{G}^*\big(\boldsymbol{\gamma}^{(t)}\big)^{-1}\mathbf{g}\big(\boldsymbol{\gamma}^{(t)}\big), \qquad t = 0, 1, \ldots, \tag{7.4}$$

where

$$\mathbf{G}^*(\boldsymbol{\gamma}) = \begin{bmatrix} -\operatorname{diag}(e^{\boldsymbol{\xi}}) & \mathbf{H}(\boldsymbol{\xi}) \\ \mathbf{H}(\boldsymbol{\xi})' & \mathbf{0} \end{bmatrix}$$

and $\mathbf{G}^*(\boldsymbol{\gamma})^{-1}$ has form (7.3).

Following an argument Aitchison and Silvey (1958) used to motivate their iterative scheme, we provide an alternative motivation for use of (7.4). For models that are relatively close to holding, we have that $n^{-1}\hat{\boldsymbol{\lambda}}$ is close to zero by (4.6). Also, the matrix $\partial\mathbf{H}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}'$ is bounded as n gets large. Thus, when $\boldsymbol{\lambda}$ is close to $\hat{\boldsymbol{\lambda}}$, we expect the matrix $(\partial\mathbf{H}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}')(n^{-1}\boldsymbol{\lambda}\otimes\mathbf{I}_s)$ to behave like it was $o(1)$. On the other hand, the matrix $n^{-1}\operatorname{diag}(e^{\boldsymbol{\xi}}) = \mathbf{W}\mathbf{D}_{\boldsymbol{\pi}} + o(1)$. Therefore, the matrix $\mathbf{G}^*$ is the dominant part of $\mathbf{G}$ in the following approximate sense:

$$n^{-1}\mathbf{G} = n^{-1}\mathbf{G}^* + \begin{bmatrix} o(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

REMARK 3. By expression (7.3), we see that the matrices used for computing statistics such as G^2, L^2, W^2, adjusted residuals and parameter estimator variances are by-products of the iterative scheme (7.4). That is, upon convergence of (7.4), one can use the block components of the final iterate estimate $\mathbf{G}(\boldsymbol{\gamma}^{(\infty)})^{-1}$ to compute these relevant statistics.

REMARK 4. Recall that the matrix $\mathbf{H}(\boldsymbol{\xi})$ has form (5.4), which involves the matrix $\mathbf{U}$. The matrix $\mathbf{U}$ is the $q \times u$ ($q = \sum_{i=1}^{z} q_i$) full column rank matrix that has range space that spans that space orthogonal to the range space of $\mathbf{X}$. Haber (1985a) points out that $\mathbf{U}$ can be calculated as $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{W}$, where $\mathbf{W}$ is a random $q \times u$ full column rank matrix. In practice, one could generate the full column rank matrix $\mathbf{W}$ (at least with high probability) using uniform random numbers.

8. Discussion. We have considered a broad class of models, namely, generalized log-linear models, that can be used to simultaneously describe several response configuration distributions of multivariate categorical responses. Generalized log-linear models are useful in many application areas, including longitudinal, rater-agreement and crossover data analyses. The large-sample behavior of many relevant statistics, such as regression (i.e., freedom) parameter ML estimators, model parameter ML estimators, goodness-of-fit statistics and adjusted residuals is described when the sampling

scheme is product-multinomial. The technique used to derive these asymptotic distributions is similar to the approach of Aitchison and Silvey (1958); the asymptotic behavior is explored within the framework of constraint equations, rather than the more commonly used freedom equations.

In contrast, McCullagh and Nelder [(1989), Section 6.5], considered the maximum likelihood approach for a generalized log-linear model when there are two binary responses; they used freedom equations and were therefore required to reparameterize the likelihood in terms of the freedom parameters. It is evident from their discussion that a straightforward generalization of their method—the freedom equation approach—to several response variables would be difficult [see also, Laird (1991)]. In particular, the freedom equation approach is applicable only if $\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}$ is a one-to-one function of $\boldsymbol{\pi}$. Glonek and McCullagh (1995) have considered this special case for a collection of models they call multivariate logistic models. A broader class of generalized linear models for multivariate categorical data is considered in Molenberghs and Lesaffre (1994). As in Glonek and McCullagh, their approach requires explicit specification of all higher-order moments so that the link is a one-to-one function of the cell probabilities. The approach we advocate in this paper —the constraint equations approach—does not require the reparameterization of the likelihood in terms of the freedom parameters and so does not require $\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}$ to be one-to-one. This is important because many models of interest (e.g., marginal homogeneity) are simpler to specify using a many-to-one generalized log-linear model function.

Existence and uniqueness results for maximum likelihood estimators exist for many special cases of these generalized log-linear models. In this paper, we assumed the existence of a unique solution to the restricted likelihood equations. Indeed, this has been shown to be the case for many examples. For instance, Haber and Brown (1986) prove that models equivalent to generalized log-linear models that imply a log-linear model structure for the joint probabilities and a linear constraint structure for the marginal distributions afford unique solutions. Pratt (1981) addresses uniqueness of ML estimators for cumulative logit models. Also, both existence and uniqueness results are well developed for log-linear and logit models [cf. Haberman (1974)]. More generally, however, results that hold for the entire class of generalized log-linear models are currently unavailable. For example, although Aitchison and Silvey (1958) have proven that a solution to the restricted likelihood equations exists with probability going to 1, the finite sample problem is not completely resolved unless $\mathbf{A}$ is an identity matrix. When $\mathbf{A}$ is an identity matrix, a sufficiency reduction is possible and existence results for standard log-linear and logit models can be utilized [cf. Haberman (1974)]. Regarding uniqueness, results that would encompass the entire class of generalized log-linear models are not currently available. Undoubtedly, this is because the constraint set implied by constraining $\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}$ to fall in a linear manifold need not be convex.

Section 6 describes several model assessment statistics. In particular, the overall goodness of fit of a model (compared to the unrestricted model) can be

measured using any of the statistics (6.1), (6.2) or (6.3). Aitchison (1962) addresses the issue of testing a hypothesis against a restricted alternative by comparing nested models using differences between statistics of the form (6.1), (6.2) or (6.3). Model residuals can be measured in many ways. We showed that (generalized) adjusted residuals are simple to compute using by-products of the fitting algorithm outlined in Section 7. Other residual competitors are the Pearson and deviance residuals [cf. Pierce and Schafer (1986)].

The maximum likelihood fitting algorithm outlined in Section 7 represents an improvement over existing algorithms. The algorithm is insensitive to starting values, which happen to be extremely simple to find. By using the log mean (or $\boldsymbol{\xi}$) parameterization, we are able to avoid out-of-range iterate estimates. Also, for this parameterization the restricted Hessian matrix (i.e., the derivative of the restricted likelihood equations) has a very simple form and can easily be inverted using standard numerical techniques. For this reason, the algorithm can be used to fit relatively large tables; of course there are limitations. We have used the algorithm to fit tables with more than 1000 cells. Finally, as a by-product of the algorithm, one can compute several relevant statistics and their asymptotic variances.

The constraint equation approach to describing the large-sample behavior for generalized log-linear models has many other benefits. For instance, because of orthogonality conditions proved in this paper, a comparison of the Poisson and multinomial generalized log-linear models is straightforward. These comparisons represent generalizations of the results of Palmgren (1981). As a special case, an alternative to the freedom equations approach to showing Palmgren's result for standard log-linear models is straightforward [Lang (1996)]; this approach has pedagogical advantages as well as technical advantages.

APPENDIX

LEMMA 3.1. *If* $\mathbf{F}(\boldsymbol{\pi})$ *is the derivative matrix corresponding to a model that satisfies A, then*

$$\left(\bigoplus_{k=1}^{K} \boldsymbol{\pi}_k'\right)\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}, \qquad \forall\, \boldsymbol{\pi} \in \Omega^{(M)}.$$

PROOF. Using matrix derivatives [MacRae (1974)],

$$\begin{aligned}\mathbf{F}(\boldsymbol{\pi}) &= \frac{\partial}{\partial \boldsymbol{\pi}}(\log(\boldsymbol{\pi}'\mathbf{A}')\mathbf{C}'\mathbf{U}) \\ &= \frac{\partial}{\partial \boldsymbol{\pi}}(\boldsymbol{\pi}'\mathbf{A}')\left\{\left(\frac{\partial}{\partial(\mathbf{A}\boldsymbol{\pi})}\log(\boldsymbol{\pi}'\mathbf{A}')\right)\mathbf{C}'\mathbf{U} + \mathbf{0}\right\} \\ &= \mathbf{A}'\mathbf{D}_{\mathbf{A}\boldsymbol{\pi}}^{-1}\mathbf{C}'\mathbf{U}.\end{aligned} \tag{A.1}$$

Therefore, one can go through the algebra to see that

$$\begin{aligned}
\left(\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k'\right)\mathbf{F}(\boldsymbol{\pi}) &= \left(\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k'\right)\mathbf{A}'\mathbf{D}_{\mathbf{A}\boldsymbol{\pi}}^{-1}\mathbf{C}'\mathbf{U}\\
&= \left(\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k'\right)[\mathbf{A}_1',\ldots,\mathbf{A}_z']\operatorname{diag}^{-1}\begin{pmatrix}\mathbf{A}_1\boldsymbol{\pi}\\ \vdots\\ \mathbf{A}_z\boldsymbol{\pi}\end{pmatrix}\bigoplus_{i=1}^{z}(\mathbf{C}_i'\mathbf{U}_i)\\
&= \left(\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k'\right)\left[\mathbf{A}_1'\mathbf{D}_{\mathbf{A}_1\boldsymbol{\pi}}^{-1}\mathbf{C}_1'\mathbf{U}_1,\ldots,\mathbf{A}_z'\mathbf{D}_{\mathbf{A}_z\boldsymbol{\pi}}^{-1}\mathbf{C}_z'\mathbf{U}_z\right]\\
&= \left(\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k'\right)\left[\bigoplus_{k=1}^{K}\mathbf{A}_{1k}'\mathbf{D}_{\mathbf{A}_1\boldsymbol{\pi}}^{-1}\mathbf{C}_1'\mathbf{U}_1,\ldots,\bigoplus_{k=1}^{K}\mathbf{A}_{zk}'\mathbf{D}_{\mathbf{A}_z\boldsymbol{\pi}}^{-1}\mathbf{C}_z'\mathbf{U}_z\right]\\
&= \left[\bigoplus_{k=1}^{K}(\boldsymbol{\pi}_k'\mathbf{A}_{1k}')\mathbf{D}_{\mathbf{A}_1\boldsymbol{\pi}}^{-1}\mathbf{C}_1'\mathbf{U}_1,\ldots,\bigoplus_{k=1}^{K}(\boldsymbol{\pi}_k'\mathbf{A}_{zk}')\mathbf{D}_{\mathbf{A}_z\boldsymbol{\pi}}^{-1}\mathbf{C}_z'\mathbf{U}_z\right]\\
&= \left[\left(\bigoplus_{k=1}^{K}\mathbf{1}_{m_1/K}'\right)\mathbf{C}_1'\mathbf{U}_1,\ldots,\left(\bigoplus_{k=1}^{K}\mathbf{1}_{m_z/K}'\right)\mathbf{C}_z'\mathbf{U}_z\right],\\
&= \mathbf{0}
\end{aligned}$$

where the last equality holds by A1 and A4. □

THEOREM 3.1. *Suppose that the model can be specified as in* (2.1) *or equivalently* (2.2) *and that it satisfies Assumptions A. Let* $\operatorname{vec}(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\tau}})$ *be the solution to the* $s+u+K$ *equations in* (3.3). *It follows that the subvector* $\operatorname{vec}(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\lambda}})$ *is the solution to the reduced set of* $s+u$ *equations*

$$\begin{bmatrix}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y}-\mathbf{N}\mathbf{1}_s+\mathbf{F}(\boldsymbol{\pi})\boldsymbol{\lambda}\\ \mathbf{f}(\boldsymbol{\pi})\end{bmatrix}=\mathbf{0}.$$

PROOF. By Lemma 3.1, if we pre-multiply the first equation in (3.3) by $(\oplus_{k=1}^{K}\boldsymbol{\pi}_k')$, we have that [since $(\oplus_{k=1}^{K}\mathbf{1}_r')\boldsymbol{\pi}=\mathbf{1}_K$]

$$\left(\bigoplus_{k=1}^{K}\mathbf{1}_r'\right)\mathbf{y}-\boldsymbol{\tau}+\mathbf{0}=\mathbf{0}.$$

That is, the undetermined multiplier $\boldsymbol{\tau}$ corresponding to the multinomial sampling constraint is no longer undetermined. In fact, it satisfies

$$\begin{pmatrix}N_1\\ \vdots\\ N_K\end{pmatrix}=\boldsymbol{\tau}.$$

In particular, we have that $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\tau}})$ is the solution to (3.3) if and only if the vector $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ is the solution to the reduced set of likelihood equations

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \left(\bigoplus_{k=1}^{K} \mathbf{1}_r\right)\begin{pmatrix} N_1 \\ \vdots \\ N_K \end{pmatrix} + \mathbf{F}(\boldsymbol{\pi})\boldsymbol{\lambda} \\ \mathbf{f}(\boldsymbol{\pi}) \end{bmatrix} = \mathbf{0},$$

which can be written simply as

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \mathbf{N}\mathbf{1}_s + \mathbf{F}(\boldsymbol{\pi})\boldsymbol{\lambda} \\ \mathbf{f}(\boldsymbol{\pi}) \end{bmatrix} = \mathbf{0}. \qquad \square$$

LEMMA 5.1. *Assume that the model matrices satisfy Assumptions A. Then, for* $\boldsymbol{\xi} = \log \mathbf{N}\boldsymbol{\pi}$, *we have that*

$$\mathbf{h}(\boldsymbol{\xi}) = \mathbf{f}(\boldsymbol{\pi}),$$

where the constraint functions $\mathbf{f}$ *and* $\mathbf{h}$ *are those used to specify* (2.2) *and* (5.2), *respectively*.

PROOF. We start by showing that $\mathbf{AN}$ can be written as $\mathbf{N}^*\mathbf{A}$, where $\mathbf{N}^*$ is a diagonal matrix

$$\begin{aligned} \mathbf{AN} &= \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_z \end{bmatrix}\mathbf{N} = \begin{bmatrix} \bigoplus_{1}^{K} \mathbf{A}_{1k} \\ \vdots \\ \bigoplus_{1}^{K} \mathbf{A}_{zk} \end{bmatrix} \bigoplus_{1}^{K} N_k \mathbf{I}_r \\ &= \begin{bmatrix} \bigoplus_{1}^{K} \mathbf{A}_{1k} N_k \\ \vdots \\ \bigoplus_{1}^{K} \mathbf{A}_{zk} N_k \end{bmatrix} = \bigoplus_{i=1}^{z} \left(\bigoplus_{k=1}^{K} N_k \mathbf{I}_{m_i/K}\right)\mathbf{A}, \\ &= \mathbf{N}^*\mathbf{A}, \end{aligned}$$

where $\mathbf{N}^* = \oplus_{i=1}^{z}(\oplus_{k=1}^{K} N_k \mathbf{I}_{m_i/K})$ is a diagonal matrix.

Also, since Assumptions A are satisfied, the vector $\mathbf{C}\log\mathbf{N}^*\mathbf{1}_s$ is in the range space of $\mathbf{X}$. This can be seen by the following argument:

$$\mathbf{N}^*\mathbf{1}_s = \begin{bmatrix} N_1\mathbf{1}_{m_1/K} \\ \vdots \\ N_K\mathbf{1}_{m_1/K} \\ N_1\mathbf{1}_{m_2/K} \\ \vdots \\ N_K\mathbf{1}_{m_2/K} \\ \vdots \\ N_K\mathbf{1}_{m_z/K} \end{bmatrix}$$

so that the vector $\mathbf{C}\log\mathbf{N}^*\mathbf{1}_s$ can be written as

$$\left(\bigoplus_{i=1}^{z}\mathbf{C}_i\right)\log\mathbf{N}^*\mathbf{1}_s = \begin{bmatrix} \mathbf{C}_1\begin{bmatrix}\log N_1\mathbf{1}_{m_1/K}\\ \vdots \\ \log N_K\mathbf{1}_{m_1/K}\end{bmatrix} \\ \vdots \\ \mathbf{C}_z\begin{bmatrix}\log N_1\mathbf{1}_{m_z/K}\\ \vdots \\ \log N_K\mathbf{1}_{m_z/K}\end{bmatrix}\end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}\log N_1\mathbf{1}_{m_1/K} \\ \mathbf{C}_{12}\log N_2\mathbf{1}_{m_1/K} \\ \vdots \\ \mathbf{C}_{1K}\log N_K\mathbf{1}_{m_1/K} \\ \vdots \\ \mathbf{C}_{zK}\log N_K\mathbf{1}_{m_z/K}\end{bmatrix}.$$

If $\mathbf{C}_{ik}$ is a zero or contrast matrix, then, by Assumption A1, $\mathbf{C}_{ik}\log N_k\mathbf{1}_{m_i/K} = \mathbf{0}$, $k = 1,\ldots,K$. Hence,

$$\mathbf{C}_i\begin{bmatrix}\log N_1\mathbf{1}_{m_i/K}\\ \vdots \\ \log N_K\mathbf{1}_{m_i/K}\end{bmatrix} = \mathbf{0}$$

is in the range space $R(\mathbf{X}_i)$.

On the other hand,if $\mathbf{C}_{ik} = \mathbf{I}_{m_i/K}$ so that by Assumptions A, $\mathbf{C}_i = \mathbf{I}_{m_i}$, then the vector

$$\mathbf{C}_i\begin{bmatrix}\log N_1\mathbf{1}_{m_i/K}\\ \vdots \\ \log N_K\mathbf{1}_{m_i/K}\end{bmatrix} = \left(\bigoplus_{k=1}^{K}\mathbf{1}_{m_i/K}\right)\begin{bmatrix}\log N_1\\ \log N_2\\ \vdots \\ \log N_K\end{bmatrix},$$

which, by definition, is in the range space $R(\oplus_{k=1}^{K}\mathbf{1}_{m_i/K})$. However, by Assumption A4, it follows that the vector

$$\mathbf{C}_i\begin{bmatrix}\log N_1\mathbf{1}_{m_i/K}\\ \vdots\\ \log N_K\mathbf{1}_{m_i/K}\end{bmatrix}$$

is in the range space $R(\mathbf{X}_i)$. Finally, since $\mathbf{X}$ has a block diagonal form as specified in Assumption A3, we have that the vector $\mathbf{C}\log\mathbf{N}^*\mathbf{1}_s$ is in $R(\mathbf{X})$.

Therefore,

$$\begin{aligned}\mathbf{h}(\boldsymbol{\xi}) &= \mathbf{U}'\mathbf{C}\log\mathbf{A}e^{\boldsymbol{\xi}} = \mathbf{U}'\mathbf{C}\log\mathbf{A}\mathbf{N}\boldsymbol{\pi}\\ &= \mathbf{U}'\mathbf{C}\log\mathbf{N}^*\mathbf{A}\boldsymbol{\pi} = \mathbf{U}'\mathbf{C}\log\mathbf{N}^*\mathbf{1}_s + \mathbf{U}'\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}\\ &= \mathbf{U}'\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}\quad\left[\text{since }\mathbf{C}\log\mathbf{N}^*\mathbf{1}_s\in R(\mathbf{X})\right]\\ &= \mathbf{f}(\boldsymbol{\pi}).\end{aligned}$$

□

LEMMA 5.2. *For models satisfying Assumptions A, the derivative matrices* $\mathbf{H}(\boldsymbol{\xi}) = \partial\mathbf{h}(\boldsymbol{\xi})'/\partial\boldsymbol{\xi}$ *and* $\mathbf{F}(\boldsymbol{\pi}) = \partial\mathbf{f}(\boldsymbol{\pi})'/\partial\boldsymbol{\pi}$ *are related according to*

$$\mathbf{H}(\boldsymbol{\xi}) = \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\boldsymbol{\pi}),$$

where $\boldsymbol{\xi} = \log\mathbf{N}\boldsymbol{\pi}$.

PROOF.

$$\begin{aligned}\mathbf{F}(\boldsymbol{\pi}) &= \frac{\partial\mathbf{f}(\boldsymbol{\pi})'}{\partial\boldsymbol{\pi}} = \frac{\partial\boldsymbol{\xi}'}{\partial\boldsymbol{\pi}}\frac{\partial\mathbf{f}(\boldsymbol{\pi})'}{\partial\boldsymbol{\xi}}\\ &= \frac{\partial\boldsymbol{\xi}'}{\partial\boldsymbol{\pi}}\frac{\partial\mathbf{h}(\boldsymbol{\xi})'}{\partial\boldsymbol{\xi}} = \frac{\partial\boldsymbol{\xi}'}{\partial\boldsymbol{\pi}}\mathbf{H}(\boldsymbol{\xi}),\end{aligned}$$

but

$$\frac{\partial\boldsymbol{\xi}'}{\partial\boldsymbol{\pi}} = \frac{\partial\log\boldsymbol{\pi}'\mathbf{N}'}{\partial\boldsymbol{\pi}} = \mathbf{D}_{\boldsymbol{\pi}}^{-1}.$$

Hence, we have that $\mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\boldsymbol{\pi}) = \mathbf{H}(\boldsymbol{\xi})$. □

LEMMA 5.3. *Let* $\boldsymbol{\xi} = \log\mathbf{N}\boldsymbol{\pi}$ *and assume that the model matrices used to specify* (2.1) *satisfy A. Then*

$$\sup_{\boldsymbol{\pi}\in\omega^{(M)}}\mathbf{y}'\log\boldsymbol{\pi} = \mathbf{y}'\log\hat{\boldsymbol{\pi}}\quad\textit{if and only if}\quad\sup_{\boldsymbol{\xi}\in\omega_{\xi}^{(M)}}\mathbf{y}'\boldsymbol{\xi} = \mathbf{y}'\hat{\boldsymbol{\xi}},$$

where $\hat{\boldsymbol{\xi}} = \log\mathbf{N}\hat{\boldsymbol{\pi}}$.

PROOF. The solution to $\sup_{\boldsymbol{\pi}\in\omega^{(M)}}\mathbf{y}'\log\boldsymbol{\pi} = \mathbf{y}'\log\hat{\boldsymbol{\pi}}$ is identical to the solution to $\sup_{\boldsymbol{\pi}\in\omega_N^{(M)}}\mathbf{y}'\log\boldsymbol{\pi} = \mathbf{y}'\log\hat{\boldsymbol{\pi}}$ since the sets $\omega^{(M)}$ and $\omega_N^{(M)}$ are identical. Now by the invariance property of MLE's, it follows that

$$\sup_{\boldsymbol{\pi}\in\omega_N^{(M)}}\mathbf{y}'\log\boldsymbol{\pi} = \mathbf{y}'\log\hat{\boldsymbol{\pi}}\quad\text{if and only if}\quad\sup_{\boldsymbol{\xi}\in\log(\mathbf{N}\omega_N^{(M)})}\mathbf{y}'\boldsymbol{\xi} = \mathbf{y}'\hat{\boldsymbol{\xi}},$$

where $\hat{\boldsymbol{\xi}} = \log \mathbf{N}\hat{\boldsymbol{\pi}}$. However, $\log(\mathbf{N}\omega_N^{(M)}) = \omega_\xi^{(M)}$, so the conclusion of the lemma follows. □

THEOREM 5.1. *Suppose the model Assumptions A hold. Then the solution* $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\lambda}})$ *to the likelihood equations* (3.5) *is identical to* $(\mathbf{N}^{-1}e^{\hat{\xi}}, \hat{\boldsymbol{\lambda}})$, *where* $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}})$ *is the solution to*

$$\begin{bmatrix} \mathbf{y} - e^{\boldsymbol{\xi}} + \mathbf{H}(\boldsymbol{\xi})\boldsymbol{\lambda} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0}.$$

PROOF. The solution to

$$\sup_{\boldsymbol{\xi} \in \omega_\xi^{(M)}} \mathbf{y}'\boldsymbol{\xi} = \mathbf{y}'\hat{\boldsymbol{\xi}}$$

can be found following the technique used in Section 3. The solution will satisfy the Lagrangian likelihood equations

$$\begin{bmatrix} \mathbf{y} + \operatorname{diag}(e^{\boldsymbol{\xi}})\left(\bigoplus_{k=1}^{K} \mathbf{1}_r\right)\boldsymbol{\gamma} + \mathbf{H}(\boldsymbol{\xi})\phi \\ \left(\bigoplus_{k=1}^{K} \mathbf{1}_r'\right)e^{\boldsymbol{\xi}} - (N_1, N_2, \ldots, N_K)' \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0}. \tag{A.2}$$

Just as the original likelihood equations could be simplified by explicitly solving for the undetermined multipliers corresponding to the multinomial sampling constraints, these reparameterized likelihood equations can be simplified. This follows since, by Lemmas 3.1 and 5.2,

$$\left(\bigoplus_{k=1}^{K} \mathbf{1}_r'\right)\mathbf{H}(\boldsymbol{\xi}) = \left(\bigoplus_{k=1}^{K} \mathbf{1}_r'\right)\mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}(\boldsymbol{\pi}) = \left(\bigoplus_{k=1}^{K} \boldsymbol{\pi}_k'\right)\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}.$$

Thus, the parameter $\boldsymbol{\gamma}$ can be solved for by pre-multiplying the first equation in (A.2) by $\oplus_{k=1}^{K}\mathbf{1}_r'$. In fact, $\boldsymbol{\gamma} = -\mathbf{1}_K$. The simplified likelihood equations are

$$\begin{bmatrix} \mathbf{y} - e^{\boldsymbol{\xi}} + \mathbf{H}(\boldsymbol{\xi})\boldsymbol{\phi} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0}. \tag{A.3}$$

However, $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\phi}})$ solves (A.3) if and only if it solves

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \mathbf{D}_{\boldsymbol{\pi}}^{-1}e^{\boldsymbol{\xi}} + \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{H}(\boldsymbol{\xi})\boldsymbol{\phi} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0},$$

where $\boldsymbol{\pi} = \mathbf{N}^{-1}e^{\boldsymbol{\xi}}$. Now since $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{h}(\boldsymbol{\xi})$ and $\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{H}(\boldsymbol{\xi}) = \mathbf{F}(\boldsymbol{\pi})$, we can rewrite these equations as

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{y} - \mathbf{N}\mathbf{1}_s + \mathbf{F}(\boldsymbol{\pi})\boldsymbol{\phi} \\ \mathbf{f}(\boldsymbol{\pi}) \end{bmatrix} = \mathbf{0}.$$

Comparing this equation to (3.5), we see that the solution results in $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\lambda}}$. So, instead of solving the equations in (3.5), we can solve the equations

$$\begin{bmatrix} \mathbf{y} - e^{\boldsymbol{\xi}} + \mathbf{H}(\boldsymbol{\xi})\boldsymbol{\lambda} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} = \mathbf{0}.$$

Finally, by Lemma 5.3, this solution $\hat{\boldsymbol{\xi}}$ satisfies $\hat{\boldsymbol{\pi}} = \mathbf{N}^{-1}e^{\hat{\boldsymbol{\xi}}}$, where $\hat{\boldsymbol{\pi}}$ is the solution to the likelihood equations (3.5) (and hence is the MLE of $\boldsymbol{\pi}$ under $[\omega^{(M)}]$). This is what we set out to show. □

THEOREM 5.2. *Suppose that the model matrices satisfy Assumptions A and that the counts are product-multinomial. Then the following results hold*:

(i) $$n^{1/2}\big(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\big) \to_D \mathrm{MVN}\Bigg(\mathbf{0}, \mathbf{W}^{-1}\mathbf{D}_{\boldsymbol{\pi}}^{-1} - \mathbf{W}^{-1}\mathbf{F}\big(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}'\mathbf{W}^{-1} - \Bigg(\bigoplus_{k=1}^{K}\mathbf{1}_r\mathbf{1}_r'\Bigg)\mathbf{W}^{-1}\Bigg);$$

(ii) $$n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to_D \mathrm{MVN}\Bigg(\mathbf{0}, \mathbf{D}_{\boldsymbol{\pi}}\mathbf{W} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}\big(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}} - \mathbf{W}\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\Bigg);$$

(iii) $$n^{1/2}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) \to_D \mathrm{MVN}\Bigg(\mathbf{0}, \mathbf{Z}_X\Bigg[\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{F}\big(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}} - \mathbf{W}\bigoplus_{k=1}^{K}\boldsymbol{\pi}_k\boldsymbol{\pi}_k'\Bigg]\mathbf{Z}_X'\Bigg),$$

where $\mathbf{Z}_X = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}_{\mathbf{A}\mathbf{W}\boldsymbol{\pi}}^{-1}\mathbf{A}$.

Here $\mathbf{F} = \mathbf{F}(\boldsymbol{\pi})$ *and the freedom parameter* $\boldsymbol{\beta}$ *is from model* (5.1). *Moreover, each of these random variables is asymptotically independent of* $\hat{\boldsymbol{\lambda}}$, *the estimator of the Lagrange multipliers.*

PROOF. (i) Since the difference $(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})$ is identical to $(\log\hat{\boldsymbol{\pi}} - \log\boldsymbol{\pi})$, which is $O_P(n^{-1/2})$, we can write this difference as

$$\hat{\boldsymbol{\xi}} - \boldsymbol{\xi} = \mathbf{D}_{\boldsymbol{\pi}}^{-1}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + O_P(n^{-1}).$$

By Theorem 4.1, $n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ has a limiting distribution that is multivariate normal with mean vector zero and variance–covariance matrix

$$\text{(A.4)} \quad \mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1} - \mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\left(\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{D}_{\boldsymbol{\pi}}\mathbf{W}^{-1} - \bigoplus_{k=1}^{K} \boldsymbol{\pi}_k\boldsymbol{\pi}_k'\mathbf{W}^{-1}.$$

A direct application of the delta method gives the desired result.

(ii) Similar to the proof of (i), the difference $n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ can be shown to be a linear combination of $n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$. Specifically,

$$\begin{aligned} n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) &= n^{1/2}n^{-1}\mathbf{N}\mathbf{N}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ &= n^{-1}\mathbf{N}n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ &= \mathbf{W}n^{1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + o_P(1). \end{aligned}$$

An application of Slutsky's theorem and the delta method gives the desired result.

(iii) First notice that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\log(\mathbf{A}\hat{\boldsymbol{\mu}})$ and that the difference $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be written as

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\log\left(\mathbf{A}n^{-1}\hat{\boldsymbol{\mu}}\right) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\log\left(\mathbf{A}n^{-1}\boldsymbol{\mu}\right).$$

A Taylor series expansion of $\mathbf{g}(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\log(\mathbf{A}\mathbf{y})$ about the vector $\mathbf{x}$ gives

$$\mathbf{g}(\mathbf{y}) = \mathbf{g}(\mathbf{x}) + \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}'}(\mathbf{y} - \mathbf{x}) + o(\|\mathbf{y} - \mathbf{x}\|),$$

where $\partial \mathbf{g}(\mathbf{x})/\partial \mathbf{x}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{D}_{\mathbf{A}\mathbf{x}}^{-1}\mathbf{A}$ and $o(\|\mathbf{y} - \mathbf{x}\|)$ is a term that converges to 0 as $\mathbf{y}$ gets close to $\mathbf{x}$. Now set $\mathbf{y} = n^{-1}\hat{\boldsymbol{\mu}}$ and $\mathbf{x} = n^{-1}\boldsymbol{\mu}$ and notice that the remainder, $o(\|\mathbf{y} - \mathbf{x}\|) = o(O_P(n^{-1/2})) = o_P(n^{-1/2})$. Therefore,

$$\begin{aligned} n^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= n^{1/2}\left[\mathbf{g}\left(n^{-1}\hat{\boldsymbol{\mu}}\right) - \mathbf{g}\left(n^{-1}\boldsymbol{\mu}\right)\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\,\mathrm{diag}^{-1}(\mathbf{A}n^{-1}\boldsymbol{\mu})\mathbf{A}\left[n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right] + o_P(1) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\,\mathrm{diag}^{-1}(\mathbf{A}\mathbf{W}\boldsymbol{\pi})\mathbf{A}\left[n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right] + o_P(1), \end{aligned}$$

where the last equality follows since

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\,\mathrm{diag}^{-1}(\mathbf{A}n^{-1}\boldsymbol{\mu})\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\,\mathrm{diag}^{-1}(\mathbf{A}\mathbf{W}\boldsymbol{\pi}) = o(1)$$

and $n^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = O_P(1)$. An application of the delta method using (ii) gives the desired result.

Finally, it was shown that each of the random variables of (i), (ii) and (iii) could be approximated by a linear combination of $(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$. Since $(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is asymptotically independent of $\hat{\boldsymbol{\lambda}}$, it follows that these other random variables must also be asymptotically independent of $\hat{\boldsymbol{\lambda}}$. □

Acknowledgments. I am grateful to the reviewers and an Editor for their many helpful comments and suggestions.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

AITCHISON, J. (1962). Large-sample restricted parametric tests. *J. Roy. Statist. Soc. Ser. B* **24** 234–250.

AITCHISON, J. and SILVEY, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29** 813–828.

AITCHISON, J. and SILVEY, S. D. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *J. Roy. Statist. Soc. Ser. B* **22** 154–171.

BECKER, M. P. and BALAGTAS, C. C. (1993). A log-nonlinear model for binary cross-over data. *Biometrics* **49** 997–1009.

BISHOP, Y., FIENBERG, S. E. and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. MIT Press.

DALE, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42** 909–917.

GILULA, Z. and HABERMAN, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* **81** 780–798.

GLONEK, G. F. V. and MCCULLAGH, P. (1995). Multivariate logistic models. J. Roy. Statist. Soc. Ser. B **57** 533–546

HABER, M. (1985a). Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.* **3** 1–10.

HABER, M. (1985b). Log-linear models for correlated marginal totals of a contingency table. *Comm. Statist. Theory 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. *J. Amer. Statist. Assoc.* **81** 477–482.

HABERMAN, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics* **29** 205–220.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.

LAIRD, N. M. (1991). Topics of likelihood-based methods for longitudinal data analysis. *Statist. Sinica* **1** 33–50.

LANG, J. B. (1996). On the comparison of multinomial and Poisson loglinear models. *J. Roy. Statist. Soc. Ser. B* **58** 253–266.

LANG, J. B. and AGRESTI, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89** 625–632.

MACRAE, E. C. (1974). Matrix derivatives with an application to an adaptive linear decision problem. *Ann. Statist.* **2** 337–346.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

MOLENBERGHS, G. and LESAFFRE, E. (1994). Marginal modelling of multivariate categorical data. Unpublished manuscript.

PALMGREN, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68** 563–566.

PIERCE, D. A. and SCHAFER, D. W. (1986). Residuals in generalized linear models. *J. Amer. Statist. Assoc.* **81** 977–986.

PRATT, J. W. (1981). Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76** 103–106.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SILVEY, S. D. (1959). The Lagrange-multiplier test. *Ann. Math. Statist.* **30** 389–407.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242