Chapter 10

The Matrix Approach to Log-Linear Models

Analysis of variance and regression analysis are both branches of linear model theory. Regression analysis and linear model theory are usually taught using matrices. It is less common to teach analysis of variance with matrices. Although standard log-linear model theory is analogous to analysis of variance, the basic results are more easily stated in matrix notation. It is assumed that the reader is familiar with the basics of using matrices.

We begin with some simple examples of writing log-linear models with matrices.

EXAMPLE 10.0.1. Consider a 3×4 table. The log-linear model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}\,, \qquad i = 1, \ldots 3\,,\ j = 1, \ldots, 4\,, \tag{1}$$

can be written in matrix form as

$$\begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{13}) \\ \log(m_{14}) \\ \log(m_{21}) \\ \log(m_{22}) \\ \log(m_{23}) \\ \log(m_{24}) \\ \log(m_{31}) \\ \log(m_{32}) \\ \log(m_{33}) \\ \log(m_{34}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{1(3)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{2(4)} \end{bmatrix}.$$

The log-linear model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{2}$$

can be written in matrix form as

$$\begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{13}) \\ \log(m_{14}) \\ \log(m_{21}) \\ \log(m_{22}) \\ \log(m_{23}) \\ \log(m_{24}) \\ \log(m_{31}) \\ \log(m_{32}) \\ \log(m_{33}) \\ \log(m_{34}) \end{bmatrix} = X \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{1(3)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{2(4)} \\ u_{12(11)} \\ u_{12(12)} \\ u_{12(13)} \\ u_{12(14)} \\ u_{12(21)} \\ u_{12(22)} \\ u_{12(23)} \\ u_{12(24)} \\ u_{12(31)} \\ u_{12(32)} \\ u_{12(33)} \\ u_{12(34)} \end{bmatrix}$$

where

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

The uniform association model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j$$

can be written

$$\begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{13}) \\ \log(m_{14}) \\ \log(m_{21}) \\ \log(m_{22}) \\ \log(m_{23}) \\ \log(m_{24}) \\ \log(m_{31}) \\ \log(m_{32}) \\ \log(m_{33}) \\ \log(m_{34}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & x_1w_1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & x_1w_2 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & x_1w_3 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & x_1w_4 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & x_2w_1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & x_2w_2 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & x_2w_3 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & x_2w_4 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & x_3w_1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & x_3w_2 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & x_3w_3 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & x_3w_4 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{1(3)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{2(4)} \\ \gamma \end{bmatrix}.$$

A matrix with only one column will be referred to as a vector. Let $x = (x_1, \ldots, x_q)'$ be a vector. Define

$$\log(x) = (\log(x_1), \log(x_2), \ldots, \log(x_q))'.$$

Consider a table with any number of dimensions that has q cells in it. For a 3×4 table, $q = 12$. For an $I \times J \times K$ table, $q = IJK$. The expected cell counts are denoted by the vector $m = (m_1, \ldots, m_q)'$. A log-linear model is a model

$$\log(m) = X\beta$$

where $\log(m)$ is a $q \times 1$ vector of unknown parameters, X is a $q \times p$ matrix with known values (often X consists entirely of 0s and 1s), and β is a $p \times 1$ vector of unknown parameters. In Example 10.0.1, the log-linear model (1) has an X matrix with 12 rows and 8 columns that consists entirely of 0s and 1s. The β vector was the 8×1 matrix $(u, u_{1(1)}, u_{1(2)}, u_{1(3)}, u_{2(1)}, u_{2(2)}, u_{2(3)}, u_{2(4)})'$. For model (2), the X matrix has 12 rows and 20 columns. The β vector is a 20×1 matrix that contains u, the $u_{1(i)}$'s, the $u_{2(j)}$'s, and the $u_{12(ij)}$'s.

EXAMPLE 10.0.2. Consider a $2 \times 3 \times 2$ table. The model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

can be written

$$\begin{bmatrix} \log(m_{111}) \\ \log(m_{112}) \\ \log(m_{121}) \\ \log(m_{122}) \\ \log(m_{131}) \\ \log(m_{132}) \\ \log(m_{211}) \\ \log(m_{212}) \\ \log(m_{221}) \\ \log(m_{222}) \\ \log(m_{231}) \\ \log(m_{232}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{3(1)} \\ u_{3(2)} \end{bmatrix}$$

$$\log(m) \qquad = \qquad X \qquad \beta\,.$$

The model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

can be written

$$\begin{bmatrix} \log(m_{111}) \\ \log(m_{112}) \\ \log(m_{121}) \\ \log(m_{122}) \\ \log(m_{131}) \\ \log(m_{132}) \\ \log(m_{211}) \\ \log(m_{212}) \\ \log(m_{221}) \\ \log(m_{222}) \\ \log(m_{231}) \\ \log(m_{232}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{3(1)} \\ u_{3(2)} \\ u_{23(11)} \\ u_{23(12)} \\ u_{23(21)} \\ u_{23(22)} \\ u_{23(31)} \\ u_{23(32)} \end{bmatrix}$$

$$\log(m) \qquad = \qquad X \qquad \beta\,.$$

EXERCISE 10.1. For a 3×4 table, write model (7.1.7) in the form $\log(m) = X\beta$.

EXERCISE 10.2. For a $2 \times 3 \times 2$ table, write the models [2][13], [13][23], [12][23][13], and [123] in the form $\log(m) = X\beta$.

One advantage of establishing results for a general log-linear model $\log(m) = X\beta$ is the flexibility of the model. Results apply to ANOVA type models for any number of dimensions. The X matrix can be any known matrix, so models that incorporate known scores for ordered categories or use predictor variables to model interactions are also special cases of the general log-linear model.

In this chapter, we present a summary of some basic results in maximum likelihood theory for log-linear models. Most of the results are presented more rigorously in Chapter 12. In Sections 1 and 2, results are presented for multinomial sampling. In Section 3, the extension to product-multinomial sampling is discussed. Section 4 discusses drawing inferences about model parameters. Section 5 examines the Newton-Raphson alternative to iterative proportional fitting for finding MLEs. Section 6 discusses the GSK method of fitting log-linear models. Section 7 considers residual analysis.

10.1 Maximum Likelihood Theory for Multinomial Sampling

Suppose we have a table with q cells, observations $n = (n_1, \ldots, n_q)'$, and the log-linear model $\log(m) = X\beta$ holds. Under multinomial sampling, the likelihood function is

$$L(p) = \frac{n_.!}{\prod_{i=1}^{q} n_i!} \prod_{i=1}^{q} p_i^{n_i}$$

where $p = (p_1, \ldots, p_q)'$. Equivalently, we can write this as a function of m because $m_i = n_. p_i$ and $n_.$ is the known sample size. In terms of the m_i's, the likelihood becomes

$$L(m) = \frac{n_.!}{\prod_{i=1}^{q} n_i!} \prod_{i=1}^{q} (m_i/n_.)^{n_i} .$$

Estimation

Maximum likelihood estimates (MLEs) are values $\hat{m}_i$ that maximize $L(m)$ subject to the constraints of our model. There are two constraints on the model: One is the log-linear structure

$$\log(m) = X\beta \quad \text{for some} \quad \beta \tag{1}$$

and the other relates to the fact that with multinomial sampling, $1 = \sum_{i=1}^{q} p_i$. This second condition is equivalent to $n_. = m_.$. Let J be a $q \times 1$ vector consisting entirely of 1s. The condition $n_. = m_.$ can be written as

$$n'J = m'J . \tag{2}$$

Rather than maximizing $L(m)$, it is simpler to maximize the log of $L(m)$,

$$\log L(m) = \log(n.!) - \sum_{i=1}^{q} \log(n_i!) + \sum_{i=1}^{q} n_i \log(m_i) - \sum_{i=1}^{q} n_i \log(n.) .$$

The only term that involves the m_i's is $\sum_{i=1}^{q} n_i \log(m_i) = n' \log(m)$, so it is enough to maximize

$$\ell(m) \equiv n' \log(m) .$$

The MLE, $\hat{m}$, is the value that maximizes $\ell(m)$ subject to conditions (1) and (2). In other words, $\hat{m}$ must have the properties that

$$\log(\hat{m}) = X\hat{\beta} \quad \text{for some } \hat{\beta} , \tag{3}$$

$$n'J = \hat{m}'J \tag{4}$$

and if $\tilde{m}$ is any other vector with $\log(\tilde{m}) = X\tilde{\beta}$ and $n'J = \tilde{m}'J$, then

$$\ell(\tilde{m}) \leq \ell(\hat{m}) .$$

It turns out that for a broad class of possible X matrices, the maximization can be performed without imposing condition (2). As will be discussed below, this occurs because the maximum of $\ell(m)$ subject only to condition (1) automatically satisfies condition (2). A standard method for finding the maximum of $\ell(m)$ subject to condition (1) is by setting appropriate partial derivatives equal to zero. It can be shown that the partial derivatives are zero at the point $\hat{m}$ that satisfies

$$n'X = \hat{m}'X, \tag{5}$$

cf. Chapter 12. Moreover, by considering the matrix of second partial derivatives, it can be shown that if $\ell(m)$ achieves its maximum, subject to the constraint $\log(m) = X\beta$, then it will be at the unique value $\hat{m}$ that satisfies conditions (3) and (5). In other words, *any value $\hat{m}$ that satisfies the (marginal) constraints (5) and the model (3) is the maximum likelihood estimate of m provided a maximum exists.* This point was made repeatedly in Section 3.2.

If X is chosen appropriately, then any $\hat{m}$ that satisfies conditions (3) and (5) automatically satisfies condition (2), i.e., satisfies (4). Before examining this claim, we introduce a very useful concept in log-linear model theory, the column space of X. The column space of X is defined to be the set

$$C(X) = \{\mu | \mu = X\beta \quad \text{for some} \quad \beta\} .$$

Thus, $C(X)$ consists of all of the possible values for $\log(m)$ that satisfy the log-linear model. Earlier, we discussed the fact that the models

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \tag{6}$$

and

$$\log(m_{ij}) = u_{12(ij)} + u_{13(ik)} \tag{7}$$

are equivalent. If we write the model in equation (6) as

$$\log(m) = X_1\beta_1$$

and the model in equation (7) as

$$\log(m) = X_2\beta_2\,,$$

it is not difficult to show that $C(X_1) = C(X_2)$. In other words, the possible values for $\log(m)$ are identical in models (6) and (7). That is why the models are equivalent.

We now return to the claim that if $\hat{m}$ satisfies conditions (3) and (5), then for an appropriate X matrix, condition (4) is also satisfied. Suppose that $J \in C(X)$. In other words, for some vector b, $J = Xb$. If $\hat{m}$ satisfies condition (5), then it follows that

$$n'J = n'Xb = \hat{m}'Xb = \hat{m}'J;$$

hence, condition (4) is satisfied. Thus, if $J \in C(X)$ and if the MLE of m exists, then we can find the MLE by finding $\hat{m}$ that satisfies conditions (3) and (5).

We will not give a detailed discussion concerning when MLEs exist; for such a discussion, see Haberman (1974a). However, we will mention one result. It is an immediate consequence of Theorem 12.2.1 that if $n_i > 0$ for all $i = 1, \ldots, q$, then the MLEs exist.

The condition imposed above on X, i.e., $J \in C(X)$, is not an onerous condition. It simply means that the model has a parameter u (with no subscripts) or that the model is equivalent to a model that contains a u term. The condition $J \in C(X)$ is also necessary for the asymptotic results discussed in Section 2 and Chapter 12. We will henceforth always assume that $J \in C(X)$.

One final point: The MLE of m does not really depend on X, it depends on $C(X)$. Any two parametrizations $\log(m) = X_1\beta_1$ and $\log(m) = X_2\beta_2$ with $C(X_1) = C(X_2)$ have exactly the same MLE of m.

EXAMPLE 10.1.1. One version of the three-dimensional saturated model is $\log(m_{ijk}) = u_{123(ijk)}$. If this is written in matrix form, $X = I_q$ where I_q is the $q \times q$ identity matrix and $q = IJK$. The conditions for MLEs become

$$\log(\hat{m}) = I_q\hat{\beta} = \hat{\beta} \quad \text{for some} \quad \hat{\beta}$$

and

$$n'I_q = \hat{m}'I_q\,.$$

Clearly, $\hat{m} = n$ satisfies the second of these equations and $\hat{\beta} = \log(n)$ satisfies the first equation. Thus, the MLE of m is $\hat{m} = n$ in a three-dimensional saturated model. In fact, this argument is valid for any saturated model. The idea of a saturated model is that there are enough parameters to explain the data perfectly. This translates to the idea that $C(X) = C(I_q) = \mathbf{R}^q$.

EXAMPLE 10.1.2. Consider a three-dimensional table and the model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} .$$

If one writes out the matrix X (cf. Example 10.0.2), it is easily seen that the condition $n'X = m'X$ is precisely $n_{...} = \hat{m}_{...}$, $n_{i..} = \hat{m}_{i..}$, $n_{.j.} = \hat{m}_{.j.}$, $n_{..k} = \hat{m}_{..k}$, $n_{.jk} = \hat{m}_{.jk}$. Several of these conditions are redundant. It is sufficient to have $n_{i..} = \hat{m}_{i..}$ and $n_{.jk} = \hat{m}_{.jk}$. The other relationships follow from these two. If the model is reparametrized as

$$\log(m_{ijk}) = u_{1(i)} + u_{23(jk)} ,$$

then the condition $n'X = m'X$ for the new X matrix gives only the conditions $n_{i..} = \hat{m}_{i..}$ and $n_{.jk} = \hat{m}_{.jk}$. The MLEs of the m_{ijk}'s are precisely the values $\hat{m}_{ijk}$ that satisfy $n_{i..} = \hat{m}_{i..}$ and $n_{.jk} = \hat{m}_{.jk}$ and can be written as

$$\log(\hat{m}_{ijk}) = \hat{u}_{1(i)} + \hat{u}_{23(jk)}$$

for some values $\hat{u}_{1(i)}$ and $\hat{u}_{23(jk)}$. It is easily seen that if $\hat{m}_{ijk} = n_{i..}n_{.jk}/n_{...}$, these conditions are satisfied.

EXERCISE 10.3.

(a) For a 3×4 table, find the conditions that the MLEs must satisfy in model (7.1.4).

(b) Repeat (a) for model (7.1.7).

TESTING HYPOTHESES

One of the tests that is allowed for three-way tables is the test of [1][2][3] versus [1][23]. If these models are written as $\log(m) = X_0\beta_0$ and $\log(m) = X\beta$, respectively, the fact that [1][23] is a larger model is reflected in the fact that $C(X_0) \subset C(X)$. Recall that for a $2 \times 3 \times 2$ table, X_0 and X where presented in Example 10.0.2.

In general, if we assume that $\log(m) = X\beta$ holds and that $C(X_0) \subset C(X)$, we can test the hypothesis

$$H_0 : \log(m) = X_0\beta_0$$

against the hypothesis

$$H_A : (H_0 \text{ is not true}) .$$

The likelihood ratio test statistic is

$$G^2 = -2[\log L(\hat{m}_0) - \log L(\hat{m})]$$

where $\hat{m}_0$ is the MLE of m under the assumption that H_0 is true and $\hat{m}$ is the MLE under the "unrestricted" model. However, in this case, the "unrestricted" model is that $\log(m) = X\beta$. It is easily seen that

$$\begin{aligned} G^2 &= -2[\ell(\hat{m}_0) - \ell(\hat{m})] \\ &= -2[n' \log(\hat{m}_0) - n' \log(\hat{m})] \\ &= 2n'[\log(\hat{m}) - \log(\hat{m}_0)] \\ &= 2\sum_{i=1}^{q} n_i \log(\hat{m}_i/\hat{m}_{0i}) \end{aligned}$$

where again $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_q)'$ is the MLE of m for $\log(m) = X\beta$ and $\hat{m}_0 = (\hat{m}_{01}, \ldots, \hat{m}_{0q})'$ is the MLE of m under the restriction that $\log(m) = X_0\beta_0$.

In fact, G^2 can be written as

$$G^2 = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i}),$$

which is our usual form. To see this equivalence, note that $\log(\hat{m}) = X\hat{\beta}$ for some $\hat{\beta}$, and because $C(X_0) \subset C(X)$, we can write $\log(\hat{m}_0) = X_0\hat{\beta}_0 = X\hat{\gamma}$ for some $\hat{\beta}_0$ and $\hat{\gamma}$. Recall that $\hat{m}'X = n'X$. By substitution,

$$\begin{aligned} G^2 = 2n'[\log(\hat{m}) - \log(\hat{m}_0)] &= 2n'[X\hat{\beta} - X\hat{\gamma}] \\ &= 2n'X[\hat{\beta} - \hat{\gamma}] \\ &= 2\hat{m}'X[\hat{\beta} - \hat{\gamma}] \\ &= 2\hat{m}'[\log(\hat{m}) - \log(\hat{m}_0)] \\ &= 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i})\,. \end{aligned}$$

10.2 Asymptotic Results

This section presents a few of the primary asymptotic results for log-linear models under multinomial sampling and mentions some applications of those results. More precise versions of these results are available in Chapter 12.

We begin by setting some notation. Let x be a $q \times 1$ vector. $D(x)$ is used to denote the $q \times q$ diagonal matrix

$$D(x) = [d_{ij}] \qquad \text{where } d_{ii} = x_i\,,\ d_{ij} = 0, \qquad i \neq j\,.$$

One diagonal matrix is used often and has a special notation:

$$D \equiv D(p) .$$

Recall that J is a $q \times 1$ vector of 1s. Define

$$A = X(X'DX)^{-1}X'D$$

and

$$A_z = J(J'DJ)^{-1}J'D$$

where it is assumed (but not really necessary) that a parametrization $\log(m) = X\beta$ has been chosen so that the inverse of $X'DX$ exists. Note that because $D(m) = n_{\cdot}D$, D can be replaced by $D(m)$ in A and A_z without changing the resulting matrices. Note that A and A_z depend on the unknown parameters p. We can estimate A and A_z simply by estimating p.

Rather than frequently writing $\log(m)$, let

$$\mu \equiv \log(m) .$$

If $\hat{m}$ is the MLE of m, $\hat{\mu} = \log(\hat{m})$ is the MLE of the μ. This follows from the *invariance of maximum likelihood estimates*; for any parameter θ and MLE $\hat{\theta}$, the MLE of a function of θ, say $f(\theta)$, is the corresponding function of the MLE, $f(\hat{\theta})$, cf. Cox and Hinkley (1974, p. 287).

The key asymptotic results about MLEs are given in the following subsections. Throughout, let $N \equiv n_{\cdot}$.

Estimation

We begin with results about the large sample distribution of the maximum likelihood estimates.

Theorem 10.2.1. Let $\mu = X\beta$ be a log-linear model for a table with q cells. Let n be the result of a multinomial sample of N observations:

(a) For N sufficiently large, $\hat{\mu} - \mu$ has the approximate distribution $N(0, [A - A_z]D^{-1}(m))$.

(b) As N gets large, $\hat{\mu} - \mu$ converges (in probability) to zero; i.e., $\hat{\mu} - \mu \xrightarrow{P} 0$.

(c) For N sufficiently large, $\hat{m} - m$ has the approximate distribution $N(0, D(m)[A - A_z])$.

(d) As N gets large, $\hat{m}/N$ converges (in probability) to p, i.e., $N^{-1}\hat{m} \xrightarrow{P} p$.

Technically, (a) and (c) deal with convergence in distribution and are similar in spirit to the Central Limit Theorem. In Chapter 11, we will have occasion to write such results as (a) $N^{\frac{1}{2}}(\hat{\mu}-\mu) \xrightarrow{L} N(0,[A-A_z]D^{-1})$ and (c) $N^{-1/2}(\hat{m}-m) \xrightarrow{L} N(0,D[A-A_z])$. The symbol $\xrightarrow{L}$ indicates convergence in distribution. The L comes from the fact that a distribution is sometimes referred to as a distributional law or simply as a law.

One interesting aspect of Theorem 10.2.1 is that, although $\hat{\mu}-\mu$ converges to zero, $\hat{\mu}$ by itself does not converge to anything. As N gets large, $\mu = \log(m) = \log(Np)$ also gets large. Although the difference $\hat{\mu}-\mu$ gets small, we cannot say that $\hat{\mu}$ converges to μ because μ *changes with* N.

Corollary 10.2.2. If the inverse of $(X'DX)$ exists and $\hat{\beta}$ satisfies $\hat{\mu} = X\hat{\beta}$, then $\hat{\beta}-\beta$ converges (in probability) to zero.

Consider the problem of drawing asymptotic inferences about a particular cell. The parameters of interest are p_i, m_i, and μ_i. We will start from the premise that estimates of the m_i's are available. These can be obtained from iterative proportional fitting as discussed in Section 3.3 or from use of the Newton-Raphson algorithm as discussed later in Section 5. Recall that

$$\hat{m} = (\hat{m}_1, \ldots, \hat{m}_q)',$$

$$\hat{\mu} = \log(\hat{m}) = (\log(\hat{m}_1), \ldots, \log(\hat{m}_q))',$$

and

$$\hat{p} = \frac{1}{N}\hat{m} = (\hat{m}_1/N, \ldots, \hat{m}_q/N)'.$$

To use Theorem 10.2.1, we need one key result. If Y is a $q \times 1$ vector with a multivariate normal distribution, i.e.,

$$Y \sim N(\xi, \Sigma),$$

and if ρ is a $q \times 1$ vector, then the scalar random variable $\rho'Y$ has a (univariate) normal distribution. In particular,

$$\rho'Y \sim N(\rho'\xi, \rho'\Sigma\rho).$$

Let $e_i' = (0, \cdots, 0, 1, 0, \cdots, 0)$ where the 1 is in the i'th place. It follows that

$$\begin{aligned} \hat{\mu}_i - \mu_i &= e_i'(\hat{\mu}-\mu), \\ \hat{m}_i - m_i &= e_i'(\hat{m}-m), \end{aligned}$$

and

$$\hat{p}_i - p_i = e_i'(\hat{p}-p).$$

Applying Theorem 10.2.1, for N large we get the approximations

$$\begin{aligned} \hat{\mu}_i - \mu_i &\sim N\left(0, e_i'[A - A_z]D^{-1}(m)e_i\right) , \\ \hat{m}_i - m_i &\sim N\left(0, e_i'D(m)[A - A_z]e_i\right) , \end{aligned}$$

and

$$\hat{p}_i - p_i \sim N\left(0, e_i'\frac{1}{N^2}D(m)[A - A_z]e_i\right) .$$

In order to use these results, we need to be able to find or at least estimate the variances. We begin with $e_i'[A - A_z]D^{-1}(m)e_i$. This value is the i'th diagonal element of $AD^{-1}(m)$ minus the i'th diagonal element of $A_zD^{-1}(m)$. To find $e_i'AD^{-1}(m)e_i$, note that

$$D^{-1}(m)e_i = \left(\frac{1}{m_i}\right)e_i,$$

so

$$e_i'AD^{-1}(m)e_i = \frac{1}{m_i}\,e_i'Ae_i\,.$$

The value $e_i'Ae_i$ is just a_{ii}, the i'th diagonal element of A. This is precisely the leverage of the i'th case. Leverages were introduced in Section 6.7 and methods for estimating them were given. The maximum likelihood estimate of

$$e_i'AD^{-1}(m)e_i = a_{ii}/m_i$$

is

$$\hat{a}_{ii}/\hat{m}_i\,.$$

The computation of $e_i'A_zD^{-1}(m)e_i$ is even simpler. For multinomial sampling,

$$A_z \equiv J(J'DJ)^{-1}J'D = JJ'D\,.$$

This follows because $J'DJ = p_{\cdot} = 1$. Moreover,

$$\begin{aligned} A_zD^{-1}(m) &= JJ'DD^{-1}(m) \\ &= JJ'\left(\frac{1}{N}\right)D(m)D^{-1}(m) \\ &= \left(\frac{1}{N}\right)JJ', \end{aligned}$$

so

$$e_i'A_zD^{-1}(m)e_i = \frac{1}{N}\,.$$

Combining results, we see that

$$e_i'[A - A_z]D^{-1}(m)e_i = \frac{a_{ii}}{m_i} - \frac{1}{N};$$

thus,

$$\hat{\mu}_i - \mu_i \sim N\left(0, \frac{a_{ii}}{m_i} - \frac{1}{N}\right) .$$

Estimating the variance leads to the approximation

$$(\hat{\mu}_i - \mu_i) \Big/ \sqrt{\frac{\hat{a}_{ii}}{\hat{m}_i} - \frac{1}{N}} \sim N(0,1) .$$

Large sample confidence intervals for μ_i follow immediately, e.g., a 95% confidence interval has the end points

$$\hat{\mu}_i \pm 1.96 \sqrt{\frac{\hat{a}_{ii}}{\hat{m}_i} - \frac{1}{N}} .$$

The $\alpha = .10$ large sample test of $H_0 : \mu_i = \mu_{i0}$ versus $H_A : \mu_i \neq \mu_{i0}$ rejects when

$$|\hat{\mu}_i - \mu_{i0}| \Big/ \sqrt{\frac{a_{ii}}{m_i} - \frac{1}{N}} > 1.645 .$$

Similar arguments lead to the asymptotic results

$$\mathrm{Var}(\hat{m}_i) = m_i a_{ii} - m_i^2/N$$

and

$$\mathrm{Var}(\hat{p}_i) = p_i a_{ii}/N - p_i^2/N .$$

Estimating the variances yields to the large sample distributions

$$\frac{\hat{m}_i - m_i}{\sqrt{\hat{m}_i \hat{a}_{ii} - \hat{m}_i^2/N}} \sim N(0,1)$$

and

$$\frac{\hat{p}_i - p_i}{\sqrt{\hat{p}_i(\hat{a}_{ii} - \hat{p}_i)/N}} \sim N(0,1) .$$

Given the distributions, inferential procedures follow in the usual way.

Just as in regression analysis, leverages fall between zero and one and the sum of all of the leverages is precisely the degrees of freedom for the model, i.e., the rank of X. The first of these facts implies that an upper bound on the variance can always be obtained by taking $a_{ii} = 1$. This is convenient because when iterative proportional fitting has been used, finding $\hat{a}_{ii}$ requires the computation of an auxiliary regression analysis. Assuming $a_{ii} = 1$ can be highly conservative because the true a_{ii} value may be much less than one. The second fact gives some idea of the extent of overestimation using $a_{ii} = 1$. If the table has $q = 24$ cells and the model has 12 degrees of freedom, the average size of the a_{ii}'s is $12/24 = \frac{1}{2}$. Thus, the variance terms based on $a_{ii} = 1$ tend to be about twice as large as they are using the estimates $\hat{a}_{ii}$.

It is interesting to note that using the upper bound $a_{ii} = 1$ is equivalent to computing the variance under the saturated model. In the saturated model, we can take $X = I$. This implies that

$$A = I(IDI)^{-1}ID = I\,.$$

Thus, for all i under the saturated model, $a_{ii} = 1$. Clearly, the use of reduced models serves to reduce the variance of estimated cell parameters.

EXAMPLE 10.2.3. In the abortion opinion data of Chapter 3 with the model [RSO][OA] (cf. Table 6.7), the cell for nonwhite males between 18 and 25 years of age who support abortion has $\hat{m}_i = 14.52$ and $\hat{a}_{ii} = .222$. The asymptotic standard error for $\hat{m}_i$ is $\sqrt{14.52(.222) - (14.52)^2/2385} = \sqrt{3.2234 - .0884} = 1.77$. An asymptotic 95% confidence interval for m_i has end points

$$14.52 \pm 1.96(1.77)\,.$$

The interval is $(11.05, 17.99)$. Similar computations lead to a 95% confidence interval for μ_i with end points

$$2.68 \pm 1.96(.123)$$

and a 90% confidence interval for p_i with end points

$$.0061 \pm 1.645(.000742)\,.$$

Besides the parameters for individual cells, the parameters of primary interest are contrasts in the μ_i's. Contrasts in the μ_i's correspond to vectors ρ in which the elements of ρ add up to zero, i.e., $\rho' J = 0$. The simplest such contrasts are log odds, but log odds ratios, the log of ratios of odds ratios, and so on, are also contrasts in the μ_i's. All of these correspond to functions $\rho'\mu$ in which ρ has a very simple structure. Given the $\hat{m}_i$'s, there is no problem in computing $\rho'\hat{\mu} = \rho' \log(\hat{m})$. The problem is in computing the variance. Finding variances for estimated contrasts is more complicated than finding them for estimates of cell parameters because contrasts involve the covariances between the estimated cell parameters. However, the fact that we are dealing with contrasts leads to one simplification based on $\rho' J = 0$.

$$\begin{aligned}
\mathrm{Var}(\rho'\hat{\mu}) &= \rho'(A - A_z)D^{-1}(m)\rho \\
&= \rho' A D^{-1}(m)\rho - \rho' A_z D^{-1}(m)\rho \\
&= \rho' X(X'D(m)X)^{-1}X'\rho - \frac{1}{N}\rho' JJ'\rho \\
&= \rho' X(X'D(m)X)^{-1}X'\rho\,.
\end{aligned}$$

Computation of the variance requires fitting the model using the Newton-Raphson algorithm, cf. Section 5. Newton-Raphson can either be used exclusively or, if the initial fit was performed using iterative proportional fitting, the auxiliary regression model of Section 6.7 can be used. Recall that the auxiliary model requires that an ANOVA type model be reparametrized as a regression model. This is so that appropriate matrix inverses can be taken. If traditional ANOVA type models are used, a simple way to generate a regression model is to drop all u terms involving index values of 1.

EXAMPLE 10.2.4. In Example 3.2.4, we examined data on automobile injuries. We found that the model of no three-factor interaction,

$$\mu_{ijk} = \log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} ,$$

fit the data very well. Below are given the data and the estimated expected cell counts based on the model.

$n_{ijk}(\hat{m}_{ijk})$		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	350 (350.49)	150 (149.51)	60 (59.51)	112 (112.49)
Ejected (i)	Yes	26 (25.51)	23 (23.49)	19 (19.49)	80 (79.51)

The regression parametrization based on dropping u terms in which any of i, j, or k equal 1 is

$$\begin{bmatrix} \hat{\mu}_{111} \\ \hat{\mu}_{121} \\ \hat{\mu}_{112} \\ \hat{\mu}_{122} \\ \hat{\mu}_{211} \\ \hat{\mu}_{221} \\ \hat{\mu}_{212} \\ \hat{\mu}_{222} \end{bmatrix} = \begin{bmatrix} \log(350.49) \\ \log(149.51) \\ \log(59.51) \\ \log(112.49) \\ \log(25.51) \\ \log(23.49) \\ \log(19.49) \\ \log(79.51) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{\gamma}_{1(2)} \\ \hat{\gamma}_{3(2)} \\ \hat{\gamma}_{2(2)} \\ \hat{\gamma}_{13(22)} \\ \hat{\gamma}_{12(22)} \\ \hat{\gamma}_{23(22)} \end{bmatrix} .$$

Because there are 8 cells and $8 - 1$ terms in the model, there is a very simple form to the matrix necessary for obtaining asymptotic variances:

$$X\left(X'D(\hat{m})X\right)^{-1}X' =$$

$$D^{-1}(\hat{m}) - (5.52816)D^{-1}(\hat{m}) \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} [1, -1, -1, 1, -1, 1, 1, -1]\, D^{-1}(\hat{m}). \quad (1)$$

The equality can be verified by direct computation of both sides. This approach requires a good matrix manipulation computer package for computing the left-hand side. The equality can also be verified by hand using orthogonality, projection operators, and the fact that $\text{rank}(X) = 7 = q - 1$. This approach requires facility with vector space concepts, cf. Christensen (1996b, p. 276).

In Example 3.2.4, for both collisions and rollovers, we were interested in the ratio of the odds of a nonsevere injury when the driver was not ejected relative to the odds of a nonsevere injury when the driver was ejected. We proceed to find a 95% confidence interval for

$$\log(m_{11k}m_{22k}/m_{12k}m_{21k})\,.$$

Recall that, based on the model of no three-factor interaction, this log odds ratio does not depend on k. The estimate of the log odds ratio is

$$.77 = \log(2.16)\,.$$

This can be arrived at in either of two ways. For $k = 1$, define the vector $\rho_1' = (1, -1, 0, 0, -1, 1, 0, 0)$ so that

$$\begin{aligned}\rho_1'\hat{\mu} &= \rho_1' \log(\hat{m}) \\ &= \log(\hat{m}_{111}\hat{m}_{221}/\hat{m}_{121}\hat{m}_{211}) \\ &= \log[(350.49)(23.49)/(149.51)(25.51)] \\ &= \log(2.16)\,.\end{aligned}$$

Otherwise, for $k = 2$, let $\rho_2' = (0, 0, 1, -1, 0, 0, -1, 1)$ so that

$$\begin{aligned}\rho_2'\hat{\mu} &= \log(\hat{m}_{112}\hat{m}_{222}/\hat{m}_{122}\hat{m}_{212}) \\ &= \log[(59.51)(79.51)/(112.49)(19.49)] \\ &= \log(2.16)\,.\end{aligned}$$

The estimated variance is

$$\rho_j' X \left(X' D(\hat{m}) X\right)^{-1} X' \rho_j = .045.$$

This can be computed directly using matrix manipulations, or it can be computed by hand using equation (1), or it can be computed from the reported standard error of $\hat{\gamma}_{12(22)}$ using the auxiliary regression (which will be reviewed in the next example). The 95% confidence interval for the log odds ratio with k fixed has the end points

$$.77 \pm 1.96\sqrt{.045}$$

and is the interval $(.35, 1.19)$. If we exponentiate the end points, we get a 95% confidence interval for the odds ratio of

$$(1.4, 3.3)\,.$$

Thus, the evidence indicates that the odds of a nonsevere injury when the driver is not ejected are between, roughly, one and a half to three times the odds of a nonsevere injury when the driver is ejected.

EXAMPLE 10.2.5. Consider a $2 \times 3 \times 2$ table and the model

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} .$$

The design matrix X for this model is given in Example 10.0.2. Suppose we are interested in the log odds

$$\log(m_{1jk}/m_{2jk}) = \mu_{1jk} - \mu_{2jk} = u_{1(1)} - u_{1(2)} .$$

Note that in this model, the odds are the same for any values of j and k. Using the notation in Example 10.0.2, write $\rho' = (1,0,0,0,0,0,-1,0,0,0,0,0)$ so that

$$\rho'\mu = \mu_{111} - \mu_{211} = u_{1(1)} - u_{1(2)} .$$

The estimate is

$$\rho'\hat{\mu} = \log(\hat{m}_{111}) - \log(\hat{m}_{211}) = \log(\hat{m}_{111}/\hat{m}_{211}),$$

but this estimate does not depend on the last two subscripts. For any j and k,

$$\rho'\hat{\mu} = \log(\hat{m}_{1jk}/\hat{m}_{2jk}) .$$

The difficult part of the analysis is in finding the variance. The variance is most easily computed by setting the problem up as a regression analysis. Write

$$\begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{122} \\ \mu_{131} \\ \mu_{132} \\ \mu_{211} \\ \mu_{212} \\ \mu_{221} \\ \mu_{222} \\ \mu_{231} \\ \mu_{232} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_{1(2)} \\ \gamma_{2(2)} \\ \gamma_{2(3)} \\ \gamma_{3(2)} \\ \gamma_{23(22)} \\ \gamma_{23(32)} \end{bmatrix}$$

$$\mu \quad = \quad W \quad \gamma .$$

The new design matrix was arrived at by eliminating every column of the old design matrix that corresponded to a u term involving $i = 1$, $j = 1$, or $k = 1$. The estimates of the γ's are the estimates of the u's subject to the

side conditions $0 = u_{1(1)} = u_{2(1)} = u_{3(1)} = u_{23(1k)} = u_{23(j1)}$ for all j and k, and

$$\rho'\mu = \mu_{111} - \mu_{211} = \begin{cases} \rho' W\gamma = -\gamma_{1(2)} \\ \rho' X\beta = u_{1(1)} - u_{1(2)} \,. \end{cases}$$

Let

$$\begin{aligned} K &= \rho'[\hat{A} - A_z]D^{-1}(\hat{m})\rho \\ &= \rho' X(X'D(\hat{m})X)^{-1}X'\rho \\ &= \rho' W(W'D(\hat{m})W)^{-1}W'\rho \end{aligned}$$

so that, asymptotically,

$$\mathrm{Var}(\rho'\hat{\mu}) = K \,.$$

The value K is easily obtained by performing an auxiliary regression, as discussed in Section 6.7. In particular, fitting

$$Y = W\gamma + e$$

with weights $\hat{m}_i$ and dependent variable

$$y_i = \log(\hat{m}_i) + (n_i - \hat{m}_i)/\hat{m}_i,$$

the regression program *will report*

$$\mathrm{SE}(\hat{\gamma}_{1(2)}) = \sqrt{\mathrm{MSE}}\, K \,.$$

Dividing by $\sqrt{\mathrm{MSE}}$ gives the correct asymptotic standard error.

Almost any good regression program allows the user to print out the matrix

$$\mathrm{Cov}(\hat{\gamma})/\mathrm{MSE} = (W'D(\hat{m})W)^{-1} \,.$$

This is the key to obtaining asymptotic variances for log-linear models. Consider the log odds ratio

$$\begin{aligned} \log(m_{i21}m_{i32}/m_{i22}m_{i31}) &= \mu_{i21} - \mu_{i22} - \mu_{i31} + \mu_{i32} \\ &= u_{23(21)} - u_{23(22)} - u_{23(31)} + u_{23(32)} \,. \end{aligned}$$

This log odds ratio does not depend on the value of i. Picking $i = 1$ for convenience, let

$$\rho' = (0, 0, 1, -1, -1, 1, 0, 0, 0, 0, 0, 0),$$

so

$$\rho'\mu = \rho' X\beta = u_{23(21)} - u_{23(22)} - u_{23(31)} + u_{23(32)} \,.$$

In the $\mu = W\gamma$ parametrization, this becomes

$$\rho'\mu = \rho' W\gamma = \gamma_{23(32)} - \gamma_{23(22)} \,.$$

There are two ways to arrive at this result. First, one can substitute the appropriate functions of the γ's in place of the μ_{ijk}'s. This leads to

$$\begin{aligned}\rho'\mu &= \mu_{121} - \mu_{122} - \mu_{131} + \mu_{132} \\ &= [\gamma_0 + \gamma_{2(2)}] - [\gamma_0 + \gamma_{2(2)} + \gamma_{3(2)} + \gamma_{23(22)}] \\ &\quad - [\gamma_0 + \gamma_{2(3)}] + [\gamma_0 + \gamma_{2(3)} + \gamma_{3(2)} + \gamma_{23(32)}] \\ &= \gamma_{23(32)} - \gamma_{23(22)}\,.\end{aligned}$$

Second, one can notice that

$$\rho'W = (0,0,0,0,0,-1,1),$$

so that

$$\rho'\mu = \rho'W\gamma = \gamma_{23(32)} - \gamma_{23(22)}\,.$$

If we write

$$\lambda' = \rho'W,$$

then the estimated large sample variance is

$$\begin{aligned}\rho'X(X'D(\hat{m})X)^{-1}X'\rho &= \rho'W(W'D(\hat{m})W)^{-1}W'\rho \\ &= \lambda'(W'D(\hat{m})W)^{-1}\lambda\end{aligned}$$

which is easily computed if the regression program provides $(W'D(\hat{m})W)^{-1}$.

Variances for other estimated log odds ratios are computed in a similar manner. Because of the model, any log odds ratio with either j or k fixed, e.g., $\log(m_{1j1}m_{2j2}/m_{1j2}m_{2j1})$, is zero by assumption. Estimates of log odds in the j or k indices, e.g., $\log(m_{ij1}/m_{ij2})$, can also be estimated and large sample variances computed. However, because of the existence of the u_{23} interaction, the log odds will depend on the value of j. These issues are considered in more detail in the next example.

EXAMPLE 10.2.6. Consider again the data on classroom behavior used in Examples 3.0.1 and 3.2.2. The data and estimated expected cell counts for the model in which behavior is independent of risk and adversity are given below.

n_{ijk} ($\hat{m}_{ijk}$)		Adversity (k) Low		Medium		High	
Risk (j)		N	R	N	R	N	R
Classroom Behavior (i)	Non.	16 (14.02)	7 (6.60)	15 (14.85)	34 (34.64)	5 (4.95)	3 (4.95)
	Dev.	1 (2.98)	1 (1.40)	3 (3.15)	8 (7.36)	1 (1.05)	3 (1.05)

The ANOVA type model is

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}\,.$$

Except for the fact that this is a $2 \times 2 \times 3$ table instead of a $2 \times 3 \times 2$ table, the model is exactly as in the previous example.

Dropping all u terms with i, j, or k equal to 1 leads to

$$\begin{bmatrix} \hat{\mu}_{111} \\ \hat{\mu}_{121} \\ \hat{\mu}_{112} \\ \hat{\mu}_{122} \\ \hat{\mu}_{113} \\ \hat{\mu}_{123} \\ \hat{\mu}_{211} \\ \hat{\mu}_{221} \\ \hat{\mu}_{212} \\ \hat{\mu}_{222} \\ \hat{\mu}_{213} \\ \hat{\mu}_{223} \end{bmatrix} = \begin{bmatrix} \log(14.02) \\ \log(6.60) \\ \log(14.85) \\ \log(34.64) \\ \log(4.95) \\ \log(4.95) \\ \log(2.98) \\ \log(1.40) \\ \log(3.15) \\ \log(7.36) \\ \log(1.05) \\ \log(1.05) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{\gamma}_{1(2)} \\ \hat{\gamma}_{2(2)} \\ \hat{\gamma}_{3(2)} \\ \hat{\gamma}_{3(3)} \\ \hat{\gamma}_{23(22)} \\ \hat{\gamma}_{23(23)} \end{bmatrix}$$

or, alternatively,

$$\hat{\mu} = W\hat{\gamma}.$$

In particular, any weighted or unweighted regression analysis provides

$$\begin{aligned} \hat{\gamma} &= 2.640, \\ \hat{\gamma}_{1(2)} &= -1.548, \\ \hat{\gamma}_{2(2)} &= -0.754, \\ \hat{\gamma}_{3(2)} &= 0.057, \\ \hat{\gamma}_{3(3)} &= -1.041, \\ \hat{\gamma}_{23(22)} &= 1.601, \\ \hat{\gamma}_{23(23)} &= 0.754. \end{aligned}$$

(Actually, these are based on more significant digits for the $\hat{m}_{ijk}$'s than were reported above.) The matrix of asymptotic variances and covariances for the $\hat{\gamma}$'s is obtained from doing the appropriate auxiliary regression. It is

	$\hat{\gamma}$	$\hat{\gamma}_{1(2)}$	$\hat{\gamma}_{2(2)}$	$\hat{\gamma}_{3(2)}$	$\hat{\gamma}_{3(3)}$	$\hat{\gamma}_{23(22)}$	$\hat{\gamma}_{23(23)}$
$\hat{\gamma}$	.0610	−.0125	−.0588	−.0588	−.0588	−.0588	−.0588
$\hat{\gamma}_{1(2)}$	−.0125	.0713	.0000	.0000	.0000	.0000	.0000
$\hat{\gamma}_{2(2)}$	−.0588	.0000	.1838	.0588	.0588	−.1838	−.1838
$\hat{\gamma}_{3(2)}$	−.0588	.0000	.0588	.1144	.0588	−.1144	−.0588
$\hat{\gamma}_{3(3)}$	−.0588	.0000	.0588	.0588	.2255	−.0588	−.2255
$\hat{\gamma}_{23(22)}$	−.0588	.0000	−.1838	−.1144	−.0588	.2632	.1838
$\hat{\gamma}_{23(23)}$	−.0588	.0000	−.1838	−.0588	−.2255	.1838	.5172

This matrix is the basis for all subsequent variance estimates.

The estimate of the log odds of nondeviant behavior is

$$\begin{aligned}\log(\hat{m}_{1jk}/\hat{m}_{2jk}) &= \hat{\mu}_{1jk} - \hat{\mu}_{2jk} \\ &= \hat{u}_{1(1)} - \hat{u}_{2(1)} \\ &= -\hat{\gamma}_{1(2)} \\ &= 1.548\,.\end{aligned}$$

The equivalence between the u parametrization and the γ parametrization is easily obtained by inspection of $W\gamma$. The asymptotic standard error is $\sqrt{.0713}$, so a 90% confidence interval has end points

$$1.548 \pm 1.645\sqrt{.0713}\,.$$

The odds of having a home situation that is not at risk depend on the adversity level. The log odds satisfy

$$\begin{aligned}\log(\hat{m}_{i1k}/\hat{m}_{i2k}) &= \hat{u}_{2(1)} + \hat{u}_{23(1k)} - \hat{u}_{2(2)} - \hat{u}_{23(2k)} \\ &= \begin{cases} -\hat{\gamma}_{2(2)}\,, & k = 1 \\ -\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(22)}\,, & k = 2 \\ -\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(23)}\,, & k = 3. \end{cases}\end{aligned}$$

The estimated value when $k = 2$ is

$$.754 - 1.601 = -.847\,.$$

With

$$\mathrm{Var}(-\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(22)}) = \mathrm{Var}(\hat{\gamma}_{2(2)}) + 2\mathrm{Cov}(\hat{\gamma}_{2(2)}, \hat{\gamma}_{23(22)}) + \mathrm{Var}(\hat{\gamma}_{23(22)}),$$

the asymptotic estimated variance is

$$.1838 - 2(.1838) + .2632 = .0794\,.$$

The interesting odds ratios involve the changes in the odds as k changes.

$$\begin{aligned}\log(m_{i11}m_{i22}/m_{i21}m_{i12}) &= u_{23(11)} - u_{23(21)} - u_{23(12)} + u_{23(22)} \\ &= \gamma_{23(22)}\,, \\ \log(m_{i11}m_{i23}/m_{i21}m_{i13}) &= \gamma_{23(23)}\,, \\ \log(m_{i12}m_{i23}/m_{i22}m_{i13}) &= \gamma_{23(23)} - \gamma_{23(22)}\,.\end{aligned}$$

The last of these has an estimate of

$$.754 - 1.601 = -.847$$

and an estimated asymptotic variance of

$$.5172 - 2(.1838) + .2632 = .4128\,.$$

A 95% confidence interval for the log odds ratio is

$$(-.2.106, .412).$$

Transforming to the original scale gives an interval for the odds ratio of

$$(.12, 1.5).$$

The odds of being not at risk for medium-adversity schools is between .12 and 1.5 times those for high-adversity schools. The large interval is related to the very small numbers available at high risk. The result does not depend on the classroom behavior.

As a matter of fact, the need for including the u_{23} terms in the model is driven by the fact that

$$\hat{\gamma}_{23(22)} = 1.601$$

with an asymptotic standard error of

$$\sqrt{.2632} = .513\,.$$

Thus, there is clear evidence that the odds of being not at risk are higher for low-adversity schools than for high-adversity schools. In fact, the odds are roughly between 2 and 13 times larger with 95% confidence.

As we have seen, there is a problem with the output from standard regression software. Using a regression parametrization

$$\mu = W\gamma,$$

the key matrix to be obtained is

$$\hat{A}D^{-1}(\hat{m}) = W(W'D(\hat{m})W)^{-1}W'$$

which does not depend on the choice of W. Unfortunately, most regression software does not report $\hat{A}D^{-1}(\hat{m})$; it only reports

$$(W'D(\hat{m})W)^{-1}\,.$$

(The fact that there are good reasons for doing this makes it no less unfortunate for our purposes.) If the software allows computation of $\hat{A}D^{-1}(\hat{m})$, then the simple structure of the ρ vectors allows simple computation of the estimated variance $\rho'W(W'D(\hat{m})W)^{-1}W'\rho$. If the software does not allow direct computation of $\hat{A}D^{-1}(\hat{m})$, then it is necessary to compute the vector $\lambda' = \rho'W$. In other words, the simple function $\rho'\mu$ must be reparametrized into $\lambda'\gamma$.

Given λ' and $(W'D(\hat{m})W)^{-1}$, the variance $\lambda'(W'D(\hat{m})W)^{-1}\lambda$ is easily computed. The problem is in identifying λ, i.e., identifying the function of γ that is equivalent to $\rho'\mu$. Even though the interesting functions $\rho'\mu$ are simple, the functions of γ get progressively more complex as the model, (i.e.,

the matrix W) gets more complex. For example, the asymptotic variance of a log odds in terms of model parameters gets progressively more complicated as the model involves more higher-order interactions, even though the log odds is an extremely simple function of μ. With a good matrix manipulation package, keeping track of the parameters can be accomplished numerically.

Asymptotic Variances for Saturated Models

In Chapter 2, an asymptotic standard error was presented for estimated log odds ratios. The standard error is a consequence of applying Theorem 10.2.1a to a saturated model. Generally, standard errors for contrasts in the μ_i's are easily obtained for saturated models. Recall that for a saturated model, $\hat{\mu} = \log(n)$. Applying Theorem 10.2.1a, $\log(n) - \mu$ is approximately $N(0, [A - A_z]D^{-1}(m))$. We wish to characterize $[A - A_z]D^{-1}(m)$. The model is saturated, i.e.,

$$A = I(IDI)^{-1}ID = I,$$

so $AD^{-1}(m) = D^{-1}(m)$. For multinomial sampling (regardless of the log-linear model),

$$A_z D^{-1}(m) = \frac{1}{N} JJ' .$$

Thus, for a saturated model with a large multinomial sample, we have the approximation

$$\log(n) - \mu \sim N\left(0, D^{-1}(m) - \frac{1}{N} JJ'\right) .$$

Let $\rho = (\rho_1, \ldots, \rho_q)'$ be a vector with $\rho' J = 0$, i.e., $\rho_\cdot = 0$, so $\rho'\mu$ is a contrast in the μ_i's. The large sample distribution of $\rho' \log(n)$ is

$$\rho' \log(n) - \rho'\mu \sim N(0, \rho'[D^{-1}(m) - (1/n_\cdot)JJ']\rho) .$$

With $\rho' J = 0$, we have

$$\rho' \log(n) - \rho'\mu \sim N(0, \rho' D^{-1}(m)\rho)$$

or, equivalently,

$$\frac{\rho' \log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(m)\rho}} \sim N(0, 1) .$$

For this distribution to be useful in drawing inferences about $\rho'\mu$, an estimate of the unknown standard deviation $\sqrt{\rho' D^{-1}(m)\rho}$ must be incorporated. By Theorem 10.2.1d, the vector n/N converges to the vector p,

so $\rho' D^{-1}(m)\rho / \rho' D^{-1}(n)\rho = \rho' D^{-1}(p)p / \rho' D^{-1}(n/N)\rho$ converges to 1 and $\sqrt{\rho' D^{-1}(m)\rho} / \sqrt{\rho' D^{-1}(n)\rho}$ converges to 1. Hence, for large samples,

$$\frac{\rho' \log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(n)\rho}} = \frac{\rho' \log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(m)\rho}} \frac{\sqrt{\rho' D^{-1}(m)\rho}}{\sqrt{\rho' D^{-1}(n)\rho}} \sim N(0,1)\,.$$

This result can be very useful, especially for examining odds ratios.

EXAMPLE 10.2.7. For the $2\times2\times2$ table of Example 3.2.4 concerning auto injuries, we were interested in whether the odds ratios $p_{111}p_{221}/p_{121}p_{211}$ and $p_{112}p_{222}/p_{122}p_{212}$ were equal. Because

$$\frac{p_{11k}p_{22k}}{p_{12k}p_{21k}} = \frac{m_{11k}m_{22k}}{m_{12k}m_{21k}},$$

the log odds ratios are

$$\log\left(\frac{m_{11k}m_{22k}}{m_{12k}m_{21k}}\right) = \mu_{11k} - \mu_{12k} - \mu_{21k} + \mu_{22k}\,.$$

The odds ratios are equal if and only if the contrast in the μ_{ijk}'s

$$(\mu_{111} - \mu_{121} - \mu_{211} + \mu_{221}) - (\mu_{112} - \mu_{122} - \mu_{212} + \mu_{222})$$

equals zero. [Note that if $\mu = (\mu_{111}, \mu_{112}, \mu_{121}, \mu_{122}, \mu_{211}, \mu_{212}, \mu_{221}, \mu_{222})'$, then $\rho' = (1, -1, -1, 1, -1, 1, 1, -1)$.] The estimated odds ratios were

$$\begin{aligned} \hat{p}_{111}\hat{p}_{221}/\hat{p}_{121}\hat{p}_{211} &= 350(23)/26(150) \\ &= 2.064 \end{aligned}$$

and

$$\begin{aligned} \hat{p}_{112}\hat{p}_{222}/\hat{p}_{122}\hat{p}_{212} &= 60(80)/19(112) \\ &= 2.256\,. \end{aligned}$$

The estimate of the contrast is

$$\log(2.064) - \log(2.256) = -0.089\,.$$

The standard error for the estimate is

$$\begin{aligned} \sqrt{\rho' D^{-1}(n)\rho} &= \sqrt{\frac{1}{350} + \frac{1}{23} + \frac{1}{26} + \frac{1}{150} + \frac{1}{60} + \frac{1}{80} + \frac{1}{19} + \frac{1}{112}} \\ &= .4268\,. \end{aligned}$$

We can now test the hypothesis that the contrast is zero. The test statistic is $-.089/.4268 = -.21$. For an α level two-sided test, $|-.21|$ is compared to $z(1-\alpha/2)$. The hypothesis that the contrasts are equal is not rejected

for any reasonable size of α. A 95% confidence interval for the contrast has limits

$$-.089 \pm (1.96)(.4268)\,.$$

The test based on the asymptotic standard error is an alternative to the likelihood ratio and Pearson chi-squared tests for no three-factor interaction.

To examine an individual cell, the term $A_z D^{-1}(m)$ must be accounted for in the covariance matrix. It is easily seen that for large samples, the appropriate distribution for $\hat{p}_{ijk}$ is

$$\frac{\hat{p}_{ijk} - p_{ijk}}{\sqrt{\hat{p}_{ijk}(1-\hat{p}_{ijk})/N}} \sim N(0,1)\,.$$

Similar results hold for $\hat{m}_{ijk}$ and $\hat{\mu}_{ijk}$.

Testing Models

Consider the problem of testing a model $\mu = X_0\beta_0$ against a larger model $\mu = X\beta$. In particular, assume that $\mu = X\beta$ is valid and examine the test of

$$H_0 : \mu = X_0\beta_0 \quad \text{for some} \quad \beta_0$$

versus

$$H_A : \mu \neq X_0\beta_0 \quad \text{for any} \quad \beta_0\,,$$

where $C(X_0) \subset C(X)$, i.e., $X_0 = XB$ for some matrix B. Let $\hat{m}$ be the MLE of m under the assumption that $\mu = X\beta$ and let $\hat{m}_0$ be the MLE of m under the assumption that $\mu = X_0\beta_0$. The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i})\,.$$

The Pearson test statistic is

$$X^2 = \sum_{i=1}^{q} (\hat{m}_i - \hat{m}_{0i})^2/\hat{m}_{0i}\,.$$

The main asymptotic results for testing hypotheses are given in the following theorem.

Theorem 10.2.8. Let $r = \text{rank}(X)$ and $r_0 = \text{rank}(X_0)$.

(a) If H_0 is true and $N \equiv n.$ is large, the following distributions are approximately valid:

$$G^2 \sim \chi^2(r - r_0)$$

and

$$X^2 \sim \chi^2(r - r_0)\,.$$

Moreover,

$$G^2 - X^2 \xrightarrow{P} 0\,.$$

(b) If H_0 is not true, then both G^2 and X^2 get arbitrarily large as the sample size increases.

It is interesting to note that the degrees of freedom for the test are $\text{rank}(X) - \text{rank}(X_0)$. This is the reason that degrees of freedom are computed exactly as in analysis of variance. In both cases, it is simply the linear structure of the model that determines the degrees of freedom.

10.3 Product-Multinomial Sampling

With a few minor changes, all of the results of Sections 1 and 2 hold for product-multinomial sampling. Suppose that we have t multinomial populations instead of just one. We can write the observations as n_{ij}, $i = 1, \ldots, t$, $j = 1, \ldots, s_i$, where s_i is the number of categories in the i'th multinomial. (Note that $q = \sum_{i=1}^{t} s_i$.) The probabilities and expected cell counts can be written similarly as p_{ij} and m_{ij}, respectively.

In place of the condition from multinomial sampling that all the probabilities in the table add to 1, cf. equation (10.1.2), we now have

$$p_{i\cdot} = 1, \qquad i = 1, \ldots, t,$$

and because $m_{ij} = n_{i\cdot} p_{ij}$, we have

$$m_{i\cdot} = n_{i\cdot}\,, \qquad i = 1, \ldots, t\,.$$

Write the vectors $n = (n_{11}, n_{12}, \ldots, n_{ts_t})'$ and $m = (m_{11}, m_{12}, \ldots, m_{ts_t})'$. Let Z be a $q \times t$ matrix of indicator variables for the t samples. Specifically, each column of Z corresponds to a different multinomial. A particular column of Z, say the i'th column, has ones in the rows corresponding to $n_{i1}, \ldots, n_{is_i}$ and zeros in all other rows. (Note that if $n_{ij} = \mu_i + e_{ij}$ was a one-way ANOVA, Z would be the design matrix for the linear model.) With this definition of Z, the condition $m_{i\cdot} = n_{i\cdot}$, $i = 1, \ldots, t$, becomes

$$n'Z = m'Z\,.$$

Suppose now that we have the log-linear model $\log(m) = \mu = X\beta$. It can be shown that maximizing the log-likelihood under product multinomial sampling is equivalent to maximizing $\ell(m) = n' \log(m)$, cf. Chapter 12.

The MLE of m must maximize $\ell(m)$ subject to the conditions

$$\log(m) = X\beta \tag{1}$$

and

$$n'Z = m'Z\,. \tag{2}$$

Just as in Section 1, if $\hat{m}$ maximizes $\ell(m)$ subject only to condition (1), then $\hat{m}$ must satisfy

$$n'X = \hat{m}'X\,. \tag{3}$$

In order to get condition (2) satisfied, we restrict our attention to models $\log(m) = X\beta$ in which $C(Z) \subset C(X)$. For such models, $Z = XB$ for some matrix B; hence, (3) implies that

$$n'Z = n'XB = \hat{m}'XB = \hat{m}'Z\,.$$

The assumption that $C(Z) \subset C(X)$ is not difficult to deal with. For an $I \times J$ table in which rows are independent multinomial samples, the condition $C(Z) \subset C(X)$ is the requirement that every log-linear model include (the equivalent of) $u_{1(i)}$ terms for rows. In an $I \times J \times K$ table in which there is an independent multinomial sample for each combination of row and layer, the condition $C(Z) \subset C(X)$ is the requirement that every log-linear model include $u_{13(ik)}$ terms or their equivalent. Note that, for example, the models

$$\log(m_{ijk}) = u_{13(ik)} + u_{123(ijk)}$$

and

$$\log(m_{ijk}) = u_{123(ijk)}$$

are equivalent models, so in spite of the fact that $\log(m_{ijk}) = u_{123(ijk)}$ does not contain $u_{13(ik)}$ terms, it does contain the equivalent of $u_{13(ik)}$ terms.

Under product-multinomial sampling, the asymptotic results of Section 2 change very little. The matrix $D(p)$ is no longer of interest. Instead, define $m^* = (m^*_{11}, \ldots, m^*_{ts_t})$ where $m^*_{ij} = n_{i\cdot}p_{ij}/n_{\cdot\cdot}$. Redefine

$$D = D(m^*)\,.$$

The matrix A is defined as before except that the new version of D is used. Also, redefine A_z as

$$A_z = Z(Z'DZ)^{-1}Z'D\,.$$

For asymptotic results, *let $N = n_{\cdot\cdot}$ get large and let $n_{i\cdot}/n_{\cdot\cdot}$ remain fixed for each i*. Write $N_i = n_{i\cdot} = m_{i\cdot}$.

Before restating the asymptotic results, note that multinomial sampling is just a special case of product-multinomial sampling. In particular, it has $t = 1$, $Z = J$ (J is a $q \times 1$ vector of 1s), $N = n_{\cdot} = n_{1\cdot} = n_{\cdot\cdot}$, and $m^* = p$.

Theorem 10.3.1. For multinomial or product-multinomial sampling, the following distributions are approximately valid when $N_1, \ldots, N_t$ are large:

(a) $\hat{\mu} - \mu \sim N(0, (A - A_z)D^{-1}(m))$,

(b) $\hat{m} - m \sim N(0, D(m)(A - A_z))$.

In addition,

(c) $\hat{\mu} - \mu \xrightarrow{P} 0$,

(d) $N^{-1}\hat{m} \xrightarrow{P} m^*$.

Estimation for product-multinomial sampling is similar to that for multinomial sampling. Theorem 10.3.1 looks identical to Theorem 10.2.1. The difference is that A_z stands for something different. Again, the only problem is in computing asymptotic variances. Under product-multinomial sampling, the variance of $\rho'\hat{\mu}$ is $\rho'[A - A_z]D^{-1}(m)\rho$. Note that

$$Z'D(m)Z = D(N_1, \cdots, N_t),$$

$$(Z'D(m)Z)^{-1} = D\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right),$$

and

$$A_z D^{-1}(m) = ZD\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right) Z'.$$

The variance of $\rho'\hat{\mu}$ is

$$\rho' AD^{-1}(m)\rho - \rho' ZD\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right) Z'\rho.$$

The second term can be computed exactly. The first term must be estimated and, even then, requires a computer to evaluate. For example, taking $\rho' = e'_{ij} = (0, \cdots, 0, 1, 0, \cdots, 0)$ with the 1 in the column corresponding to the ij cell, Theorem 10.3.1 yields

$$\hat{\mu}_{ij} - \mu_{ij} = e'_{ij}(\hat{\mu} - \mu) \sim N\left(0, \frac{a_{ij,ij}}{m_{ij}} - \frac{1}{N_i}\right)$$

where $a_{ij,ij}$ is the diagonal element of A corresponding to the ij cell.

Similarly,

$$\mathrm{Var}(\hat{m}_{ij} - m_{ij}) = m_{ij}a_{ij,ij} - m_{ij}^2/N_i$$

and with $p_{ij} = m_{ij}/N_i$,

$$\begin{aligned} \mathrm{Var}(\hat{p}_{ij} - p_{ij}) &= p_{ij}a_{ij,ij}/N_i - p_{ij}^2/N_i \\ &= p_{ij}(a_{ij,ij} - p_{ij})/N_i. \end{aligned}$$

If $\rho'\mu$ is a log odds or a log odds ratio that happens to be computed entirely within a particular multinomial, then $\rho' Z = 0$ and

$$\rho'[A - A_z]D^{-1}(m)\rho = \rho' AD^{-1}(m)\rho.$$

This is computed exactly as in Section 2. Unfortunately, it again requires a computer to evaluate.

For testing hypotheses, the large sample results appropriate for product-multinomial sampling are given in the following theorem.

Theorem 10.3.2. Assume $\mu = X\beta$ and let X_0 be a matrix with $C(Z) \subset C(X_0) \subset C(X)$. Let $\text{rank}(X) = r$ and $\text{rank}(X_0) = r_0$. For testing $H_0 : \mu = X_0\beta_0$ for some β_0 versus $H_A : \mu \neq X_0\beta_0$ for any β_0, under multinomial or product-multinomial sampling, if N is large, then the following approximate distributions hold:

(a) if H_0 is true, $G^2 \sim \chi^2(r - r_0)$,

(b) if H_0 is true, $X^2 \sim \chi^2(r - r_0)$,

also,

(c) if H_0 is true, $G^2 - X^2 \xrightarrow{P} 0$,

(d) if H_0 is false, G^2 and X^2 tend to infinity as N gets large.

Note that by (c), if H_0 is true, the difference between G^2 and X^2 can be used as an indication of how good the large sample approximation is. If H_0 is not true, then G^2 and X^2 need not be equivalent in large samples.

10.4 Inference for Model Parameters

Thus far, we have been primarily concerned with estimation of m and μ. It may be of interest to estimate the parameter vector β in the log-linear model $\mu = X\beta$. Estimates of β are obtained as in analysis of variance and regression, except that instead of performing operations on the data (y values), the operations are performed on $\hat{\mu}$.

Suppose that $\text{rank}(X) = p$ so that $\mu = X\beta$ is a regression model. $\hat{\beta}$ satisfies

$$\hat{\mu} = X\hat{\beta},$$

so

$$(X'X)^{-1}X'\hat{\mu} = (X'X)^{-1}X'X\hat{\beta} = \hat{\beta}\,.$$

The MLE of $\hat{\beta}$ is obtained by performing a regression on $\hat{\mu}$. (In fact, any weighted regression will give the same $\hat{\beta}$.)

Essentially the same argument holds for ANOVA type models. If one imposes side conditions on the parameters (something the author is loathe to do), then estimates of the parameters in ANOVA models are available. For example, in the model $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ if the side

conditions $u_{1(\cdot)} = u_{2(\cdot)} = u_{12(i\cdot)} = u_{12(\cdot j)} = 0$ are imposed and if we denote $w_{ij} = \hat{\mu}_{ij}$, then

$$\begin{aligned} \hat{u} &= \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{1(i)} &= \bar{w}_{i\cdot} - \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{2(j)} &= \bar{w}_{\cdot j} - \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{12(ij)} &= w_{ij} - \bar{w}_{i\cdot} + \bar{w}_{\cdot j} + \bar{w}_{\cdot\cdot}\,. \end{aligned}$$

Again, these are precisely the estimates obtained by doing an ANOVA on $\hat{\mu}$.

Tests and confidence intervals for functions $\rho' X\beta$ can be obtained from the asymptotic distribution

$$\frac{\rho'\hat{\mu} - \rho' X\beta}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} \sim N(0,1)\,.$$

For example, an asymptotic 95% confidence interval for $\rho' X\beta$ has limits $\rho'\hat{\mu} \pm 1.96\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}$ and an $\alpha = .05$ test of $H_0 : \rho' X\beta = 0$ versus $H_A : \rho' X\beta \neq 0$ rejects if

$$\frac{\rho'\hat{\mu}}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} > 1.96$$

or if

$$\frac{\rho'\hat{\mu}}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} < -1.96\,.$$

EXAMPLE 10.4.1. In this and the previous three sections, a lot of machinery has been developed for analyzing log-linear models. In this example, we apply the matrix approach to the analysis of model (6.2.2) in Example 6.2.6. Our analysis also employs the data from Example 6.2.1 as summarized in Example 6.2.5.

In matrix form, model (6.2.2) can be written as

$$\begin{bmatrix} \log(m_{1111}) \\ \log(m_{1112}) \\ \log(m_{1121}) \\ \log(m_{1122}) \\ \log(m_{1211}) \\ \log(m_{1212}) \\ \log(m_{1221}) \\ \log(m_{1222}) \\ \log(m_{2111}) \\ \log(m_{2112}) \\ \log(m_{2121}) \\ \log(m_{2122}) \\ \log(m_{2221}) \\ \log(m_{2222}) \end{bmatrix} = X \begin{bmatrix} \lambda \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \\ \lambda_{23} \\ \lambda_{24} \\ \lambda_{34} \\ \lambda_{123} \\ \lambda_{124} \\ \lambda_{134} \\ \lambda_{234} \\ \lambda_{1234} \end{bmatrix}$$

where

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

The columns of the design matrix X can be identified as $X_0, \ldots, X_{1234}$ with the subscript of X identical to the subscript of the corresponding λ term. (X_0 corresponds to λ.) Note that, say, X_{12} can be obtained by multiplying together the elements of X_1 and X_2. Similarly, the elements of X_{134} can be obtained by multiplying together the elements of X_1, X_3, and X_4. In fact, any column with more than one subscript can be obtained by multiplying together the appropriate columns with one subscript.

Another important fact is that any two columns, for example X_{12} and X_{134}, have the property that $X_{12}'X_{134} = 0$. Any column, say X_{12}, has $X_{12}'X_{12} = 16 = q$, so we have $\frac{1}{16}X_{12}'X\beta = \lambda_{12}$. The estimate of λ_{12} is $\frac{1}{16}X_{12}'X\hat{\beta} = (1/16)X_{12}'\hat{\mu} = (1/16)(\hat{\mu}_{11\cdot\cdot} - \hat{\mu}_{12\cdot\cdot} - \hat{\mu}_{21\cdot\cdot} + \hat{\mu}_{22\cdot\cdot}) = 4(\bar{w}_{11\cdot\cdot} - \bar{w}_{12\cdot\cdot} - \bar{w}_{21\cdot\cdot} + \bar{w}_{22\cdot\cdot})$ where $w_{hijk} = \log(n_{hijk})$ because the model is saturated.

The variance of $\frac{1}{16}X_{12}'\hat{\mu}$ is $\left(\frac{1}{16}\right)^2 X_{12}'[D^{-1}(m) - A_z D^{-1}(m)]X_{12}$ for large samples. If the parameter λ_{12} is not forced into the model to deal with product-multinomial sampling, then $X_{12}'A_z D^{-1}(m)X_{12} = 0$, so the asymptotic variance is

$$\left(\frac{1}{q}\right)^2 X_{12}'D^{-1}(m)X_{12} = \left(\frac{1}{q}\right)^2 \sum_{hijk}\left(\frac{1}{m_{hijk}}\right)$$

where $q = 16$. The estimated asymptotic variance of $\hat{\lambda}_{12}$ is

$$\widehat{\mathrm{Var}}(\hat{\lambda}_{12}) = \left(\frac{1}{q}\right)^2 \sum_{hijk}\left(\frac{1}{n_{hijk}}\right).$$

In fact, the same asymptotic variance applies to all of the λ terms that are not forced into the model.

Using the asymptotic distribution

$$\frac{\hat{\lambda}_{12} - \lambda_{12}}{\sqrt{\left(\frac{1}{q^2}\right)\sum_{hijk}\left(\frac{1}{n_{hijk}}\right)}} \sim N(0,1),$$

a test of $H_0 : \hat{\lambda}_{12} = 0$ is based on comparing the test statistic

$$\frac{\hat{\lambda}_{12} - 0}{\sqrt{\left(\frac{1}{q^2}\right) \sum_{hijk} \left(\frac{1}{n_{hijk}}\right)}}$$

to a $N(0,1)$ distribution. Using the numbers in Example 6.2.5, we see that

$$|\hat{\lambda}_{TW}| = 0.914/16 = .0571\,,$$

the standard error is

$$1.307/16 = .0817\,,$$

and the test statistic is

$$\frac{.0571 - 0}{.0817} = 0.70\,,$$

just as reported in Example 6.2.5. There is very little evidence that $\lambda_{TW} \neq 0$.

Similarly, an asymptotic 95% confidence interval for λ_{TW} has end points

$$.0571 \pm 1.96(.0817)\,.$$

10.5 Methods for Finding Maximum Likelihood Estimates

In general, some sort of iterative technique is necessary to find MLEs for log-linear models. The two commonly used methods are *iteratively reweighted least squares* and iterative proportional fitting. Iterative proportional fitting was discussed in Section 3.3. It works only for ANOVA type models. Fitting of general log-linear models is usually performed using iteratively reweighted least squares.

Iteratively Reweighted Least Squares

Maximum likelihood estimates for log-linear models can be found by performing a sequence of weighted linear regressions. This method is an application of the *Newton-Raphson algorithm.*

Given a vector function $f(\beta)$, Newton-Raphson is a method for finding a solution to $f(\beta) = 0$. It begins with an initial guess of β, say β_0. Newton-Raphson then defines a sequence of β's, say $\beta_1, \beta_2, \ldots$, that converge to a value $\hat{\beta}$ that satisfies $f(\hat{\beta}) = 0$. The sequence is defined recursively; we begin with an initial value β_0 and define β_{t+1} given the value of β_t. Specifically, let $df(\beta)$ be the matrix of partial derivatives of the vector-valued function $f(\beta)$. By Taylor's theorem, if β_t and β_{t+1} are close to each other and $\delta_t = \beta_{t+1} - \beta_t$, then the approximate equality

$$f(\beta_{t+1}) \doteq f(\beta_t) + [df(\beta_t)]\delta_t$$

holds. We are seeking a zero of $f(\beta)$ so Newton-Raphson sets

$$0 = f(\beta_t) + [df(\beta_t)]\delta_t$$

so that

$$\delta_t = -[df(\beta_t)]^{-1} f(\beta_t) .$$

With $\delta_t = \beta_{t+1} - \beta_t$, we have

$$\beta_{t+1} = \beta_t + \delta_t .$$

Consider a log-linear model $\mu = X\beta$ where X is a $q \times p$ matrix with $\text{rank}(X) = p$. Note that any log-linear model can be reparametrized so that $\text{rank}(X) = p$. The MLE of m will be the same regardless of the parametrization. We wish to find the maximum of the function $\ell(m)$. In particular, this can be done by setting appropriate partial derivatives of $\ell(m)$ equal to zero. The Newton-Raphson method can be used to find the zero of the partial derivative vector.

Before applying the Newton-Raphson method, we set some notation. If $x = (x_1, \ldots, x_q)'$, write $e^x = (e^{x_1}, \ldots, e^{x_q})'$. With $\log(m) = X\beta$, m is a function of β. Write $\log(m(\beta)) = X\beta$ and $m(\beta) = e^{X\beta}$. In applying Newton-Raphson, we find $\hat{\beta}$ with $f(\hat{\beta}) = 0$ where

$$f(\beta) = d\ell(e^{X\beta})$$

and $d\ell(e^{X\beta})$ is the matrix of partial derivatives of $\ell(e^{X\beta})$ with respect to the vector β. It follows that $\hat{m} = e^{X\hat{\beta}}$ will maximize $\ell(m)$ subject to the constraint that $\log(\hat{m}) = X\hat{\beta}$ for some $\hat{\beta}$.

It is shown in Chapter 12 that $f(\beta_t) = X'(n - m(\beta_t))$ and that $df(\beta_t) = -X'D(m(\beta_t))X$; thus,

$$\delta_t = [X'D(m(\beta_t))X]^{-1}X'(n - m(\beta_t))$$

and

$$\beta_{t+1} = \beta_t + [X'D(m(\beta_t))X]^{-1}X'(n - m(\beta_t)) .$$

The value β_{t+1} can be obtained from β_t simply by doing a weighted regression analysis. Let

$$Y \equiv X\beta_t + [D(m(\beta_t))]^{-1}(n - m(\beta_t)) . \tag{1}$$

If we fit the regression model

$$Y = X\beta + e\,,\ E(e) = 0\,,\ \text{Cov}(e) = [D(m(\beta_t))]^{-1}\,,$$

the estimate of β is

$$\begin{aligned} \beta_{t+1} &= [X'D(m(\beta_t))X]^{-1}X'D(m(\beta_t))Y \\ &= \beta_t + \delta_t . \end{aligned}$$

The matrix $D(m(\beta_t))$ is diagonal, so this is the simplest form of weighted regression and can be performed on most standard regression programs. The weights are simply the individual values of the vector $m(\beta_t)$.

This method of finding MLEs, because it consists of a series of weighted regressions in which the weights continually change, is called *iteratively reweighted least squares*. The method does not depend on any particular choice of X except for the condition that $\text{rank}(X) = p$. Any log-linear model can be reparametrized so that X has full column rank, i.e., $\text{rank}(X) = p$, so the method is perfectly general.

10.6 Regression Analysis of Categorical Data

In this section, we present an alternative to maximum likelihood, namely the *weighted least squares* method of fitting log-linear models. This method was introduced by Grizzle, Starmer, and Koch (1969). It consists of fitting a linear model (regression model) to the logs of the counts while also using the counts as weights. We begin by explaining and illustrating the method. Mathematical justifications are given at the end of the section.

Recall that for a saturated model, $\hat{m} = n$ is the MLE. For large samples, Theorem 10.3.1 applies and, because $A = I$, we have the approximation

$$\log(n) \sim N(\mu, D^{-1}(m) - A_z D^{-1}(m)) .$$

If we assume a log-linear model

$$\mu = X\beta ,$$

then

$$\log(n) \sim N(X\beta, D^{-1}(m) - A_z D^{-1}(m)) ,$$

which can be rewritten as

$$\log(n) = X\beta + e, \qquad e \sim N(0, D^{-1}(m) - A_z D^{-1}(m)) . \tag{1}$$

This is just a linear model, but it has an unusual covariance matrix for the errors. Most commonly in regression analysis, it is assumed that $\text{Cov}(e) = \sigma^2 I$. Courses on applied regression analysis often deal with weighted least squares, where $\text{Cov}(e) = \sigma^2 D(w)$ and w is some $q \times 1$ vector of known constants. This covariance structure can be handled very easily. In particular, most computer programs for doing regression analysis can handle this form of weighted regression. Unfortunately, the covariance matrix for model (1) is more complicated.

As will be discussed later in the subsection on Mathematical Justifications, estimates in model (1) are precisely the same as estimates in

$$\log(n) = X\beta + e , \qquad e \sim N(0, D^{-1}(m)) . \tag{2}$$

This is much closer to the standard form of $\sigma^2 D(w)$. There are two differences. One is that in model (2) there is no variance σ^2 to be estimated; we know that $\sigma^2 = 1$. The second difference is that w is supposed to be known but m is not known. This problem is evaded by estimating m from the saturated model. Thus, the regression method is to fit the model

$$\log(n) = X\beta + e\,, \qquad e \sim N(0, D^{-1}(n))\,. \tag{3}$$

This procedure has essentially the same asymptotic properties as maximum likelihood estimation.

Although we are using model (3) as a device for fitting the log-linear model, our real model is model (1). Model (3) gives a valid estimate for β, but it cannot be used for the entire analysis. Fortunately, when considering the most interesting parameters in β, model (3) can be used to construct asymptotically valid tests and confidence intervals. In particular, this works for parameters that are not forced into the model to account for the sampling scheme. Remember, we assume that $C(Z) \subset C(X)$ where Z is the matrix of indicators for the product-multinomial samples, cf. Section 3. Any log-linear model can be reparametrized so that $X = [Z, X_1]$ and $\beta' = [\alpha', \beta_1']$. The parameter vector α consists of parameters that are forced into the model to account for the sampling scheme. For drawing inferences about β_1, model (3) gives valid tests and confidence intervals.

Because $\sigma^2 = 1$, when drawing inferences about model (3) one uses tests based on the normal distribution and the chi-square distribution rather than the t distribution and the F distribution. When performing chi-square tests, the test statistic is the numerator sum of squares from the usual F statistic with the appropriate number of degrees of freedom. Again, inferences must be restricted to parameters that are not forced into the model.

EXAMPLE 10.6.1. *Drug Comparisons.*
The hypothetical data presented below has been analyzed in Koch, Imrey, Freeman, and Tolley (1976). They also mention other references. Three drugs A, B, and C were given to each of 46 subjects. The response of each subject to each drug was noted as favorable (F) or unfavorable (U). Assume a multinomial sampling scheme. The data are

	Drug B	F		U	
	Drug C	F	U	F	U
Drug A	F	6	16	2	4
	U	2	4	6	6

First, consider fitting the log-linear model [AB][C] by fitting the corre-

sponding linear model

$$\begin{bmatrix} \log(6) \\ \log(16) \\ \log(2) \\ \log(4) \\ \log(2) \\ \log(4) \\ \log(6) \\ \log(6) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta \\ (\alpha\beta) \\ \gamma \end{bmatrix} + e \qquad (4)$$

using the weights (6,16,2,4,2,4,6,6). The parameters α, β, $\alpha\beta$, and γ can be described as main effects for drugs A and B, the $A \times B$ interaction, and the drug C main effect. A regression program gave the following results for fitting this model:

Regression Output

	Coefficient	Std Error	t
μ	1.5662	.1335	11.73
α	.1436	.1300	1.11
β	.1436	.1300	1.11
$\alpha\beta$	.5128	.1294	3.96
γ	−.3055	.1201	−2.54

Sum of squared errors (SSE) = 1.7348
Degrees of freedom error (dfE) = 3
Mean squared error (MSE) = .5783

As discussed above, the regression program acts as if there is a scale parameter σ that needs to be estimated. For log-linear models, the scale parameter is one, so the regression output must be modified to remove the adjustments for scale. This consists of dividing the regression standard errors and multiplying the t values by $(\text{MSE})^{1/2}$. Doing this gives

GSK Estimates

Coefficient	Estimate	Std Error	z
μ	1.5662	—	—
α	.1436	.1710	.844
β	.1436	.1710	.844
$\alpha\beta$	.5128	.1702	3.011
γ	−.3055	.1579	−1.932

The z values can be compared to the standard normal distribution for an asymptotic test of whether the coefficients are zero. The fact that no

standard error is reported for μ is due to the fact that μ is forced into the model by the multinomial sampling.

SSE is not used in the standard errors of coefficients, but it is used for testing different models. For example, fitting the model [A][B], i.e.,

$$\begin{bmatrix} \log(6) \\ \log(16) \\ \log(2) \\ \log(4) \\ \log(2) \\ \log(4) \\ \log(6) \\ \log(6) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta \end{bmatrix} + e \tag{5}$$

with the counts as weights, gives

$$\text{SSE[model (5)]} = 14.017$$

with 5 degrees of freedom. To test model (5) against model (4) [i.e., to test $H_0 : \alpha\beta = \gamma = 0$], compare the difference in the error sums of squares, $14.0173 - 1.7348 = 12.2825$, to a chi-square distribution with $5 - 3 = 2$ degrees of freedom. Of course, to test the significance of one parameter, either the chi square or the normal test can be used. The tests are identical.

Saturated models present some different features. Saturated models must fit perfectly; so for such a model, the SSE is zero. The fact that the saturated model has SSE $= 0$ causes a problem in finding standard errors and z values for regression coefficients. The regression output will try to use a scale parameter of zero, so the regression standard errors will all be reported as zero and the t values will be reported as infinite. It also follows that for any model other than the saturated model, the SSE reported in the regression output provides a direct test of lack of fit, i.e., a test of the model against the saturated model, when compared to a chi-square distribution with dfE degrees of freedom. For example, in the model [AB][C], comparing 1.7348 to a $\chi^2(2)$ provides a test for lack of fit.

Finally, many regression programs give additional output on the sums of squares for the different coefficients such as Sum of Squares explained by each variable in the order they are entered into the model. For model (4), this is

Due to	*df*	SS
Regression	4	18.3901
α	1	4.4353
β	1	1.6723
$\alpha\beta$	1	8.5381
γ	1	3.7444

The test of $H_0 : \gamma = 0$ can be performed by comparing 3.7444 to a chi-squared distribution with 1 degree of freedom. The test of $H_0 : \gamma = \alpha\beta = 0$ can be performed by comparing $8.5381 + 3.7444 = 12.2825$ to a chi square with 2 degrees of freedom. Both of these tests are equivalent to previously discussed versions of the tests.

Three things should be noted about the estimation technique of Grizzle, Starmer, and Koch (GSK). First, the method consists of performing one step of the Newton-Raphson algorithm. If the initial guess in the Newton-Raphson equation (10.5.1) is taken as $X\beta_0 = \log(n)$, then one iteration gives the GSK estimate. Second, the GSK method of estimation depends on an asymptotic result. It is only asymptotically that model (1) is valid. Maximum likelihood, on the other hand, is a valid method of estimation for any sample size. Similarly, likelihood ratio statistics are reasonable statistics on which to base tests for any sample size. With maximum likelihood, all procedures will be based on sufficient statistics. Only the distributions depend on large samples. Finally, the GSK method has trouble with observations that are zero. Taking the log of zero is usually a problem.

Koch et al. (1976) propose a compromise between maximum likelihood and weighted least squares. Suppose we wish to fit some model that cannot be conveniently fitted by iterative proportional fitting, say

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + \gamma_i k\,. \tag{6}$$

If software is available for iterative proportional fitting but not for iteratively reweighted least squares, perform the maximum likelihood fit of a slightly larger ANOVA type model, say

$$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)}\,.$$

The estimated vector $\hat{m}$ obtained from this can be used in place of n in the GSK procedure. The analysis follows the standard GSK methods. The compromise essentially provides GSK with better starting values.

Mathematical Justifications

There are two things that require justification. First, that estimates are the same in models (1) and (2) and, second, that estimates of functions of β_1 have the same variance in models (1) and (2).

Why do models (1) and (2) have the same estimates? Note that

$$\begin{aligned} A_z D^{-1}(m) &= Z(Z'D(m)Z)^{-1}Z'D(m)D^{-1}(m) \\ &= Z(Z'D(m)Z)^{-1}Z'\,. \end{aligned}$$

Thus, the covariance matrix in model (1) is

$$D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'\,.$$

The covariance matrix in model (2) can be written

$$D^{-1}(m) = [D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'] + Z(Z'D(m)Z)^{-1}Z' .$$

Because $C(Z) \subset C(X)$, this is precisely the condition needed to apply Theorem 10.1.3 in Christensen (1996b). The theorem implies that best linear unbiased estimates in models (1) and (2) are identical.

The idea behind Christensen's theorem is that because $C(Z) \subset C(X)$, for some (non-negative definite) matrix B, the covariance matrix of (1) can be written

$$D^{-1}(m) - XBX' .$$

Model (2) is equivalent to model (1) but with an additional independent error term added in. In particular, model (2) is equivalent to

$$\begin{aligned} \log(n) &= X\beta + (e + e_0) , \\ e &\sim N(0, D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z') , \\ e_0 &\sim N(0, Z(Z'D(m)Z)^{-1}Z') , \end{aligned} \tag{7}$$

where e and e_0 are independent. The covariance matrix for the entire error is

$$\begin{aligned} \mathrm{Cov}(e + e_0) &= D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z' + Z(Z'D(m)Z)^{-1}Z' \\ &= D^{-1}(m) . \end{aligned}$$

The trick involves the covariance matrix associated with e_0. With probability one, $e_0 \in C(Z(Z'D(m)Z)^{-1}Z') \subset C(Z) \subset C(X)$, cf. Christensen (1996b, Lemma 1.3.5). Because $e_0 \in C(X)$, we are adding error that we cannot distinguish from the mean $X\beta$. The only thing an unbiased estimate can do to such error is ignore it. Thus, the estimates with the additional error e_0 and the estimates without the additional error are identical.

We now examine the fact that estimates of estimable functions of β_1 have the same variance in models (1) and (2).

Estimates are the same in models (1) and (2), so using model (2) and standard linear model results, we have

$$\hat{\mu} = X\hat{\beta} = A\log(n)$$

where $A = X(X'D(m)X)^{-1}X'D(m)$. A is a projection operator onto $C(X)$; this means that $AX = X$, so, in particular, $AA = A$ and, because $C(Z) \subset C(X)$, $AZ = Z$ and $AA_z = A_z$.

Again, write $\mu = X\beta = Z\alpha + X_1\beta_1$. Consider an estimable function of β_1, say $\lambda'\beta_1$. For this to be estimable, by definition we must have $\lambda'\beta_1 = \rho'\mu = \rho'Z\alpha + \rho'X_1\beta_1$ for some $q \times 1$ vector ρ. In particular, we must have $\rho'Z = 0$ and $\rho'X_1 = \lambda'$. Now consider the asymptotic variance of $\lambda'\hat{\beta}_1$

under model (1),

$$\begin{aligned}\mathrm{Var}(\lambda'\hat{\beta}_1) &= \mathrm{Var}(\rho'\hat{\mu}) \\ &= \mathrm{Var}(\rho' A \log(n)) \\ &= \rho' A[D^{-1}(m) - A_z D^{-1}(m)]A'\rho \\ &= \rho' A D^{-1}(m) A'\rho - \rho' A A_z D^{-1}(m) A'\rho \\ &= \rho' A D^{-1}(m) A'\rho\,.\end{aligned}$$

The last equality follows from the fact that

$$\rho' A A_z = \rho' A_z = \rho' Z (Z' D(m) Z)^{-1} Z' D(m) = 0$$

because $\rho' Z = 0$.

The variance of $\lambda'\hat{\beta}_1$ under model (2) is

$$\begin{aligned}\mathrm{Var}(\lambda'\hat{\beta}_1) &= \mathrm{Var}(\rho' A \log(n)) \\ &= \rho' A[D^{-1}(m)]A'\rho,\end{aligned}$$

so the variances are the same under models (1) and (2). Thus, for estimable functions of β_1, the standard error reported from fitting model (2) is identical to the true standard error which is computed using model (1). For estimating functions that involve α, model (2) cannot be used to obtain standard errors.

In practice, neither model (1) nor (2) can be used because the covariance matrices involve the unknown parameter vector m. Model (3) substitutes the estimate n for m in the covariance matrix of (2). The same substitution in model (1) gives the most proper usable form for the GSK analysis. As above, the two models give the same estimate and the same standard errors for estimates of β_1. For drawing inferences about $\lambda'\beta_1$, use the approximate distribution

$$\frac{\lambda'\hat{\beta}_1 - \lambda'\beta_1}{\sqrt{\rho' D^{-1}(n)\rho}} \sim N(0,1)\,.$$

In particular, a large sample 95% confidence interval for $\lambda'\beta_1$ has end points

$$\lambda'\hat{\beta}_1 \pm 1.96\sqrt{\rho' D^{-1}(n)\rho}\,.$$

An α level test of $H_0 : \lambda'\hat{\beta}_1 = 0$ rejects H_0 when

$$\frac{|\lambda'\hat{\beta}_1|}{\sqrt{\rho' D^{-1}(n)\rho}} > z(1-\alpha/2)\,.$$

In summary, model (3) can be fitted very simply because it has a diagonal covariance matrix. Model (3) gives valid estimates of β. Model (3) yields valid estimates of the variance for parameters that are not forced into the model to deal with the sampling scheme. However, there are constraints

on the forced parameters due to the sampling scheme that do not appear in model (3). These constraints reduce the variability to which the forced parameters are subject. Thus, instead of a covariance matrix $D^{-1}(n)$, the appropriate covariance matrix has a term subtracted from $D^{-1}(n)$ to reduce certain aspects of the variability.

10.7 Residual Analysis and Outliers

Residuals are used in regression analysis to check normality, look for serial correlation, examine possible lack of fit, look for heteroscedasticity of variances, identify outliers, and generally to examine whether the assumptions of the regression model appear to be appropriate. Addressing many of these issues is somewhat less appropriate in analysis of variance. For example, appropriate tests for lack of fit are readily available and the question of serial correlation comes up less frequently.

For log-linear models, we will be interested in residuals primarily for identifying outliers and checking approximate normality. We define residuals by analogy with regression analysis.

In a regression model

$$Y = X\beta + e, \tag{1}$$

the residuals $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)$ are defined as the difference between the observations and their estimated expected values. Symbolically,

$$\hat{e} = Y - X\hat{\beta} = (I - H)Y,$$

where $\hat{\beta} = (X'X)^{-1}X'Y$ is the least squares estimate of β and $H = X(X'X)^{-1}X'$. If

$$e \sim N(0, \sigma^2 I)\,,$$

then

$$\hat{e} \sim N(0, \sigma^2(I - H))\,. \tag{2}$$

In most applications of residual analysis, the residuals are standardized before they are used. For example, in checking for outliers, we are checking for residuals that have unusually large absolute values. How large does a residual have to be before it is large enough to cause concern? If we standardize residuals so that they have a variance of about one, then we have a handle on what it means to have a large residual.

There are two methods of standardizing residuals that have been commonly used. One method is a crude standardization

$$\tilde{r}_i = \hat{e}_i/\hat{\sigma},$$

where $\hat{\sigma}^2$ is the mean squared error from fitting model (1). Recall that the object of standardization is to make the variance of the residual about 1.

Clearly, we should be dividing the residual by an estimate of its standard deviation. The standard deviation of $\hat{e}_i$ is $\sigma\sqrt{1-h_{ii}}$, where h_{ii} is the i'th diagonal element of H. The problem with the crude standardized residuals is that they ignore H. Instead of using the correct distribution (2), crude standardized residuals behave as if $\hat{e} \sim N(0, \sigma^2 I)$. This is the correct distribution for $e = Y - X\beta$, but it ignores the fact that $\hat{\beta}$ is estimated in $\hat{e} = Y - X\hat{\beta}$. In other words, the crude standardized residuals give just that: a very crude standardization. The only advantage to the crude standardized residuals is that they do not require the computation of the h_{ii} values.

The second method of standardizing residuals consists simply of doing it right. The standard deviation of $\hat{e}_i$ is $\sigma\sqrt{1-h_{ii}}$, so define the standardized residual as

$$r_i = \hat{e}_i \Big/ \hat{\sigma}\sqrt{1-h_{ii}}\,.$$

We now argue similarly for log-linear models. Again define the residuals as the difference between the observations and their estimated expected values. Symbolically, the residuals are

$$\hat{e}_i = n_i - \hat{m}_i\,. \tag{3}$$

The need for standardizing these residuals is so glaring that it is almost unheard of to define residuals as in (3). To repeat an intuitive argument given earlier, suppose we have a cell in which $n_i = 7$ and $\hat{m}_i = 2$, then $\hat{e}_i = 7 - 2 = 5$, which is not a very good fit. Now, suppose $n_i = 107$ and $\hat{m}_i = 102$. Again, $\hat{e}_i = 5$, but $\hat{m}_i$ seems to fit n_i quite well. With our standard sampling schemes, variability tends to be large when the numbers n_i and $\hat{m}_i$ are large. For example, under multinomial sampling, for each i, $\mathrm{Var}(n_i) = Np_i(1-p_i) = m_i(N-m_i)/N$. Unless p_i is very close to zero or one, the variance is large when n_i, and implicitly N, are large.

In order to standardize the residuals, we need a relationship similar to (2) for log-linear models. This relationship is essentially that, for large samples, the approximate distribution is

$$n - \hat{m} \sim N(0, D(m)(I-A))\,.$$

A more formal statement is given in the following theorem in which we make explicit the dependence of n and m on the sample size N.

Theorem 10.7.1. For multinomial, or product-multinomial sampling, if the log-linear model $\mu = X\beta$ holds, then as $N \to \infty$,

$$N^{-1/2}(n_N - \hat{m}_N) \xrightarrow{L} N(0, D(I-A))$$

where $D = D(m^*)$, m^* is defined as in Section 3, $A = X(X'DX)^{-1}X'D$, and N is $n_{\cdot}$ for multinomial sampling and $n_{\cdot\cdot}$ for product-multinomial sampling.

Proof. An argument similar to that in the proof of Lemma 12.3.3 gives

$$N^{1/2}[N^{-1}\hat{m}_N - N^{-1}m_N - DAD^{-1}N^{-1}(n_N - m_N)] \xrightarrow{P} 0\,.$$

[This is obtained by doing a Taylor expansion of the function $\hat{m}(\cdot)$ rather than $\hat{\mu}(\cdot)$.] Adding and subtracting $N^{-1/2}n_N$ and multiplying by -1 gives

$$\begin{aligned} &N^{1/2}[N^{-1}(n_N - \hat{m}_N) - N^{-1}(n_N - m_N) + DAD^{-1}N^{-1}(n_N - m_N)] \\ &\qquad = [N^{-1/2}(n_N - \hat{m}_N) - (I - DAD^{-1})N^{-1/2}(n_N - m_N)] \xrightarrow{P} 0\,. \end{aligned}$$

It follows that $N^{-1/2}(n_N - \hat{m}_N)$ and $(I - DAD^{-1})N^{-1/2}(n_N - m_N)$ have the same asymptotic distribution. By Theorem 12.3.1,

$$N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D - DA_z).$$

Some algebra shows that

$$(I - DAD^{-1})N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D(I - A)).$$

□

For large samples, Theorem 10.7.1 gives the approximation

$$n - \hat{m} \sim N(0, D(\hat{m})(I - A(\hat{m})))$$

where $A(\hat{m}) = X(X'D(\hat{m})X)^{-1}X'D(\hat{m})$. We are now in a position to define both standardized residuals and crude standardized residuals. *Crude standardized residuals* are defined by ignoring the fact that m is estimated or, in other words, by ignoring the matrix $A(\hat{m})$. Thus,

$$\tilde{r}_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i}}\,.$$

In discussions of residuals for contingency tables, these values are often called the residuals or the standardized residuals. In previous chapters, these were referred to as the *Pearson residuals*. As in linear models, the primary advantage of the crude standardized residuals is that they do not require the computation of the diagonal elements of a complicated matrix depending on X. Note also that the sum of the squared crude residuals is precisely the Pearson test statistic for lack of fit.

The *standardized residuals* are defined as

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i(1 - \hat{a}_{ii})}}$$

where $\hat{a}_{ii}$ is the i'th diagonal element of the square matrix $A(\hat{m})$. In some discussions of residuals, $\hat{a}_{ii}$ is defined to be the i'th diagonal element of

$D(\sqrt{\hat{m}})X(X'D(\hat{m})X)^{-1}X'D(\sqrt{\hat{m}})$. It is easily seen that the diagonal elements of these matrices are the same. These standardized residuals are also known as *adjusted residuals*. The term "adjusted" was introduced to distinguish these from the crude standardized residuals because the crude standardized residuals are often referred to as standardized residuals.

Given the standardized residuals, we can check for normality. Although maximum likelihood estimates and tests based on the likelihood ratio test statistics make sense with any sample size, we have discussed particular confidence intervals and tests that assume the validity of asymptotic distribution theory. We would like to know if this assumption is reasonable. One way to check is to see whether the standardized residuals really seem to be normally distributed. As in regression analysis, we can check this assumption by doing a normal (rankit) plot or a Shapiro-Francia test, cf. Christensen (1996a, Section 2.4). Note that the validity of these procedures depends on having a valid log-linear model.

Another way to check the validity of the asymptotic distributions is by comparing G^2 and X^2. If the asymptotic approximations are good and the model is true, then G^2 and X^2 should be about equal.

If the asymptotic distributions do not seem to be valid, we have a problem. One possible solution is simply to accept the fact that significance levels and confidence coefficients given by asymptotics are very crude. If our assumed sampling schemes are appropriate, the point estimates and test statistics are reasonable, but without valid distributions only crude conclusions can be made. The conditional approaches discussed in Section 3.5 or Bayesian methods similar to Chapter 13 can also be used here. Finally, another possibility is to try to incorporate a more realistic sampling scheme than the simple multinomial and product-multinomial schemes considered here.

The other primary use of standardized residuals is in identifying outliers. Standardized residuals are asymptotically distributed as $N(0,1)$, so we can test whether residuals really have mean zero. Typically, we would be interested in the standardized residuals with largest absolute values. This is equivalent to testing all residuals, so a multiple comparison method would be appropriate. The Bonferroni method is easy to apply (cf. Christensen, 1996a, Sections 6.2, 7.9 or Christensen, 1996b, Section 5.3). We declare that a case is an outlier if

$$|r_i| > z(1-\alpha/2q)$$

where $z(\eta)$ is the η'th percentile of a standard normal distribution, α is the size of the test, and q is the number of cells in the table.

An alternative method for identifying outliers is based on *exploratory data analysis*. This method does not rely on asymptotic distributions. Treating the standardized residuals as a sample, compute the quartiles and the interquartile range (IQR). Any case with a residual more than (1.5)IQR from the nearest quartile is considered an outlier.

Rather than using an ad hoc test for outliers based on standardized residuals, we can construct a likelihood ratio test. Suppose our model is

$$\mu = X\beta\,. \tag{4}$$

Without loss of generality, consider testing whether the observation in the q'th cell, n_q, is an outlier. An outlier in the q'th cell can be modeled by fitting a separate parameter to the cell. Let $v_q = (0,\ldots,0,1)'$ and consider the model

$$\mu = X\beta + v_q\gamma\,. \tag{5}$$

The likelihood ratio test of this model against the reduced model (4) is a test of whether the q'th cell is an outlier.

Typically, we would want to examine each cell for being an outlier; thus, we need q likelihood ratio test statistics. Again, applying the Bonferroni method for multiple comparisons would be appropriate.

Computing each of the q likelihood ratio test statistics requires an iterative procedure for obtaining estimates in models like model (5). In computing estimates of m, μ, and β in model (5), we can use estimates from model (4) as starting values. To reduce costs, we might stop after just one step of the iterative procedure. For these one-step procedures, closed forms for the likelihood ratio test can be obtained. Unfortunately, there are several possible approaches to deriving one-step approximations and it is not clear which, if any of them, work well. The remainder of this section is devoted to deriving a one-step approximation to Cook's distance.

Derivation of Cook's Distance

Rewrite model (5) as

$$\mu_{[q]} = \log(m_{[q]}) = X\beta + v_q\gamma\,,$$

where $\mu_{[q]}$ and $m_{[q]}$ are used to distinguish the parameters in model (5), where cell q may be an outlier, from the parameters in model (4). The MLE of $m_{[q]}$ must satisfy

$$\begin{pmatrix} X' \\ v_q' \end{pmatrix} n = \begin{pmatrix} X' \\ v_q' \end{pmatrix} \hat{m}_{[q]}$$

where $\log(\hat{m}_{[q]}) \in C(X, v_q)$. In the discussion below, a subscript (q) indicates that the row corresponding to case q has been deleted from a matrix or vector, so $X = \begin{bmatrix} X_{(q)} \\ x_q' \end{bmatrix}$. Notice that $C(X, v_q) = C\left(\begin{bmatrix} X_{(q)} & 0 \\ 0 & 1 \end{bmatrix}\right)$. Thus, $\hat{m}_{[q]}$ satisfies

$$\begin{bmatrix} X_{(q)}' & 0 \\ 0 & 1 \end{bmatrix} n = \begin{bmatrix} X_{(q)}' & 0 \\ 0 & 1 \end{bmatrix} \hat{m}_{[q]}$$

or, equivalently, writing $\hat{m}'_{[q]} = (\hat{m}'_{[q](q)}, \hat{m}_{[q]q})$,

$$X'_{(q)}n_{(q)} = X'_{(q)}\hat{m}_{[q](q)}$$

and

$$n_q = \hat{m}_{[q]q}\,.$$

Thus, $\hat{m}_{[q](q)}$ can be obtained by fitting the model, say $\mu_{(q)} = X_{(q)}\beta_{(q)}$, in which cell q has been deleted and $\beta_{(q)}$ denotes the new parameter vector that applies to this model. In particular, $\hat{\mu}_{[q](q)} = \hat{\mu}_{(q)}$.

A natural version of Cook's distance that is appropriate for log-linear models is

$$C_q(X'D(\hat{m})X, p) = \frac{(\hat{\beta} - \hat{\beta}_{(q)})'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}_{(q)})}{p}.$$

This is the same measure as used in Section 6.7, but is written in a different form. Note that $X'D(\hat{m})X$ is the inverse of the estimated asymptotic covariance matrix for $\hat{\beta}$ under model (4) with Poisson sampling. A one-step version of Cook's distance is

$$C_q^1(X'D(\hat{m})X, p) = \frac{(\hat{\beta} - \hat{\beta}^1_{(q)})'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}^1_{(q)})}{p}$$

where $\hat{\beta}^1_{(q)}$ is a one-step approximation to $\hat{\beta}_{(q)}$.

Using the Newton-Raphson method with a starting value of $\hat{\beta}$ and a result similar to Proposition 13.5.1 in Christensen (1996b) on the inverse of a sum of matrices, the one-step estimate is

$$\begin{aligned}
\hat{\beta}^1_{(q)} &= \hat{\beta} + [X_{(q)}D(\hat{m}_{(q)})X_{(q)}]^{-1}X'_{(q)}[n_{(q)} - \hat{m}_{(q)}] \\
&= \hat{\beta} + [X'D(\hat{m})X - \hat{m}_q x_q x'_q]^{-1}[X'(n - \hat{m}) - x_q(n_q - \hat{m}_q)] \\
&= \hat{\beta} + [X'D(\hat{m})X - \hat{m}_q x_q x'_q]^{-1}[-x_q(n_q - \hat{m}_q)] \\
&= \hat{\beta} - \left[(X'D(\hat{m})X)^{-1} + \frac{\hat{m}_q}{1 - \hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q x'_q(X'D(\hat{m})X)^{-1}\right] \\
&\quad \times [x_q(n_q - m_q)] \\
&= \hat{\beta} - \frac{1}{1 - \hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q(n_q - \hat{m}_q)\,.
\end{aligned}$$

The equation

$$\hat{\beta}^1_{(q)} = \hat{\beta} - \frac{n_q - \hat{m}_q}{1 - \hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q$$

leads to a computational formula for the one-step version of Cook's distance. The one-step version can be written as

$$C_q^1(X'D(\hat{m})X, p) = \frac{(\hat{\beta} - \hat{\beta}^1_{(q)})'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}^1_{(q)})}{p}$$

$$= \frac{1}{p}\frac{\hat{a}_{qq}}{(1-\hat{a}_{qq})^2}\frac{(n_q-\hat{m}_q)^2}{\hat{m}_q} = \frac{1}{p}r_q^2\frac{\hat{a}_{qq}}{1-\hat{a}_{qq}}.$$

As mentioned just prior to Subsection 6.7.1, this definition of Cook's distance has weaknesses. Primarily, it does not take into account the marginal constraints imposed by multinomial or product-multinomial sampling, so it is most appropriate for Poisson sampling. In general, the implications of deleting a cell in a multinomial distribution are hard to grasp. Anderson (1992) has a valuable idea. For multinomial sampling, rather than merely deleting cell i, he proposes looking at the probabilities than an observation occurs in the other cells conditional on the observation not appearing in cell i. He then develops a version of Cook's distance that can be written in terms of the standardized residuals. Unfortunately, Anderson (1992) uses standardized residuals that seem to conflict with Theorem 10.7.1; i.e., he seems to have a different asymptotic variance for $N^{-1/2}(n_i-\hat{m}_i)$. The problem appears to be that he does not use the term A_z in Theorem 10.3.1b. [Of course, the really fun thing about this is that the reader gets to wonder who is making the mistake. Anderson's standardized residuals are based on Rao (1973, p. 394). Theorem 10.7.1 is, I believe, identical to a result in Haberman (1974a).] In any case, the reported results should be used with care. In other work, Thomas and Cook (1989, 1990) discuss influence for generalized linear models (which include log-linear models, cf. Chapter 9).

10.8 Exercises

EXERCISE 10.8.1. Show that for a 3×3 table with $n_{11}=n_{13}$, $n_{31}=n_{33}$, and $n_{\cdot 1}=n_{\cdot 3}$ that the $\hat{m}$'s for the independence model are also the $\hat{m}$'s for the uniform association model.

EXERCISE 10.8.2. Waite (1911) reports data on classifications of general intelligence made for students from a secondary school in London. Classifications were made after two different school terms and each student was classified by two instructors. Classifications were based on Pearson's criteria as explained in Exercise 2.6.3. The data are given in Table 10.1. The entry for Term 1, row C, column D of 55 indicates that there were 55 people who were classified as C by one instructor and D by the other. We want to develop interesting log-linear models for such data. For the moment, consider only Term 1. A model of interest is that the two teachers have the same marginal distribution of assigning various classifications and that they assign them independently. Write the marginal probabilities for the categories C, D, E, F, and G as p_1, p_2, p_3, p_4, and p_5, respectively. What are the table probabilities p_{ij} in terms of the marginal probabilities? Take logs of the p_{ij}'s to identify a log-linear model. Write the log-linear model $\log(m_{ij})=\alpha_i+\beta_j$ for these data in matrix form. Incorporate the

restrictions $\alpha_i = \beta_i$ to get a model $\log(m) = X\gamma$ and fit the model to the data of Table 10.1. What conclusions can you reach about the data?

TABLE 10.1. Intelligence Classifications

	Term 1				
	C	D	E	F	G
C	13	55	29	1	0
D		123	326	24	0
E			421	253	17
F				107	31
G					5

	Term 2				
	C	D	E	F	G
C	17	51	17	1	0
D		129	479	46	5
E			700	343	28
F				109	72
G					21

EXERCISE 10.8.3. Use the saturated model and Theorem 10.2.1 to find the large sample distribution of a multinomial sample in terms of its category probabilities.

EXERCISE 10.8.4. *The Delta Method.*
Let v_N be a sequence of $q \times 1$ random vectors and suppose that

$$\sqrt{N}(v_N - \theta) \xrightarrow{L} N\big(0, \Sigma(\theta)\big).$$

Suppose that $F(\cdot)$ is a differentiable function taking q vectors into r vectors. Let dF be the $r \times q$ matrix of partial derivatives of F. Then

$$\sqrt{N}\big(F(v_N) - F(\theta)\big) \xrightarrow{L} N\big(0, dF\,\Sigma(\theta)\,dF'\big).$$

For technical reasons, it is advantageous to assume that dF and $\Sigma(\theta)$ are also continuous. For mathematical details, see Bishop, Fienberg, and Holland (1975, Section 14.6.3).

Assuming the saturated model for a 2×2 table, use Theorem 10.2.1c and the delta method to find an asymptotic standard error and asymptotic confidence intervals for the odds ratio. Show that the intervals do not change if based on Theorem 10.3.1b. Apply this method to the data of Example 2.1.1 to get a 95% interval. How does this interval compare to the interval given at the end of the subsection on The Odds Ratio in Section 2.1?

EXERCISE 10.8.5. Use the delta method of the previous exercise to show that if

$$\sqrt{N}(v_N - \theta) \xrightarrow{L} N\big(0, \Sigma(\theta)\big),$$

then for any $r \times q$ matrix A,

$$\sqrt{N}(Av_N - A\theta) \xrightarrow{L} N\big(0, A\Sigma(\theta)A'\big).$$

Show that if $\mathrm{Cov}(\sqrt{N}v_N) = \Sigma(\theta)$, then $\mathrm{Cov}(\sqrt{N}Av_N) = A\Sigma(\theta)A'$.

EXERCISE 10.8.6. *Testing Marginal Homogeneity in a Square Table.* For an $I \times I$ table, the hypothesis of marginal homogeneity is

$$H_0 : \; p_{i\cdot} = p_{\cdot i}\,,$$

$i = 1, \ldots, I$. A natural statistic for testing this hypothesis is the vector $d = (n_{1\cdot} - n_{\cdot 1}, \ldots, n_{I\cdot} - n_{\cdot I})'$ Clearly,

$$\mathrm{E}(d_i) = p_{i\cdot} - p_{\cdot i}\,.$$

Use the results of Exercise 1.6.5 to show that

$$\mathrm{Var}(d_i) = N\left[(p_{i\cdot} + p_{\cdot i} - 2p_{ii}) - (p_{i\cdot} - p_{\cdot i})^2\right]$$

and

$$\mathrm{Cov}(d_h, d_i) = N\left[(p_{hi} + p_{ih}) + (p_{h\cdot} - p_{\cdot h})(p_{i\cdot} - p_{\cdot i})\right].$$

Use the previous exercise to find the large sample distribution of d. Show that $\Pr(J'd = 0) = 1$. Show that the asymptotic covariance matrix of d has rank $I - 1$ and thus is not invertible. It is well known that if $Y \sim N(0, V)$ with V an $s \times s$ nonsingular matrix, then $Y'V^{-1}Y \sim \chi^2(s)$. Use this fact along with the asymptotic distribution of d to obtain a test of the hypothesis of marginal homogeneity. Apply the test of marginal homogeneity to the data of Exercise 2.6.10.

Chapter 11

The Matrix Approach to Logit Models

In this chapter, we again discuss logistic regression and logit models, but here we use the matrix approach of Chapter 10. Section 1 discusses the equivalence of logit models and log-linear models. This equivalence is used to arrive at results on estimation and testing. Because the data in a typical logistic regression correspond to very sparse data in a contingency table, the asymptotic results of Section 10.2 are not appropriate. Section 6 presents results from Haberman (1977) that *are* appropriate for logistic regression models. Section 2 discusses model selection criteria for logistic regression. Direct fitting of logit models is considered in Section 3. The appropriate maximum likelihood equations and Newton-Raphson procedure are given. Section 4 indicates how the weighted least squares model-fitting procedure is applied to logit models. Models appropriate for response variables with more than two categories are examined in Section 5. Finally, Section 7 considers the discrimination problem.

11.1 Estimation and Testing for Logistic Models

In general, if the dependent variable has only two categories, regardless of the number of predictor variables, the table can be considered as a two-dimensional table with two columns (one column for each category of the dependent variable). In this structure, all predictor variables are being pooled into the rows of the table. For t distinct sets of predictor variables,

we can write the $2t \times 1$ vector of observations as

$$n = (n_{11}, n_{21}, \ldots, n_{t1}, n_{12}, \ldots, n_{t2})'$$

with similar notations for p, m, and μ. A logistic model is a linear model for the values $\log(p_{i1}/p_{i2})$. Note that we are modeling the log odds of category 1 compared to category 2. These are the log odds of observing category 1 given that the observation falls in row i. If each row constitutes an independent binomial, then we are simply modeling the log odds for the various binomials.

Logistic models are nothing more than log-linear models. All of the results of Chapter 10 apply to logistic models. We now consider the exact nature of this equivalence for prospective studies.

Let $\eta = \big(\log(p_{11}/p_{12}), \log(p_{21}/p_{22}), \ldots, \log(p_{t1}/p_{t2})\big)'$ be the vector of log odds. A linear logistic model is a model $\eta = X\beta$, where X is a $t \times k$ matrix. Define

$$L' = [I_t, -I_t]$$

and note that for prospective studies, $\log(p_{i1}/p_{i2}) = \log(m_{i1}/m_{i2}) = \log(m_{i1}) - \log(m_{i2})$, so

$$\eta = L'\mu\,.$$

Thus, the logistic model can be written as

$$L'\mu = X\beta\,.$$

Now define a log-linear model

$$\mu = X_*\xi$$

where

$$X_* = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \quad \text{and} \quad \xi = \begin{bmatrix} \gamma \\ \beta \end{bmatrix}.$$

It is easily seen that if $\mu = X_*\xi$, then $L'\mu = X\beta$. It is only moderately more difficult to see (cf. Section 12.4 and Christensen, 1996b, Section 3.3) that

$$\{\mu | L'\mu = X\beta\} = C(X_*)\,.$$

Thus, the restriction on μ imposed by the logistic model $\eta = X\beta$ is precisely the same as the log-linear model $\mu = X_*\xi$. In other words, *the logistic model $\eta = X\beta$ is identical to the log-linear model $\mu = X_*\xi$.*

Unfortunately, there can be problems with the asymptotic results of Chapter 10 when applied to logistic models. The asymptotic results of Section 10.2 are based on the assumption of a fixed number of cells q in the table. This number is $q = 2t$. It is assumed that the sample size in each cell gets large. Often, logistic models are used in situations that are more similar to regression than analysis of variance. In such cases, any additional

observations obtained typically correspond to new rows of the design matrix X. This implies the addition of a new row to the $t \times 2$ table; thus, the assumption of a fixed number of cells is invalidated. These issues are dealt with in Section 6.

In practice, the data are fixed and neither the sample sizes nor the number of cells increases. Both remain constant. If the number of observations in each cell is reasonably large, the usual asymptotic theory should work adequately. If the number of cells is large relative to the number of observations, new asymptotic results are required as a basis for statistical inference.

Frequently, when the dependent variable has two categories, the data are collected so that for each unique set of predictor variables (i.e., for each row of the $t \times 2$ table), the counts are independent and have a binomial distribution. As in Section 10.4, the existence of product-multinomial sampling (product-binomial, in this instance) restricts us to a special form for the design matrix of the log-linear model. In particular, it is required that $C(Z) \subset C(X_*)$ where Z is a matrix of indicators for the rows of the $t \times 2$ table. With $\mu = (\mu_{11}, \ldots, \mu_{t1}, \mu_{12}, \ldots, \mu_{t2})'$,

$$Z = \begin{bmatrix} I_t \\ I_t \end{bmatrix}.$$

Since

$$X_* = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix},$$

it is clearly the case that $C(Z) \subset C(X_*)$. It follows that hypothesizing a logistic model automatically makes the model appropriate for product-binomial sampling.

Moreover, when fitting logistic models, it suffices to imagine fitting a log-linear model to a table with product-binomial sampling. This mental device of imagining product-binomial sampling assures that the structure of the log-linear model implies the existence of a valid logistic model. To see this, note that a quite arbitrary model appropriate for product-binomial sampling can be written with the design matrix

$$\begin{bmatrix} I_t & X_1 \\ I_t & X_2 \end{bmatrix}.$$

However, $C\left(\begin{bmatrix} I_t & X_1 \\ I_t & X_2 \end{bmatrix}\right) = C\left(\begin{bmatrix} I_t & X_1 - X_2 \\ I_t & 0 \end{bmatrix}\right)$ where $\begin{bmatrix} I_t & X_1 - X_2 \\ I_t & 0 \end{bmatrix}$ has the form given earlier for logistic models.

Estimation

We now consider the problem of estimation in a logistic model $\eta = X\beta$. Recall that X is $t \times k$ and that η and β are t and k vectors, respectively.

In particular, consider estimation of a linear function $\rho'\eta$, where ρ is an arbitrary $t \times 1$ vector. Note that $\rho'\eta = \rho' X\beta$, so these functions can be considered as functions of β. If $\text{rank}(X) = k$ and e_i is a t vector of 0s with a 1 in the ith position, then choosing $\rho' = e_i'(X'X)^{-1}X'$ gives $\rho' X\beta = \beta_i$, where $\beta = (\beta_1, \ldots, \beta_k)'$.

Write

$$X_L = \begin{bmatrix} X \\ 0 \end{bmatrix},$$

where X_L is $2t \times k$ and 0 is a $t \times k$ matrix of zeros. As above, $\eta = X\beta$ is equivalent to

$$\begin{aligned} \mu = X_*\xi &= \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} = [Z, X_L] \begin{bmatrix} \gamma \\ \beta \end{bmatrix} \\ &= Z\gamma + X_L\beta, \end{aligned}$$

where β is identical in the logistic and log-linear models.

Using invariance of the MLEs, the MLE of $\rho'\eta$ comes directly from the MLE of μ. Because $\eta = L'\mu$,

$$\rho'\hat{\eta} = \rho' L'\hat{\mu}.$$

In terms of estimating β, we get

$$\begin{aligned} \rho' X\hat{\beta} &\equiv \rho'\hat{\eta} \\ &= \rho' L'\hat{\mu} \\ &= \rho' L'(Z\hat{\gamma} + X_L\hat{\beta}) \\ &= \rho' L' X_L\hat{\beta} \\ &= \rho' X\hat{\beta}. \end{aligned}$$

Thus, $\rho' X\beta$ can be estimated in either the logistic model or the log-linear model and the estimates are identical.

To form tests and confidence intervals for $\rho' X\beta$, we need a distribution for $\rho' X\hat{\beta}$. Asymptotically, for any vector ρ_*,

$$\frac{\rho_*'\hat{\mu} - \rho_*' X_*\xi}{\sqrt{\rho_*'(A - A_z)D^{-1}(m)\rho_*}} \sim N(0,1) \tag{1}$$

and

$$\begin{aligned} A - A_z &= X_*(X_*' D X_*)^{-1} X_*' D - Z(Z'DZ)^{-} Z'D \\ &= X_*(X_*' D(m) X_*)^{-1} X_*' D(m) - Z'(Z' D(m) Z)^{-1} Z' D(m), \end{aligned}$$

so

$$(A - A_z)D^{-1}(m) = X_*(X_*' D(m) X_*)^{-1} X_*' - Z(Z' D(m) Z)^{-1} Z'.$$

For estimating $\rho' X\beta$ in a logistic model, let $\rho_*' = \rho' L'$, so $\rho_*'\hat{\mu} = \rho'\hat{\eta}$ and $\rho_*' X_* \xi = \rho' X\beta$. To apply (1), we need to find $\rho_*'(A - A_z)D^{-1}(m)\rho_*$. In the appendix to this section, it is shown that

$$\rho_*'(A - A_z)D^{-1}(m)\rho_* = \rho' X[X'D(b)X]^{-1}X'\rho\,, \tag{2}$$

where

$$b = (b_1, \ldots, b_t)'$$

and

$$b_i = m_{i1}m_{i2}/(m_{i1} + m_{i2})\,.$$

Taking $\hat{b}_i = \hat{m}_{i1}\hat{m}_{i2}/n_{i\cdot}$ and $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_t)'$ gives, asymptotically,

$$\frac{\rho'\hat{\eta} - \rho' X\beta}{\sqrt{\rho' X[X'D(\hat{b})X]^{-1}X'\rho}} \sim N(0,1)\,.$$

Tests and confidence intervals follow in the usual way. In particular, using $\rho' = e_i' = (0, \cdots, 0, 1, 0, \cdots, 0)$, for large samples

$$\frac{\log(\hat{p}_{i1}/\hat{p}_{i2}) - \log(p_{i1}/p_{i2})}{\sqrt{\hat{a}_{ii}/\hat{b}_i}} \sim N(0,1)\,, \tag{3}$$

where $\hat{a}_{ii}$ is the leverage as found in Section 4.3 as modified by Subsection 4.4.1. The asymptotic variance for log odds ratios can also be computed. Except for the difference in the weights b_i, the value

$$\mathrm{Var}(\rho'\hat{\eta}) = \rho' X\,(X'D(b)X)^{-1}\,X'\rho$$

looks like that used in Section 10.2. Methods for evaluating variances are similar.

Using the delta method of Exercise 10.8.4, the logistic transform, (3), and writing $N_i = n_{i1} + n_{i2}$, asymptotic inferences for p_{ij} are based on

$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})\hat{a}_{ii}/N_i}} \sim N(0,1)\,,$$

cf. Exercise 11.8.5

Testing Hypotheses

Assume a logistic model $\eta = X\beta$ and consider the problem of testing a reduced model $\eta_0 = X_0\beta_0$ against $\eta = X\beta$, where $C(X_0) \subset C(X)$. This test can be performed by testing log-linear models. The full model corresponds to $\mu = X_*\xi$. We can write the reduced model as

$$\mu_0 = X_{*0}\xi_0,$$

where

$$X_{*0} = [Z, X_{L0}]$$

and

$$X_{L0} = \begin{bmatrix} X_0 \\ 0 \end{bmatrix} .$$

The degrees of freedom for the chi-square test is $\text{rank}(X_*) - \text{rank}(X_{*0}) = \text{rank}(X) - \text{rank}(X_0)$. The likelihood ratio test statistic is

$$G^2 = 2 \sum_{i=1}^{t} \sum_{j=1}^{2} \hat{m}_{ij} \log(\hat{m}_{ij}/\hat{m}_{0ij}) .$$

Appendix

To simplify notation somewhat, let

$$D_m \equiv D(m).$$

In this appendix, we wish to show equation (2), i.e., for $\rho'_* = \rho' L'$,

$$\rho'_* X_* [X'_* D_m X_*]^{-1} X'_* \rho - \rho'_* Z [Z' D_m Z]^{-1} Z' \rho_* = \rho' X [X' D(b) X]^{-1} X' \rho.$$

The algebra necessary for this demonstration is quite nasty. We break it up into several parts.

Lemma 11.1.1. $\rho'_* Z [Z' D_m Z]^{-1} Z' \rho_* = 0.$

Proof. $\rho'_* Z = \rho' L' Z$ but $L' Z = 0.$ □

Now, all we have to deal with is

$$\rho'_* X_* [X'_* D_m X]^{-1} X'_* \rho .$$

We will rewrite this using a perpendicular projection operator (cf. Christensen, 1996b, Appendix B) and then use a property of perpendicular projection operators to derive the result. Let

$$D_m^{1/2} = D(\sqrt{m_{11}}, \ldots, \sqrt{m_{t2}})$$

so that

$$\begin{aligned} & \rho'_* X_* [X'_* D_m X_*]^{-1} X'_* \rho_* \\ &= \rho' L' X_* [X'_* D_m X_*]^{-1} X'_* L \rho \\ &= \rho' L' D_m^{-1/2} \left[D_m^{1/2} X_* [X'_* D_m X_*]^{-1} X'_* D_m^{1/2} \right] D_m^{-1/2} L \rho \qquad (4) \\ &= \rho' L' D_m^{-1/2} P D_m^{-1/2} L \rho , \end{aligned}$$

where $P = D_m^{1/2} X_*[X_*' D_m X_*]^{-1} X_*' D_m^{1/2}$. The matrix P is the perpendicular projection operator onto

$$C(D_m^{1/2} X_*) = C(D_m^{1/2} Z, D_m^{1/2} X_L).$$

We need one property of the perpendicular projection operator (cf. Christensen, 1996b, Section 9.2), namely

$$P = M_1 + M_2,$$

where

$$M_1 = D_m^{1/2} Z[Z' D_m Z]^{-1} Z' D_m^{1/2} \tag{5}$$

and

$$M_2 = (I - M_1) D_m^{1/2} X_L [X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L]^{-1} \times X_L' D_m^{1/2} (I - M_1). \tag{6}$$

From (4), we see that

$$\rho_*' X_* [X_*' D_m X_*]^{-1} X_*' \rho = \rho' L' D_m^{-1/2} M_1 D_m^{-1/2} L\rho + \rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L\rho\,.$$

The first term on the right-hand side vanishes.

Lemma 11.1.2. $\rho' L' D_m^{-1/2} M_1 D_m^{-1/2} L\rho = 0.$

Proof. Note that $0 = L'Z = L' D_m^{-1/2} [D_m^{1/2} Z]$. Using formula (5) for M_1, we see that $0 = L' D_m^{-1/2} M_1$. □

By Lemmas 11.1.1 and 11.1.2, we have reduced the demonstration of equation (2) to showing

$$\rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L\rho = \rho' X [X' D(b) X]^{-1} X' \rho\,. \tag{7}$$

Again, we break the demonstration into parts.

Lemma 11.1.3. $\rho' L' D_m^{-1/2} (I - M_1) D_m^{1/2} X_L = \rho' X.$

Proof. As in the proof of Lemma 11.1.2, $L' D_m^{-1/2} M_1 = 0$. Thus,

$$\begin{aligned} \rho' L' D_m^{-1/2} (I - M_1) D_m^{1/2} X_L &= \rho' L' D_m^{-1/2} D_m^{1/2} X_L \\ &= \rho' L' X_L \\ &= \rho' X\,. \end{aligned}$$

□

Define $m_1 = (m_{11}, \ldots, m_{t1})'$ and $m_2 = (m_{12}, \ldots, m_{t2})'$.

Lemma 11.1.4. $X_L' D_m Z = X' D(m_1)$.

Proof.

$$\begin{aligned} X_L' D_m Z &= [X', 0'] \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix} \begin{bmatrix} I_t \\ I_t \end{bmatrix} \\ &= X' D(m_1)\,. \end{aligned}$$

□

A similar argument yields

Lemma 11.1.5. $X_L' D_m X_L = X' D(m_1) X$.

We need two additional results.

Lemma 11.1.6. $[Z' D_m Z] = D(m_1 + m_2)$.

Proof.

$$\begin{aligned} Z' D_m Z &= [I_t, I_t] \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix} \begin{bmatrix} I_t \\ I_t \end{bmatrix} \\ &= D(m_1) + D(m_2) \\ &= D(m_1 + m_2). \end{aligned}$$

□

Lemma 11.1.7. $X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L = X' D(b) X$.

Proof. Using Lemmas 11.1.4, 11.1.5, and 11.1.6 gives

$$\begin{aligned} & X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L \\ &= X_L' D_m X_L - X_L' D_m Z [Z' D_m Z]^{-1} Z' D_m X_L \\ &= X' D(m_1) X - X' D(m_1) D^{-1} (m_1 + m_2) D(m_1) X \\ &= X' [D(m_1) - D(m_1) D^{-1} (m_1 + m_2) D(m_1)] X \\ &= X' D(b) X\,. \end{aligned}$$

□

We can now obtain equation (7). Using (6), Lemmas 11.1.3 and 11.1.7 give

$$\rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L \rho = \rho' X [X' D(b) X]^{-1} X' \rho$$

and we are done.

11.2 Model Selection Criteria for Logistic Regression

The purpose of this section is to show that not only do the model selection criteria of Section 3.6 apply to logistic regression, but that they have interpretations similar to those in normal theory regression.

For a logistic regression model $\eta = X\beta$ or, equivalently,

$$\mu = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

write the likelihood ratio test statistic for testing the model against the saturated model as $G^2(X)$. Let J be a $t \times 1$ matrix of one's.

A natural definition of R^2 for logit models gives precisely the same definition as for log-linear models. In defining R^2 for logistic regression, the smallest interesting model is typically the model that contains only the intercept

$$\eta = J\gamma.$$

The corresponding log-linear model is

$$\mu = \begin{bmatrix} I_t & J \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \gamma \end{bmatrix} \tag{1}$$

which is equivalent to

$$\mu = X_0\delta \equiv \begin{bmatrix} I_t & J & 0 \\ I_t & 0 & J \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \end{bmatrix}. \tag{2}$$

These are equivalent because the μ vectors that can be obtained from the models are identical. Any μ from model (1) can be obtained from model (2) by taking $\alpha = \xi$, $\gamma_1 = \gamma$, and $\gamma_2 = 0$. Conversely, any μ from model (2) can be obtained from model (1) by taking $\xi = \alpha + J\gamma_2$ and $\gamma = \gamma_1 - \gamma_2$. This is really just a demonstration that $C(X_0)$ is identical to the column space of the model matrix in (1). If we think of logistic regression as the analysis of a $t \times 2$ table, model (2) is

$$\log(m_{ij}) = \alpha_i + \gamma_j$$

which is the model of independence. So comparing the logistic model $\eta = X\beta$ to the logistic intercept model $\eta = J\gamma$ is the same as comparing the log-linear equivalent of $\eta = X\beta$ to the independence model.

The $G^2(X)$ statistic is the same whether considering logit models or their log-linear equivalents. If $k = \text{rank}(X)$, the degrees of freedom are the number of cells in the table minus the rank of the log-linear model design matrix, $2t - (t + k) = t - k$. Note that this can also be viewed as the number

of cases (number of independent binomials) minus the rank of the logistic model design matrix, just like the degrees of freedom error in a normal theory model. The independence model is equivalent to fitting an intercept in logistic regression, so $G^2(X_0)$ has degrees of freedom $2t - t - 1 = t - 1$. Note that this is the number of cases minus 1 for the intercept, just like the error for fitting only an intercept in normal theory regression.

The log-linear model definition

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)}$$

(cf. Section 3.6) makes perfect sense when applied to logistic regression models. The definition of Adj R^2 when applied to logit models gives

$$\text{Adj } R^2 = 1 - \frac{G^2(X)/(t-k)}{G^2(X_0)/(t-1)}\,.$$

Finally, Akaike's information criterion suggests picking X to minimize

$$\begin{aligned} A_x &= G^2(X) - \big[(2t) - 2\big(t+k\big)\big] \\ &= G^2(X) + 2k. \end{aligned}$$

Because t is fixed, minimizing A_x is equivalent to minimizing

$$A_x - q \equiv A_x - 2t = G^2(X) - 2[t-k]\,.$$

As illustrated in Section 4.1, it is common to report A^*, the information relative to a full model.

11.3 Likelihood Equations and Newton-Raphson

When dealing with logit models, some simplification occurs in the likelihood equations and the Newton-Raphson algorithm. Write the log-linear model version of the logit model $\eta = X\beta$ as

$$\mu = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix}\,, \tag{1}$$

where X is a $t \times k$ matrix. Write $m_j = (m_{1j}, \ldots, m_{tj})'$ and $n_j = (n_{1j}, \ldots, n_{tj})'$ for $j = 1, 2$, so $m' = (m_1', m_2')$ and $n' = (n_1', n_2')$. Also write $N = (N_1, \ldots, N_t)'$, where

$$N_i = n_{i1} + n_{i2}\,.$$

As in Chapter 10, the likelihood equations are

$$\begin{bmatrix} I_t' & I_t' \\ X' & 0 \end{bmatrix} (n - m) = 0\,.$$

Noting that $n_2 = N - n_1$, we get the equations

$$\begin{bmatrix} I_t' & I_t' \\ X' & 0 \end{bmatrix} \begin{bmatrix} n_1 - m_1 \\ N - n_1 - m_2 \end{bmatrix} = 0$$

or

$$\begin{bmatrix} (n_1 - m_1) + (N - n_1 - m_2) \\ X'(n_1 - m_1) \end{bmatrix} = 0\,,$$

which simplifies to

$$\begin{bmatrix} N - (m_1 + m_2) \\ X'(n_1 - m_1) \end{bmatrix} = 0\,. \tag{2}$$

We are seeking values $\hat{\gamma}$ and $\hat{\beta}$ that give solutions to equation (2). We now show that the value of $\hat{\gamma}$ can be determined by the value of $\hat{\beta}$. Write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}$$

and $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_t)'$; then,

$$\begin{aligned} \hat{m}_{i1} &= \exp[\hat{\gamma}_i + x_i'\hat{\beta}]\,, \\ \hat{m}_{i2} &= \exp[\hat{\gamma}_i]\,. \end{aligned} \tag{3}$$

Because $\hat{\gamma}$ and $\hat{\beta}$ provide a solution to (2), $N_i = \hat{m}_{1i} + \hat{m}_{2i}$ and

$$\begin{aligned} N_i &= \exp[\hat{\gamma}_i + x_i'\hat{\beta}] + \exp[\hat{\gamma}_i] \\ &= e^{\hat{\gamma}_i}[1 + \exp(x_i'\hat{\beta})]\,, \end{aligned}$$

so

$$e^{\hat{\gamma}_i} = N_i/[1 + \exp(x_i'\hat{\beta})] \tag{4}$$

and

$$\hat{\gamma}_i = \log(N_i/[1 + \exp(x_i'\hat{\beta})])\,.$$

This completes the demonstration.

Because $\hat{\gamma}$ is a function of $\hat{\beta}$, the likelihood equations can be reduced to the bottom half of (2), which is solely a function of β. The top half of (2) is satisfied for any $\hat{\beta}$ by taking $\hat{\gamma}$ as indicated above. To write the bottom half of (2) as a function of β, we need only write m_1 as a function of β. From equations (3) and (4)

$$\begin{aligned} m_{i1} &= e^{\gamma_i} e^{x_i'\beta} \\ &= (N_i/[1 + e^{x_i'\beta}])e^{x_i'\beta} \\ &= N_i \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}\,. \end{aligned} \tag{5}$$

We can use the Newton-Raphson algorithm to find a solution to the likelihood equations

$$X'(n_1 - m_1) = 0\,.$$

The highlights of using Newton-Raphson are given below. More detailed discussions are given in Chapter 10 and Section 12.4. To apply Newton-Raphson, we need the derivative of $X'(n_1 - m_1(\beta))$ with respect to β, i.e., the matrix of partial derivatives with respect to the β_j's. Using the chain rule, this is just $-X'$ times the matrix of partial derivatives of $m_1(\beta)$ with respect to β. Note that

$$\begin{aligned} m_1(\beta) &= [m_{11}(\beta), \ldots, m_{t1}(\beta)]' \\ &= \left[N_1 e^{x_1'\beta}/(1 + e^{x_1'\beta}), \ldots, N_t e^{x_t'\beta}/(1 + e^{x_t'\beta})\right]' . \end{aligned}$$

The partial derivative of $m_{1i}(\beta)$ with respect to β_j is

$$\begin{aligned} \frac{\partial m_{i1}(\beta)}{\partial \beta_j} &= N_i x_{ij} e^{x_i'\beta}(1 + e^{x_i'\beta})^{-1} \\ &\quad + N_i e^{x_i'\beta}(-1)(1 + e^{x_i'\beta})^{-2} x_{ij} e^{x_i'\beta} \qquad (6) \\ &= N_i x_{ij} \left[\frac{e^{x_i'\beta}}{(1 + e^{x_i'\beta})} - \left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}\right)^2\right] . \end{aligned}$$

Recalling equation (5), define

$$p_i \equiv m_{i1}/N_i = e^{x_i'\beta}\Big/(1 + e^{x_i'\beta})\,. \qquad (7)$$

For product-binomial sampling, p_i is the probability of an observation in the ith row occurring in the first column of the table. Substituting p_i into (6), the partial derivative is

$$\begin{aligned} \frac{\partial m_{i1}(\beta)}{\partial \beta_j} &= N_i x_{ij}(p_i - p_i^2) \\ &= N_i p_i (1 - p_i) x_{ij}\,. \end{aligned}$$

The matrix of partial derivatives can be written as

$$\begin{aligned} dm_1(\beta) &= \begin{bmatrix} N_1 p_1(1-p_1)x_{11} & \cdots & N_1 p_1(1-p_1)x_{1k} \\ \vdots & & \vdots \\ N_t p_t(1-p_t)x_{t1} & \cdots & N_t p_t(1-p_t)x_{tk} \end{bmatrix} \\ &= D(N_i p_i(1-p_i))X\,. \end{aligned}$$

The matrix of partial derivatives for $X'(n - m(\beta))$ is then

$$X' dm_1(\beta) = X' D(N_i p_i (1 - p_i)) X\,.$$

Note that $N_i p_i(1-p_i) = b_i$ from (11.1.2).

We can now apply the Newton-Raphson algorithm. Given a current estimate β_s, the next estimate is

$$\beta_{s+1} = \beta_s + \delta_s\,,$$

where

$$\delta_s = [X'D(N_i p_i(1-p_i))X]^{-1}X'(n_1 - m_1(\beta_s))$$

and p_i in $N_i p_i(1-p_i)$ is actually $p_i = p_i(\beta_s)$ as defined by equation (7).

As with log-linear models, the estimates can be found by doing a series of weighted regressions. Let $y_i = x_i'\beta_s + [n_{i1} - m_{i1}(\beta_s)]/N_i p_i(1-p_i)$, so that

$$Y = X\beta_s + D(N_i p_i(1-p_i))^{-1}(n_1 - m_1(\beta_s))\,.$$

If the weighted regression model

$$Y = X\beta + e\,,\ \mathrm{E}(e) = 0\,,\ \mathrm{Cov}(e) = D^{-1}(N_i p_i(1-p_i))$$

is fitted, then the estimate of β is

$$\begin{aligned}\beta_{s+1} &= [X'D(N_i p_i(1-p_i))X]^{-1}X'D(N_i p_i(1-p_i))Y \\ &= \beta_s + \delta_s\,.\end{aligned}$$

11.4 Weighted Least Squares for Logit Models

The methods introduced by Grizzle, Starmer, and Koch (1969) are actually quite general and can be applied to logit models as well as log-linear models, cf. Section 10.6. As before, asymptotic properties of the GSK method are often the same as maximum likelihood, but the small sample justification is less compelling. Moreover, for some small samples, the GSK method cannot be used at all unless ad hoc modifications to the data are introduced.

As with log-linear models, the GSK method amounts to performing one step of the Newton-Raphson algorithm. For a logit model

$$\eta = X\beta\,,$$

where

$$\begin{aligned}\eta &= [\mu_{11} - \mu_{21}, \ldots, \mu_{1t} - \mu_{2t}]' \\ &= [\log(p_{11}/p_{21}), \ldots, \log(p_{1t}/p_{2t})]'\,.\end{aligned}$$

The model

$$Y = X\beta + e\,,\ \mathrm{E}(e) = 0\,,\ \mathrm{Cov}(e) = D^{-1}(N_i\hat{p}_i(1-\hat{p}_i))$$

is fitted where

$$\hat{p}_i \equiv \hat{p}_{i1} = n_{i1}/N_i \tag{1}$$

and

$$Y = [y_1, \ldots, y_t]'$$

with

$$y_i = \log[\hat{p}_i/(1-\hat{p}_i)]\,.$$

The justification for the GSK procedure is asymptotic. The estimates obtained are asymptotically optimal if the N_i's are all large. As the justification for the GSK procedure is based on its large sample optimality, there is no apparent reason to use GSK for small samples. Moreover, it is not clear that the sort of asymptotic arguments which involve large numbers of small samples can be extended to the GSK approach.

Perhaps the most obvious difficulty with the GSK approach is that if $\hat{p}_i$ is either zero or one, y_i is not defined. When $N_i = 1$, y_i is always undefined. Of course, the corresponding weight from the inverse of the covariance matrix is also zero, so one could argue that GSK simply ignores such cases. But this could result in ignoring great quantities of data. For the Chapman data of Section 4.1, all of the data would be ignored. An ad hoc correction for this problem has been proposed, which is simply to substitute for any value $\hat{p}_i$ that is 0 or 1, the values $\hat{p}_i \pm \epsilon_i$, where the substitution forces $\hat{p}_i$ to be between 0 and 1 and ϵ_i is some small number. It is frequently suggested that any values $n_{ij} = 0$ be replaced by $n_{ij} = 0.5$.

It may be noted that the $\hat{p}_i$'s given in (1) are also the natural starting values for the Newton-Raphson algorithm and that small samples also require that $\hat{p}_i$'s of 0 or 1 be adjusted before they can be used. The key difference is that the justification for MLEs does not depend on properties of the starting values. Any starting values that lead to MLEs are perfectly acceptable. The GSK method depends crucially on the initial estimates.

The details of a GSK logit model analysis will not be given because, for uniformly large samples, they are exactly analogous to the log-linear model analysis given in Section 10.6. For large N_i's, the SSE can be used to give a chi-square test for lack of fit. The degrees of freedom for the test are the degrees of freedom for error. Models can be compared by comparing sums of squares for error. Reported standard errors must be corrected for the root mean square error; t statistics must be corrected for the root mean square error and compared to the standard normal distribution. The only difference is that a valid standard error exists for the intercept.

11.5 Multinomial Response Models

An integral point in the definition of logistic regression models is that the response variable has only two categories. The model posits a linear mean

structure for $\log(m_{i1}/m_{i2})$. As discussed in the previous chapter, if the response variable has more than two categories, it is by no means clear how to extend the logistic regression model to deal with the additional categories. It may then be somewhat surprising to find that it is clear how to extend the log-linear model version of the logistic regression model to more than two categories. We will discuss the appropriate log-linear model for a three-category response model. Extensions to responses with more than three categories follow the same pattern.

Before beginning with three-category responses, we reconsider the nature of the log-linear model for a two-category response. The model is

$$\mu = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{1}$$

This model lacks symmetry in the categories. The model matrix has an X submatrix corresponding to the first category of each response, but for the second category, the submatrix is zero. A model that treats the first and second categories on the same basis is the model

$$\mu = \begin{bmatrix} I & X & 0 \\ I & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \end{bmatrix}. \tag{2}$$

The important thing to note about model (2) is that it is equivalent to model (1). Any vector μ from model (1) can be obtained from (2) and vice versa. To see this, note that

$$C\left(\begin{bmatrix} I & X \\ I & 0 \end{bmatrix}\right) = C\left(\begin{bmatrix} I & X & 0 \\ I & 0 & X \end{bmatrix}\right).$$

Any two models with the same column space are equivalent models. Model (2) is a reparametrization of (1).

Given model (2), there is an obvious generalization to a three-category response. Simply take

$$\mu = \begin{bmatrix} I & X & 0 & 0 \\ I & 0 & X & 0 \\ I & 0 & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \tag{3}$$

where $\mu = (\mu_{11}, \ldots, \mu_{t1}, \mu_{12}, \ldots, \mu_{t2}, \mu_{13}, \ldots, \mu_{t3})'$. As model (2) is equivalent to model (1), it is easily seen that model (3) is equivalent to

$$\mu = \begin{bmatrix} I & X & 0 \\ I & 0 & X \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Writing

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}$$

gives model (3) as

$$\log(m_{ij}) = \mu_{ij} = \alpha_i + x_i'\gamma_j \tag{4}$$

where $i = 1, \ldots, t$ and $j = 1, 2, 3$. With this notation, it becomes a simple matter to consider models such as

$$\log(m_{ij}/m_{i\,j+1}) = x_i'(\gamma_j - \gamma_{j+1}), \qquad j = 1, \ldots, J-1,$$

or

$$\log(m_{ij}/m_{iJ}) = x_i'(\gamma_j - \gamma_J), \qquad j = 1, \ldots, J-1,$$

as were discussed in Chapter 4. Note that if X contains a column of ones, model (4) includes a term for column effects; thus, both the row and column margins of the $t \times 3$ table are fixed.

11.6 Asymptotic Results

We now consider asymptotic results for log-linear models contained in Haberman (1977). The results are quite general. In particular, we will show that they give the standard asymptotics for log-linear models and we will show how they apply to logistic regression. In Section 11.1, no explicit discussion was given concerning the nature of the asymptotic theory used to justify the results on estimation and testing. If the expected count in each cell approaches infinity, the usual asymptotic theory applies. Unfortunately, in regression settings, this is often not appropriate. For regression problems, a more reasonable approach is to allow additional observations to have distinct values of the predictor variables rather than having them occur at values of the predictor variables that have already occurred. These additional observations with new predictors constitute new cells in the table, so the table itself is getting larger. We need to think in terms of the convergence of a sequence of models. Haberman's results do this. They apply when fitting log-linear models to many situations in which there are few observations relative to the number of cells in the table. In particular, they satisfy our need for asymptotic results appropriate to logistic regression. Frequently, they do not apply when testing a log-linear model against the saturated model with data from a large sparse multinomial distribution. For asymptotic results that apply to this case, see Koehler (1986). Other results on large sparse multinomials are contained in Koehler and Larntz (1980), Simonoff (1983, 1985, 1986), and Zelterman (1987). The mathematics in this section are the most sophisticated in the book (with the possible exception of Chapter 12).

Before discussing Haberman's asymptotic results, we set notation for a fixed sample size problem. Consider a log-linear model

$$\mu = X\beta \tag{1}$$

where $\mu = \log(m)$. To deal with the sampling scheme, assume that $C(Z) \subset C(X)$. *We are interested in the asymptotic distribution of those estimable functions of β that do not depend on the parameters that are forced into the model to deal with the sampling scheme.* In other words, we are interested in functions $\gamma'\mu = \gamma'X\beta$ for which $\gamma'Z = 0$. Moreover, to avoid trivial cases, we assume that $\gamma'X \neq 0$. The MLE of μ is denoted $\hat{\mu}$. The asymptotic variance of $\gamma'\hat{\mu}$ will be related to the function

$$\sigma^2(\gamma'\hat{\mu}) = \gamma' A(m) D^{-1}(m)\gamma$$

where $A(m) = X(X'D(m)X)^{-1}X'D(m)$. This function can be estimated by

$$\hat{\sigma}^2(\gamma'\hat{\mu}) = \gamma' A(\hat{m}) D^{-1}(\hat{m})\gamma\,.$$

Also of interest are the asymptotic distributions of the Pearson test statistic X^2 and the likelihood ratio test statistic G^2 for testing model (1) against a reduced model

$$\mu = W\delta \tag{2}$$

where $C(Z) \subset C(W) \subset C(X)$. Let $r = \text{rank}(X) - \text{rank}(Z)$ and $s = \text{rank}(W) - \text{rank}(Z)$.

The asymptotic results require one more concept. Let

$$A_z = Z(Z'D(m)Z)^{-1}Z'D(m)$$

and let $\mathcal{N}(A_z)$ be the null space of A_z, i.e.,

$$\mathcal{N}(A_z) = \{x | A_z x = 0\}.$$

If we write a vector in $C(X)$ as $x = (x_1, \ldots, x_q)'$, define d to be

$$d = \sup\left\{|x_i|/\sqrt{x'D(m)x} \,:\, x \in \mathcal{N}(A_z) \cap C(X)\right\}. \tag{3}$$

To get asymptotic results, consider a sequence of log-linear models indexed by t. Thus, the log-linear models are

$$\mu_t = X_t\beta_t$$

where $\mu_t = \log(m_t)$ and $C(Z_t) \subset C(X_t)$. Our estimable functions of interest are $\gamma_t'\mu_t$, where $\gamma_t'Z_t = 0$. The MLE of $\gamma_t'\mu_t$ is $\gamma_t'\hat{\mu}_t$. Similarly,

$$\sigma^2(\gamma_t'\hat{\mu}_t) = \gamma_t' A_t(m_t) D^{-1}(m_t)\gamma_t$$

and

$$\hat{\sigma}^2(\gamma_t'\hat{\mu}_t) = \gamma_t' A_t(\hat{m}_t) D^{-1}(\hat{m}_t)\gamma_t\,.$$

The reduced model of interest in testing models is

$$\mu_t = W_t\delta_t$$

with $C(Z_t) \subset C(W_t) \subset C(X_t)$. The ranks are $r_t = \text{rank}(X_t) - \text{rank}(Z_t)$ and $s_t = \text{rank}(W_t) - \text{rank}(Z_t)$. Note that r_t is also the rank of $\mathcal{N}(A_{z_t}) \cap C(X_t)$, cf. Proposition 11.6.5. Finally,

$$d_t = \sup\left\{|x_i|/\sqrt{x'D(m_t)x} \,:\, x \in \mathcal{N}(A_{z_t}) \cap C(X_t)\right\}\,.$$

Obviously, the standard asymptotic results will not hold for all sequences of log-linear models. There must be some restrictions on the models. The restriction involves d_t.

Theorem 11.6.1. If $r_t d_t \to 0$ as $t \to \infty$, then

(a) $[\gamma_t'\hat{\mu}_t - \gamma_t'\mu_t]\Big/\sqrt{\sigma^2(\gamma_t'\hat{\mu}_t)} \xrightarrow{L} N(0,1)$,

(b) $\hat{\sigma}^2(\gamma_t'\hat{\mu}_t)/\sigma^2(\gamma_t'\hat{\mu}_t) \xrightarrow{P} 1$.

Corollary 11.6.2. If $r_t d_t \to 0$ as $t \to \infty$, then

$$[\gamma_t'\hat{\mu}_t - \gamma_t'\mu_t]\Big/\sqrt{\hat{\sigma}^2(\gamma_t'\hat{\mu}_t)} \xrightarrow{L} N(0,1)\,.$$

Let G^2 and X^2 be the likelihood ratio and Pearson test statistics for testing model (2) against model (1).

Theorem 11.6.3. If $r_t d_t \to 0$ and $r_t - s_t \to f$ as $t \to \infty$ and if $\mu_t \in C(W_t)$ for $t \geq 0$, then

(a) $G_t^2 \xrightarrow{L} \chi^2(f)$,

(b) $X_t^2 \xrightarrow{L} \chi^2(f)$,

(c) $G_t^2 - X_t^2 \xrightarrow{P} 0$.

These results imply the usual asymptotic results, cf. Chapter 12. In the usual results, replace the index t by the sample size N. For all N, we have $Z_N = Z$, $W_N = W$, $X_N = X$, and $\gamma_N = \gamma$. Moreover, $r_N = r$, $r_N - s_N = r - s$, and $m_N = Nm^*$, where m^* is defined as in Section 10.3.

With these adjustments, Theorems 1 and 3 give the standard results. For the theorems to apply, we need $r_N d_N \to 0$. Because r_N is fixed, this is simply the condition that $d_N \to 0$. To see that $d_N \to 0$, use the Cauchy-Schwartz inequality. Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ where the 1 is in the ith place. Thus, for any vector x, $x_i = e_i' x$.

Proposition 11.6.4. As $N \to \infty$, $d_N \to 0$.

Proof. By Cauchy-Schwartz, for $x \in C(X)$

$$\begin{aligned}(e_i'x)^2 &= ([e_i'D(1/\sqrt{m_N})][D(\sqrt{m_N})x])^2 \\ &\leq (e_i'D(1/m_N)e_i)(x'D(m_N)x)\,.\end{aligned}$$

By the definition of d_N,

$$\begin{aligned}d_N^2 &\leq \sup\{(e_i'x)^2/x'D(m_N)x \,:\, x \in C(X)\} \\ &\leq \max_i e_i'D(1/m_N)e_i \\ &= N^{-1}\max_i e_i'D(1/m^*)e_i\,.\end{aligned}$$

Because $\max_i e_i'D(1/m^*)e_i$ is a fixed positive constant, $d_N^2 \to 0$; thus, $d_N \to 0$. □

A primary use of these theorems is in their application to logistic regression models. In logistic regression, the model is $\eta = X\beta$, where $\eta = (\log(m_{11}/m_{21}), \ldots, \log(m_{1t}/m_{2t}))'$. The equivalent log-linear model is

$$\mu = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \equiv [Z, X_L] \begin{bmatrix} \alpha \\ \beta \end{bmatrix}\,.$$

As discussed in Section 1, rather than letting the m_{ij}'s get large (i.e., taking additional observations in existing cells), it is more realistic to let the number of cells get large while retaining the same basic structure in the logistic regression model. In other words, as the sample size increases, we add new rows to the design matrix while the expected counts in each cell are allowed to remain small.

Setting the notation for this case, μ_t is a $2t \times 1$ matrix,

$$\mu_t = \begin{bmatrix} I_t & X_t \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} \tag{4}$$

where I_t is a $t \times t$ identity matrix, 0 is a $t \times p$ matrix of zeros, and X_t is a $t \times p$ design matrix for the logistic regression model. We assume that for t large enough, the matrix X_t has $\text{rank}(X_t) = p$. For testing (4) against a reduced model, write the reduced model as

$$\mu_t = \begin{bmatrix} I_t & X_{0t} \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \eta_t \end{bmatrix} \tag{5}$$

where $C(X_{0t}) \subset C(X_t)$ and $\text{rank}(X_{0t}) = p_0$. Note that the full model is a log-linear model where the rank of the design matrix is $t + p$ and $r_t = (t+p) - t = p$. Similarly, for the reduced model, $s_t = (t+p_0) - t = p_0$.

As a practical matter, we never really have a sequence of models. We have one model. Thus, t is fixed (it is the number of binomial populations in the logistic regression). To apply Theorem 11.6.1, we need $r_t d_t \to 0$; thus, if rd is small for our model, we use the approximation

$$\frac{\gamma'\hat{\mu} - \gamma'\mu}{\sqrt{\hat{\sigma}^2(\gamma'\hat{\mu})}} \sim N(0,1)\,.$$

To apply Theorem 11.6.3, we need $r_t d_t \to 0$ and $(r_t - s_t) \to f$. In our current setup, $r_t - s_t = p - p_0$, which is fixed; so, again, if rd is small and the reduced model is adequate, we use the approximations

$$G^2 \sim \chi^2(p - p_0)$$

and

$$X^2 \sim \chi^2(p - p_0)\,.$$

It remains for us to get a handle on what conditions are necessary to have $r_t d_t \to 0$. Before doing that, we comment on why lack of fit tests often work poorly for logistic regression. The standard test for lack of fit of a logistic regression model is to test a model against the saturated log-linear model. To simplify notation, we will use model (5) as the model to be tested, but now, instead of testing model (5) against a model with similar structure like model (4), we test it against the saturated model. The saturated model can be written

$$\mu_t = \begin{bmatrix} I_t & I_t \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}\,.$$

The difference between the saturated model and model (4) is that X_t is replaced by I_t. Whereas $\text{rank}(X_t) = p$, we now have $\text{rank}(I_t) = t$. With the saturated model as the full model, we have $r_t = t$. For Theorem 11.6.3 to apply, we need $r_t d_t \to 0$ and, for some value f, $r_t - s_t \to f$. First, $r_t - s_t = t - p_0 \to \infty$, so there is no appropriate number of degrees of freedom for the asymptotic test. More importantly, because the condition $r_t d_t \to 0$ is the condition used to ensure that the model behaves well asymptotically and because the saturated model has $r_t d_t = t d_t$, for the saturated model to behave well asymptotically it must have $d_t \to 0$ very rapidly. Under standard sampling schemes, this does not happen and Theorem 11.6.3 does not apply. This is not to say that the lack of fit test will always work poorly. If the expected counts in each cell are large, we do not need to appeal to the sequence of models argument and, thus, the usual asymptotic results give an approximate χ^2 distribution for the lack of fit test. However, if expected cell counts are not large, there is no reason to believe that the asymptotic lack of fit test will work well and experience indicates that it does not work well.

A condition under which $r_t d_t \to 0$ for a sequence of logistic regression models is that the logistic regression leverages a_{ii} all approach zero while the b_i's do not, cf. equation (11.1.2). The remainder of this section is devoted to mathematical details associated with this demonstration. It requires a facility with linear models comparable to that developed in Christensen (1996b). The primary result is finding an explicit form for $(A - A_z)D^{-1}(m)$.

With $r_t = p$ for all t, it suffices to show that $d_t \to 0$. To do this, we will characterize d for an arbitrary logistic regression model.

The expected cell counts are $m = (m_{11}, m_{21}, \ldots, m_{t1}, m_{12}, \ldots, m_{t2})'$. Write $m_j = (m_{1j}, \ldots, m_{tj})$ for $j = 1, 2$ so that $m' = (m_1', m_2')$. For a logistic regression model, write

$$W \equiv X_* = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix},$$

so the symbol W is playing the role generally reserved for X in a log-linear model. Write $A = W(W'D(m)W)^{-1}W'D(m)$.

For logistic regression, the value of d is defined as the sup of a function of w over all w's in $\mathcal{N}(A_z) \cap C(W)$. First, we need to characterize $\mathcal{N}(A_z) \cap C(W)$.

Proposition 11.6.5. $\mathcal{N}(A_z) \cap C(W) = C(A - A_z)$.

Proof. A is a projection operator onto $C(W)$. In particular, for $w \in C(W)$, $Aw = w$ and $AA = A$. Similarly, A_z is a projection operator onto $C(Z)$. Moreover, $AA_z = A_z$.

If $w \in \mathcal{N}(A_z) \cap C(W)$, then $(A - A_z)w = Aw - A_z w = Aw = w$; thus, $w \in C(A - A_z)$.

If $w \in C(A - A_z)$, clearly $w \in C(W)$. We need to show that $w \in \mathcal{N}(A_z)$. The matrix $(A - A_z)$ is a projection operator, so $(A - A_z)w = w$ and, thus, $w = (A - A_z)w = Aw - A_z w = w - A_z w$, so $A_z w = 0$. □

We now examine the behavior of d. For $i = 1, \ldots, 2t$, let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ where the 1 is in the ith place. By (3),

$$d = \sup \left\{ \max_i |e_i' x| / \sqrt{x' D(m) x} \; : \; x \in C(A - A_z) \right\}.$$

Because $x \in C(A - A_z)$ can be written as $x = (A - A_z)b$ for some b, write

$$d = \sup \left\{ \max_i |e_i'(A - A_z)b| / \sqrt{b'(A - A_z)'D(m)(A - A_z)b} \; : \; \text{all } b \right\}.$$

Note that

$$|e_i'(A - A_z)b|$$

$$
\begin{aligned}
&= |e_i'(A - A_z)(A - A_z)b| \\
&= |[e_i'(A - A_z)D^{-1}(\sqrt{m})][D(\sqrt{m})(A - A_z)b]| \\
&\leq \sqrt{e_i'(A - A_z)D^{-1}(m)(A - A_z)'e_i}\sqrt{b'(A - A_z)'D(m)(A - A_z)b}
\end{aligned}
$$

where the inequality is just the Cauchy-Schwartz inequality. It follows that

$$d \leq \max_i \sqrt{e_i'(A - A_z)D^{-1}(m)(A - A_z)'e_i}\,.$$

Proposition 11.6.6. $(A - A_z)D^{-1}(m)(A - A_z)' = (A - A_z)D^{-1}(m)$.

Proof. Using the definitions of A and A_z, the fact that $(A - A_z)Z = 0$, and that $AZ = Z$, so $Z'A' = Z'$, we find

$$
\begin{aligned}
&(A - A_z)D^{-1}(m)(A - A_z)' \\
&\quad = (A - A_z)D^{-1}(m)D(m)[W(W'D(m)W)^{-1}W' - Z(Z'D(m)Z)^{-1}Z'] \\
&\quad = (A - A_z)[W(W'D(m)W)^{-1}W' - Z(Z'D(m)Z)^{-1}Z'] \\
&\quad = (A - A_z)[W(W'D(m)W)^{-1}W'] \\
&\quad = A[W(W'D(m)W)^{-1}W'] - A_z[W(W'D(m)W)^{-1}W'] \\
&\quad = [W(W'D(m)W)^{-1}W'] \\
&\qquad - Z(Z'D(m)Z)^{-1}Z'D(m)[W(W'D(m)W)^{-1}W'] \\
&\quad = AD^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'A' \\
&\quad = AD^{-1}(m) - Z(Z'D(m)Z)^{-1}Z' \\
&\quad = AD^{-1}(m) - A_zD^{-1}(m) \\
&\quad = (A - A_z)D^{-1}(m).
\end{aligned}
$$

□

Using Proposition 11.6.6, we now have

$$d \leq \max_i \sqrt{e_i'(A - A_z)D^{-1}(m)e_i}\,. \tag{6}$$

So far, we have not used the logistic regression structure of the log-linear model. We now use the special structure of logistic regression to further characterize inequality (6).

Proposition 11.6.7.

$$A_z = \begin{bmatrix} D(m_1)D^{-1}(m_1 + m_2) & D(m_2)D^{-1}(m_1 + m_2) \\ D(m_1)D^{-1}(m_1 + m_2) & D(m_2)D^{-1}(m_1 + m_2) \end{bmatrix}.$$

Proof. This follows immediately from Lemma 11.1.6, the definitions

of Z and A_z, and the fact that

$$D(m) = \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix}. \qquad \square$$

To simplify notation, for vectors $v = (v_1, \ldots, v_q)'$ and $u = (u_1, \ldots, u_q)'$, let $vu = (v_1u_1, \ldots, v_qu_q)'$ and $v/u = (v_1/u_1, \ldots, v_q/u_q)'$. As in (11.1.2), $b \equiv m_1m_2/(m_1 + m_2)$.

Proposition 11.6.8.

$$\begin{aligned} &A - A_z \\ &= \begin{bmatrix} D(m_2/(m_1+m_2))X \\ -D(m_1/(m_1+m_2))X \end{bmatrix} [X'D(b)X]^{-1} [X'D(b)\,,\ -X'D(b)]\,. \end{aligned}$$

Proof. The perpendicular projection operator onto $C(D(\sqrt{m})W)$ is $D(\sqrt{m})AD^{-1}(\sqrt{m}) \equiv M_w$ and the perpendicular projection operator onto $C\left(D(\sqrt{m})Z\right)$ is $D(\sqrt{m})A_zD^{-1}(\sqrt{m}) \equiv M_z$. It follows that

$$\begin{aligned} M_w - M_z &= D(\sqrt{m})AD^{-1}(\sqrt{m}) - D(\sqrt{m})A_zD^{-1}(\sqrt{m}) \\ &= D(\sqrt{m})(A - A_z)D^{-1}(\sqrt{m})\,. \end{aligned}$$

If we can find $M_w - M_z$, then $A - A_z = D^{-1}(\sqrt{m})(M_w - M_z)D(\sqrt{m})$.

The matrix $M_w - M_z$ is the perpendicular projection operator onto

$$C\left((I - M_z)D(\sqrt{m}) \begin{bmatrix} X \\ 0 \end{bmatrix}\right),$$

cf. Christensen (1996b, Sections 9.1, 9.2).

$$\begin{aligned} &(I - M_z)D(\sqrt{m}) \begin{bmatrix} X \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} D(\sqrt{m_1})X \\ 0 \end{bmatrix} - \begin{bmatrix} D(m_1\sqrt{m_1}/(m_1+m_2))X \\ D(m_1\sqrt{m_2}/(m_1+m_2))X \end{bmatrix} \\ &= \begin{bmatrix} D(\sqrt{m_1}m_2/(m_1+m_2))X \\ -D(\sqrt{m_2}m_1/(m_1+m_2))X \end{bmatrix}. \end{aligned}$$

Multiplying out to get the perpendicular projection operator and simplifying gives the result. $\square$

Finally, the main result is

Proposition 11.6.9.

$$\begin{aligned} (A - A_z)D^{-1}(m) = &\begin{bmatrix} D(m_2/(m_1+m_2))X \\ -D(m_1/(m_1+m_2))X \end{bmatrix} [X'D(b)X]^{-1} \\ &\times [X'D(m_2/(m_1+m_2))\,,\ -X'D(m_1/(m_1+m_2))] \end{aligned}$$

Proof. Multiply $A - A_z$ by $D^{-1}(m)$. □

We can now examine the exact nature of $e_i'(A - A_z)D^{-1}(m)e_i$. Write

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_t' \end{bmatrix};$$

then for $i = 1, \ldots, t$, let $j = i$ and

$$e_i'(A - A_z)D^{-1}(m)e_i = [m_{j2}/(m_{j1} + m_{j2})]^2 x_j'[X'D(b)X]^{-1}x_j\,,$$

for $i = t + 1, \ldots, 2t$, let $j = i - t$ and

$$e_i'(A - A_z)D^{-1}(m)e_i = [m_{j1}/(m_{j1} + m_{j2})]^2 x_j'[X'D(b)X]^{-1}x_j\,.$$

If these terms approach zero for a sequence of logistic regression models, then inequality (6) implies that Theorems 11.6.1 and 11.6.3 hold. In practice, if these terms are small for all i, then Theorems 11.6.1 and 11.6.3 should provide reasonable approximate distributions. The key (cf. Exercise 11.1) is that the X matrix needs to have the property that $x_i'[X'D(b)X]^{-1}x_i$ is small for all i.

This condition can also be related to linear model theory. If the elements of m_1 and m_2 are bounded above zero, then the terms

$$x_i'[X'D(b)X]^{-1}x_i$$

will converge to zero if and only if the terms $x_i'[X'X]^{-1}x_i$ converge to zero. The condition that the terms $x_i'[X'X]^{-1}x_i$ all converge to zero is known as Huber's condition. This is the condition assumed by Arnold (1981) to show that the usual distributions hold asymptotically for linear models with non-normal independent errors. Similarly, if Huber's condition holds for a logistic regression model, then the usual asymptotic results for logistic regression hold. Note that Huber's condition is sufficient to imply that asymptotic results hold; it is not a necessary condition.

EXERCISE 11.1. Show for logistic regression that $d_t \to 0$ as the maximum leverage goes to zero.

11.7 Discrimination, Allocation, and Retrospective Data

The Chapman data of Section 4.1 are the result of a *prospective* study; a large number of people were sampled and they were classified by whether

they had experienced a coronary incident and by their values on the variables age, systolic blood pressure, diastolic blood pressure, cholesterol, height, and weight. Only 26 of the 200 people had coronary incidents, so most of the information in the data is about people who did not have coronaries.

Retrospective studies are commonly used to examine events that are relatively rare, like coronary incidents. They address the problem of having a sample that contains few observations on the rare event. Consider a response variable with three levels: no incident, mild coronary incident, and severe coronary incident. One might take a sample of 125 people with no incidents, a sample of 40 people with mild incidents, and a sample of 35 people with severe coronary incidents. Thus, the sample sizes in the rare event categories are fixed by design. Once again, the case variables age, systolic blood pressure, diastolic blood pressure, cholesterol, height, and weight can be measured for each of the 200 individuals. When the response categories are also the populations sampled, it is easier to get substantial numbers of observations in each response category. Prospective and retrospective studies were discussed earlier at the beginning of Chapter 4.

While all of the case variables discussed above are really continuous, there are only a finite number of values that one could actually measure for any of the variables. For example, one often measures age in integer values of years and height in integer values of inches. Moreover, there are upper bounds on these values. Similar limitations based on the accuracy of the measuring instruments exist for all continuous variables. Thus, there are only a finite number of combinations of the case variables that can be considered. Call this very large but finite number S. The retrospective study described above yields a $3 \times S$ table in which each of the three rows is an independent multinomial sample. We wish to model these multinomials so that we can explain the data and, perhaps more importantly, predict the population into which a new case would fall when only the information on the case variables is available. The modeling problem can be thought of as *discriminating* among the three populations. The prediction problem is one of *allocating* new cases to the appropriate population.

Consider the allocation problem in more detail. Write the case variables as a vector $x_i = (Ag_i, S_i, D_i, Ch_i, H_i, W_i)'$, $i = 1, \ldots, S$. The value p_{hi} is the probability under population h of being in the category determined by x_i. Here, $h = 1, 2, 3$ and $i = 1, \ldots, S$. Write

$$f(x_i|h) = p_{hi}\,.$$

The function $f(x_i|h)$ is just the discrete density (probability mass function) of population h. The value of h is a parameter. Given a new case with known case variables x in one of the S observable categories, we can view $f(x|h)$ as a likelihood function. The maximum likelihood allocation rule assigns the new case x to the population h that has the largest value for the likelihood.

The only problem with this procedure is that the probabilities $f(x|h)$

are not known. They have to be estimated from the data. S is typically extremely large, so there are typically many more parameters (probabilities) to be estimated than observations with which to estimate them. Some sort of additional assumptions must be made in order to proceed.

One way to proceed is to abandon the fact that the observed data are discrete and assume a continuous density $f(x|h)$ as a model for the observations. If the underlying but unobservable case variables are continuous, this is extremely reasonable. In fact, it is so reasonable that people often overlook the fact that it is an approximation to what is properly a discrete distribution for the observations. The problem is not in approximating a discrete distribution with a continuous distribution but in finding a continuous distribution that provides an appropriate model. Traditionally, the case variables have been modeled with a multivariate normal distribution. For multivariate normals, estimating the mean vector and the covariance matrix for each population leads to natural estimates for the $f(x|h)$'s. This approach to discrimination and allocation is originally due to R. A. Fisher (1936). More recently, nonparametric density estimation has been used to model the distributions, cf. Seber (1984, Section 6.5).

Rather than invoking continuity, another way to proceed is to cut down the number of categories to a manageable size. Rather than using all S categories, one can restrict attention to the x values that were actually observed. In our hypothetical example, if all the x_i's are distinct, this restriction yields a 3×200 table. The x_i's are frequently distinct when any of the case variables are continuous, but if they are not distinct, it simply reduces the number of columns in the table. We will assume that the x_i's are distinct.

The original sampling scheme for the $3 \times S$ table was product-multinomial in the rows and the standard method of analysis, as illustrated in Section 4.7, also treats the 3×200 table as product-multinomial in the rows. Alas, restricting the table invalidates this product-multinomial sampling scheme for the reduced table. For example, under product-multinomial sampling, there is a positive probability of getting zeros for all three of the counts in a given column. We are using only the x_i's that are actually observed, so every column must have at least one observation in it.

There are two ways to analyze the data in the 3×200 table. The standard method illustrated in Section 4.7 is both a *partial likelihood* analysis and an *extended likelihood* analysis. Alternatively, one can perform a *conditional likelihood* analysis to obtain information on some parameters.

Partial likelihood analysis depends on having a likelihood that can be factored into the product of two terms. One term, the partial likelihood, must involve all of the parameters of interest and only those parameters. The second term involves only nuisance parameters. Without loss of generality, assume that the actual observations occur in the first 200 categories. For each population h, let $p_h = (p_{h1}, \ldots, p_{hS})'$ be the probability vector

and let $N_h \equiv n_{h\cdot}$ be the sample size. The likelihood is

$$\begin{aligned} L(p_1, p_2, p_3) &= \prod_{h=1}^{3} \prod_{i=1}^{S} p_{hi}^{n_{hi}} \\ &= \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \left\{ \prod_{h=1}^{3} \prod_{i=201}^{S} p_{hi}^{n_{hi}} \right\} \\ &= \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \end{aligned}$$

where the last equality holds because for $i > 200$, $n_{hi} = 0$ and any log-linear model has $p_{hi} > 0$ for all h and i. With all the zero counts, the likelihood cannot be maximized subject to the condition of positive cell probabilities. However, this function is also the partial likelihood involving only the parameters p_{hi}, $h = 1, 2, 3,\ i = 1, \ldots, 200$. As such, it can be maximized.

Technically, write

$$L(p_1, p_2, p_3) = \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \cdot \Gamma(p_{hi} : h = 1, 2, 3;\ i = 201, \ldots, S)$$

where

$$\Gamma(p_{hi} : h = 1, 2, 3;\ i = 201, \ldots, S) \equiv 1.$$

We have factorized the likelihood appropriately, so the partial likelihood for $p_{hi} : h = 1, 2, 3,\ i = 1, \ldots, 200$ is

$$\prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} .$$

Obviously, the log of the partial likelihood is

$$\sum_{h=1}^{3} \sum_{i=1}^{200} n_{hi} \log(p_{hi}) .$$

We now incorporate a log-linear model into the analysis. Assume a typical multinomial response model for the parameters

$$m_{hi} = N_h p_{hi}$$

that consists of

$$\log(m_{hi}) = \alpha_i + x_i' \gamma_h \tag{1}$$

for all h and i, cf. model (11.5.4). Dropping a constant that depends only on the N_h's, the log-likelihood becomes

$$\begin{aligned}\ell(\gamma_1,\gamma_2,\gamma_3,\alpha_i,i=1,\ldots,S) &= \sum_{h=1}^{3}\sum_{i=1}^{S} n_{hi}\log(m_{hi}) \\ &= \sum_{h=1}^{3}\sum_{i=1}^{S} n_{hi}\left[\alpha_i + x_i'\gamma_h\right] \\ &= \sum_{h=1}^{3}\sum_{i=1}^{200} n_{hi}\left[\alpha_i + x_i'\gamma_h\right]\end{aligned}$$

where, again, the last equality follows from the fact that $n_{hi} = 0$ for $i = 201,\ldots,S$. In the $3 \times S$ table, the row totals are fixed by the product-multinomial sampling scheme. The standard analysis of the 3×200 table also treats the rows as product-multinomials, so the row totals are fixed. Fixing the row totals requires the inclusion of main effects for rows in the log-linear model. These can be incorporated into the $x_i'\gamma_h$ terms. Write $x_i' = (1, x_{i2},\ldots,x_{ip})$, where the case variables are $x_{i2},\ldots,x_{ip}$. Then with $\gamma_h = (\gamma_{h1},\ldots,\gamma_{hp})'$, the intercept parameters γ_{h1} are the row main effects.

For a partial likelihood analysis, observe that the log-likelihood is the sum of two terms, one of which depends on the parameters of interest $\gamma_1,\gamma_2,\gamma_3,\alpha_i$, $i = 1,\ldots,200$, and another, which in this case is identically equal to zero, that depends only on $\alpha_i, i = 201,\ldots,S$. (The second function is identically zero, so it depends on anything we want it to depend on.) A partial likelihood analysis then obtains estimates of $\gamma_1,\gamma_2,\gamma_3,\alpha_i, i = 1,\ldots,200$, by maximizing the term that involves only those parameters. Of course, the only difference between the log partial likelihood and the log-likelihood is that the log partial likelihood is considered as a function of fewer parameters. In particular, the log partial likelihood is exactly the same as the log-likelihood for the 3×200 table under product-multinomial sampling. Thus, the MLEs from the standard analysis are maximum partial likelihood estimates for the full $3 \times S$ table.

Another productive way to use the log-likelihood is to consider *extended maximum likelihood estimates*, cf. Haberman (1974a, p. 402). An estimate $\hat{m}$ is an extended maximum likelihood estimate if the log-likelihood $\ell(m)$ converges to its supremum as m converges to $\hat{m}$. In this setup, the usual MLEs from the 3×200 table are extended MLEs. The log-likelihood function only depends on $\gamma_1,\gamma_2,\gamma_3,\alpha_i$, $i = 1,\ldots,200$, so the reduced table MLEs together with $\hat{p}_{hi} = 0$ for $h = 1,2,3$, $i = 201,\ldots,S$ maximize the full table log-likelihood function subject to the constraints $\hat{p}_{h\cdot} = 1$. The only problem is that log-linear models do not allow $\hat{p}_{hi} = 0$ for any h, i. Allowing extended MLEs removes the problem.

Whether the justification is partial likelihood or extended likelihood, we arrive at an analysis based on the MLEs for the 3×200 table with product-

multinomial sampling in the rows. Details of the analysis are given in the next subsection.

The *conditional likelihood analysis* simply defines the likelihood in terms of the conditional distribution of the 3×200 table given that these were the only 200 columns of the $3 \times S$ table that were observed. The conditional likelihood of the 3×200 table is

$$\prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \Big/ \sum_{r} \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{r_{hi}} \tag{2}$$

where the sum is over all $3 \times S$ tables of counts $r = (r_{11}, \ldots, r_{3S})'$ with

$$r_{h\cdot} = n_{h\cdot} = N_h, \qquad h = 1, 2, 3,$$

$$r_{\cdot i} = n_{\cdot i} = 1, \qquad i = 1, \ldots, 200,$$

$$r_{\cdot i} = 0, \qquad i = 201, \ldots, S.$$

It is not difficult to see that the conditional likelihood does not depend on the α_i's or the intercept terms γ_{h1}, cf. Exercise 11.8.4. Thus, any inferences that require estimates of these quantities cannot be made using the conditional likelihood approach. In particular, it will be seen in the next subsection that allocation of observations depends on the vectors γ_h, including the components γ_{h1}. Thus, *the conditional likelihood approach cannot be used for allocation.*

The key early paper on logistic discrimination was written by Anderson (1972). Anderson and Blair (1982) clarified several aspects of the theory and introduced another basis for analysis: penalized maximum likelihood. Some other relevant works are Farewell (1979), Prentice and Pyke (1979), and Breslow and Day (1980).

The Partial Likelihood Analysis

We have a vector of allocation variables $x_i' = (x_{i1}, \ldots, x_{ip})$ that are observed on each of t individuals. Thus, far in the section, we have always used $t = 200$, but the conclusions hold for any value of t. In addition to observing the x_i's, we know to which of the three populations each individual belongs. Our goal is to use the information on these t individuals in order to allocate future individuals into an appropriate population.

To do this, we set up a model similar to the standard multinomial response model. Let

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}.$$

Our data are

$$n = (n_{11}, \ldots, n_{1t}, n_{21}, \ldots, n_{2t}, n_{31}, \ldots, n_{3t})'$$

where $n_{hi} = 1$ if the ith case belongs to population h and $n_{hi} = 0$ otherwise. For a prospective multinomial response model, the $3 \times t$ table either is or can be considered to be the result of taking t independent trinomial samples and can be analyzed by standard logistic regression. With the retrospective sampling appropriate for discrimination problems, the sampling scheme is the result of taking three independent multinomial samples each with S categories where S is large and unknown. As discussed above, for the purpose of estimation the samples can be treated as three independent multinomials with t categories. *The standard prospective approach assumes that column totals of the $3 \times t$ table are fixed by the sampling scheme, whereas the retrospective (discrimination) approach assumes that the row totals are fixed by the sampling scheme.*

With a $3 \times t$ table in which the row totals are fixed, indicator variables must be included in the model to ensure that the estimated row totals equal the observed row totals. This is accomplished by requiring that the X matrix include a column of 1s (or its equivalent). In other words, for logistic models, the sampling scheme requires that models for discrimination data include intercepts. It is a common practice to include an intercept in multinomial response models, so the fact that an intercept is *required* for retrospective data is easily overlooked.

The log-linear model $\log(m_{hi}) = \alpha_i + x_i'\gamma_h$ for the $3 \times t$ table is written in matrix form as

$$\log(m) = \mu = \begin{bmatrix} I_t & X & 0 & 0 \\ I_t & 0 & X & 0 \\ I_t & 0 & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}.$$

This model is of exactly the same form as a standard multinomial response model and is fit in exactly the same way. The difference is in the interpretation of the underlying probabilities. In prospective sampling, the multinomial expected cell counts are $m_{hi} = n_{\cdot i} p_{hi}$ where $p_{\cdot i} = 1$. For retrospective sampling,

$$m_{hi} = n_{h\cdot} p_{hi} = N_h p_{hi}$$

where

$$p_{h\cdot} = 1\,.$$

In particular,

$$\log(p_{hi}) = \alpha_i + x_i'\gamma_h - \log(n_{h\cdot})\,.$$

The MLE of $\log(p_{hi})$, under the device of treating the sampling as product-multinomial in the rows of the $3 \times t$ table, is both the maximum partial likelihood estimate and an extended maximum likelihood estimate of $\log(p_{hi})$ in the full $3 \times S$ table.

The maximum likelihood allocation method applied to the observed data assigns case i to the population h with $\log(p_{hi}) = \max_k\{\log(p_{ki})\}$. The estimated maximum likelihood allocation method assigns i to the population

h with

$$\log(\hat{p}_{hi}) = \max_k\{\log(\hat{p}_{ki})\}. \tag{3}$$

Note that equation (3) is equivalent to

$$\hat{\alpha}_i + x_i'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{\hat{\alpha}_i + x_i'\hat{\gamma}_k - \log(n_{k\cdot})\}$$

which is equivalent to

$$x_i'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{x_i'\hat{\gamma}_k - \log(n_{k\cdot})\} \ . \tag{4}$$

Equation (4) does not depend on the $\hat{\alpha}_i$'s; thus, the allocation procedure depends on the individual only through the value of x_i.

Equation (4) can also be used as the basis for classifying new cases from an unknown population into one of the three possible populations. If the new case has observation vector x, the estimated maximum likelihood allocation rule is to classify the new case into population h if

$$x'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{x'\hat{\gamma}_k - \log(n_{k\cdot})\} \ .$$

This result depends on the fact that the allocation rule is really a function of the likelihood ratios of the various populations. The likelihood ratios depend only on x, the γ's, and the $\log(n_{h\cdot})$'s. All of these are either known or can be estimated. For a given value of x, the corresponding value of α does not enter into the analysis. Moreover, as illustrated in Section 4.7, if the likelihood ratios can be estimated, the posterior probabilities can also be estimated.

To evaluate how well the model discriminates between populations, check to see how often the cases in the data are allocated to the correct population. In other words, when case i is really in population h, see how often the probability that case i comes from population h is larger than the probabilities that case i comes from any of the other populations. Because the evaluation is carried out on the same data that generated the estimates of the p_{hi}'s, the results of the evaluation will be biased in favor of the discrimination method; i.e., the method will look better than it really is. See Section 4.7 for more discussion of this problem.

11.8 Exercises

EXERCISE 11.8.1. Analyze the data of Exercise 8.4.1 as a logistic regression with nodal involvement as the response. Include the investigation of higher-order interactions in your analysis. The original investigator was particularly interested in whether acid was a valuable predictor of nodal involvement.

EXERCISE 11.8.2. *Asymptotic Inference for the $LD(50)$.*
In Exercise 4.8.9, models and methods for estimating the $LD(50)$ were discussed. Use the delta method of Exercise 10.8.4 to obtain an asymptotic standard error for the $LD(50)$. Using the data of Exercise 4.8.11, give a 99% confidence interval for the $LD(50)$.

EXERCISE 11.8.3. *Fieller's Method for the $LD(50)$.*
Fieller's method is an alternative to the delta method for obtaining an asymptotic confidence interval for the $LD(50)$, cf. Exercise 11.8.2. Fieller's method is thought to be less sensitive to the high correlation that is typically present between $\hat{\alpha}$ and $\hat{\beta}$. From standard results, one can obtain the estimated asymptotic variance and covariance for $\hat{\alpha}$ and $\hat{\beta}$; thus, for any fixed but unknown value w, an asymptotic standard error for $\hat{\alpha} + \hat{\beta}w$ is readily available as a function of w. Denote this standard error by $\hat{\sigma}(w)$. If $\alpha + \beta w$ is some known value Q, a 99% confidence region for w can be obtained from

$$\begin{aligned} .99 &= \Pr\left(-2.5758 \leq \frac{(\hat{\alpha} + \hat{\beta}w) - Q}{\hat{\sigma}(w)} \leq 2.5758\right) \\ &= \Pr\left((\hat{\alpha} + \hat{\beta}w - Q)^2 - 2.5758^2\,\hat{\sigma}^2(w) \leq 0\right). \end{aligned}$$

The 99% confidence region consists of all values of w that satisfy

$$(\hat{\alpha} + \hat{\beta}w - Q)^2 - 2.5758^2\hat{\sigma}^2(w) \leq 0.$$

Show how to use the quadratic formula to find the end points of the region. Under what conditions is the confidence region an interval? What other possibilities exist? How does this result apply to estimating the $LD(50)$? Using the data of Exercise 4.6.11, give a 99% confidence interval for the $LD(50)$.

EXERCISE 11.8.4. Show that the conditional likelihood given by equation (11.7.1) and display (11.7.2) does not depend on the α_i's or the intercept terms γ_{h1}. Here, $x_i' = (1, x_{i2}, \ldots, x_{ik})$ and $\gamma = (\gamma_{h1}, \ldots, \gamma_{hk})'$.

EXERCISE 11.8.5. Use the delta method of Exercise 10.8.4, the logistic transform, and (11.1.3) to show that

$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})\hat{a}_{ii}/N_i}} \sim N(0,1)\,,$$

where $N_i = n_{i1} + n_{i2}$.