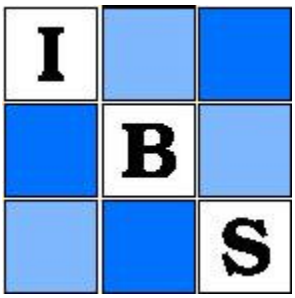

BIOMETRICS *30*, 589–600
December 1974

LOG-LINEAR MODELS FOR FREQUENCY TABLES WITH ORDERED CLASSIFICATIONS[1]

SHELBY J. HABERMAN

Department of Statistics,
University of Chicago, Chicago, Illinois 60637, U.S.A.
and
University of California, Berkeley, California

SUMMARY

Log-linear models are proposed for use with frequency tables with ordered classifications. Procedures are given for selection of models, determination of maximum likelihood (ML) equations, computation of ML estimates, and determination of asymptotic variances and tests.

1. INTRODUCTION

Contingency tables with ordered classifications frequently arise in statistical practice. Various methods for examining such tables are explored by Yates [1948], Armitage [1955], Mantel [1963], Kendall and Stuart ([1967] p.p. 562–78), and Williams and Grizzle [1972]. These techniques generally involve attempts to assign scores to the categories of the table and then analyze the data as if they were continuous. An alternate approach which exploits the actual distributional structure of these tables may be made by use of the general log-linear models discussed by Haberman [1973].

The present paper considers a class of log-linear models appropriate for the analysis of $r \times c$ contingency tables with ordered row and column categories. Methods are presented for selection of models, computation of ML estimates, performance of likelihood ratio or Pearson chi-square tests, and calculation of asymptotic variances. Although the discussion in this paper is confined to two-way contingency tables, no difficulty exists in applying the methods advocated here to m-way contingency tables.

2. DECOMPOSITION OF A TWO-WAY CONTINGENCY TABLE INTO ORTHOGONAL COMPONENTS

In this paper, an $r \times c$ contingency table $\mathbf{n}$ is considered, where $\mathbf{n} = \{n_{ij} : i = 1, \cdots, r, j = 1, \cdots, c\}$ has an expected value $\mathbf{m} = \{m_{ij}\}$ such that $m_{ij} > 0$, $i = 1, \cdots, r$, $j = 1, \cdots, c$. It is assumed that either (A) the cells of the table are independent Poisson random variables, (B) the entire table is a single multinomial sample, (C) each row is an independent multinomial sample, or (D) each column is an independent multinomial sample. To describe the relationship between the variables represented by the rows and columns of the table, the log-mean vector $\boldsymbol{\mu} = \{\log m_{ij}\}$ is decomposed into linear, quadratic, and

[1] This research was carried out in the Department of Statistics, University of Chicago, under partial support by Research Grant No. NSF GS 31967X from the Division of the Social Sciences of the National Science Foundation.

higher order components just as the expected values of the observations in a two-way analysis of variance are decomposed into such components (see Cochran and Cox [1957] p.p. 164–7).

In this decomposition, scores u_i , $i = 1, \cdots, r$, and v_j , $j = 1, \cdots, c$, are assigned to the respective categories of the row and column variables of the table. Orthogonal polynomials f_i , $i = 1, \cdots, r$, for the row scores are constructed as in Kendall and Stuart ([1967] p.p. 356–61) so that f_i has degree $i - 1$, and the vectors $\mathbf{x}^{(i)} = \{f_i(u_k): k = 1, \cdots, r\}$, $i = 1, \cdots, r$, are an orthonormal basis of R^r. Similarly, orthogonal polynomials g_j , $j = 1, \cdots, c$, for the scores v_j , $j = 1, \cdots, c$, are constructed so that g_j has degree $j - 1$ and the vectors $\mathbf{y}^{(j)} = \{g_j(v_l): l = 1, \cdots, c\}$, $j = 1, \cdots, c$, are an orthonormal basis of R^c. If the scores are evenly spaced, these vectors may be found by use of tables in Fisher and Yates [1963]. Given the vectors $\mathbf{x}^{(i)}$, $i = 1, \cdots, r$, and $\mathbf{y}^{(j)}$, $j = 1, \cdots, c$, the vector $\boldsymbol{\mu}$ may be uniquely written as

$$\boldsymbol{\mu} = \sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij} \mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}, \tag{1}$$

where the direct product of vectors $\mathbf{x} \in R^r$ and $\mathbf{y} \in R^c$ is defined by the equation

$$\mathbf{x} \otimes \mathbf{y} = \{x_i y_j\}$$

(see Halmos [1958] p. 174). The vectors $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $i = 1, \cdots, r$, $j = 1 \cdots, c$, are an orthonormal basis of the space W of $r \times c$ arrays. In (1), the coefficients α_{i1} , $i = 2, \cdots, r$, are the row effects, with α_{21} the linear row effect, α_{31} the quadratic row effect, etc. Similarly, the coefficients α_{1j} , $j = 2, \cdots, c$, are the column effects. The interaction effects are the coefficients α_{ij} , $i = 2, \cdots, r$, $j = 2, \cdots, c$, with α_{22} the linear by linear component of the interaction. A decomposition based on the logarithms of the expected cell counts is chosen since independence of the row and column variables is equivalent to the condition that the interaction coefficients $\alpha_{ij} = 0$, $i = 2, \cdots, r$, $j = 2, \cdots, c$; and since the categories of the column variable are equally probably given any category of the row variable if and only if $\alpha_{ij} = 0$, $i = 2, \cdots, r$, $j = 2, \cdots, c$ (see Bishop [1969] and Goodman [1970]). In addition, if $r > 2$ and $c = 2$, then the logit model

$$\log (m_{i1}/m_{i2}) = \beta_0 + \beta_1 u_i \quad i = 1, \cdots, r$$

for some β_0 and β_1 holds if and only if $\alpha_{i2} = 0$, $i = 3, \cdots, r$.

In the log-linear models examined in this paper, the restriction is imposed that $\alpha_{ij} = 0$ for $(i, j) \in S$, where S is a subset of $\{(i, j): 1 \leq i \leq r, 1 \leq j \leq c\}$ such that the following conditions are satisfied:

(E) If $(i, j) \in S$, $k \geq i$, $l \geq j$, then $(k, l) \in S$.

(F) The sets S and N are disjoint, where $N = \{(1, 1)\}$ under sampling conditions (A) and (B), $N = \{(i, 1): 1 \leq i \leq r\}$ under condition (C), and $N = \{(1, j): 1 \leq j \leq c\}$ under condition (D).

Condition (E) imposes the restriction that whenever a lower-order effect is assumed 0, then all higher-order effects are assumed 0. It is analogous to the hierarchy requirement of Bishop [1969] and Goodman [1970]. Condition (F) is needed to permit the same methods of analysis to be applied with any of the sampling schemes considered (see Haberman [1973]). The class of models considered in this paper is large enough to include both the independence and equiprobability models of Bishop [1969] and Goodman [1970] and models such as logit models which make explicit use of the ordering of the categories of the row and column variables.

3. MAXIMUM LIKELIHOOD ESTIMATION

For models considered in this paper, the log-mean vector $\boldsymbol{\mu}$ or the vector of means $\mathbf{m}$ is readily estimated by means of maximum likelihood, since the respective ML estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{m}}$ are easily calculated by the Newton–Raphson algorithm and have standard asymptotic properties. Results of this section follow from Haberman's [1974] more general analysis of log-linear models, so proofs of assertions are omitted. A more detailed description of results used is given in the Appendix.

Two equivalent computational procedures are available. In the first procedure, let $T = \{(i, j): 1 \leqq i \leqq r, 1 \leqq j \leqq c, (i, j) \notin S\}$ have f elements. Let $\boldsymbol{\mu}^{(0)} = \{\log n_{ij}\}$ and $\boldsymbol{\alpha}^{(\nu+1)} = \{\alpha_{ij}{}^{(\nu+1)}: (i, j) \in T\}$, $\nu \geqq 0$. Suppose that M is the rc by f matrix with columns $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $(i, j) \in T$, and $D(\mathbf{m})$ is the rc by rc diagonal matrix with diagonal elements m_{ij}, $1 \leqq i \leqq r$, $1 \leqq j \leqq r$. The Newton–Raphson algorithm generates successive approximations $\boldsymbol{\mu}^{(\nu)}$ for $\boldsymbol{\mu}$, where $\nu \geqq 1$, by means of the equations

$$m_{ij}{}^{(\nu)} = \exp(\mu_{ij}{}^{(\nu)}), \tag{2}$$

$$\tilde{\mu}_{ij}{}^{(\nu)} = \mu_{ij}{}^{(\nu)} + (n_{ij} - m_{ij}{}^{(\nu)})/m_{ij}{}^{(\nu)}, \tag{3}$$

$$\boldsymbol{\alpha}^{(\nu+1)} = [M'D(\mathbf{m}^{(\nu)})M]^{-1}M'D(\mathbf{m}^{(\nu)})\tilde{\boldsymbol{\mu}}^{(\nu)}, \tag{4}$$

$$\boldsymbol{\mu}^{(\nu+1)} = M\boldsymbol{\alpha}^{(\nu+1)}, \tag{5}$$

where M' is the transpose of M, and $\mathbf{m}^{(\nu)} = \{m_{ij}{}^{(\nu)}\}$.

If $(i, j) \in T - N$, then the asymptotic variance of $\hat{\alpha}_{ij}$, the ML estimate of α_{ij}, may be estimated by the (i, j)th diagonal element of $[M'D(\hat{\mathbf{m}})M]^{-1}$.

Alternatively, one may suppose that S has d elements and M_0 is the rc by d matrix with columns $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $(i, j) \in S$. One then defines $m_{ij}{}^{(\nu)}$ by (2) and $\tilde{\mu}_{ij}{}^{(\nu)}$ by (3) but lets

$$\mathbf{b}^{(\nu+1)} = [M_0'\{D(\mathbf{m}^{(\nu)})\}^{-1}M_0]^{-1}M_0'\tilde{\boldsymbol{\mu}}^{(\nu)} \tag{6}$$

and

$$\boldsymbol{\mu}^{(\nu+1)} = \tilde{\boldsymbol{\mu}}^{(\nu)} - \{D(\mathbf{m}^{(\nu)})\}^{-1}M_0\mathbf{b}^{(\nu+1)}. \tag{7}$$

If $\hat{\boldsymbol{\alpha}} = \{\hat{\alpha}_{ij} : (i, j) \in T\}$ is desired, then the relationship $\hat{\boldsymbol{\alpha}} = M'\hat{\boldsymbol{\mu}}$ is used. If $(i, j) \in T - N$, the estimated asymptotic variance of $\hat{\alpha}_{ij}$ is the (i, j)th diagonal element of

$$M'\{D(\hat{\mathbf{m}})\}^{-1}M - [M'\{D(\hat{\mathbf{m}})\}^{-1}M_0][M_0'\{D(\hat{\mathbf{m}})\}^{-1}M_0]^{-1}[M'\{D(\hat{\mathbf{m}})\}^{-1}M_0]'.$$

Given a ML estimate $\hat{\mathbf{m}}$ for the model $\alpha_{ij} = 0$ for $(i, j) \in S$, the model may be tested by use of the likelihood ratio chi-square statistic

$$L^2(S) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log(n_{ij}/\hat{m}_{ij}) \tag{8}$$

or the Pearson chi-square statistic

$$X^2(S) = \sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij} - \hat{m}_{ij})^2/\hat{m}_{ij}\ . \tag{9}$$

In both cases, the number of degrees of freedom is d, the number of elements of S.

4. SELECTION OF MODELS

In most applications in which models of the form $\alpha_{ij} = 0$ for $(i, j) \in S$ are considered, there are many possible choices of S. The most attractive choices are those which use the

fewest parameters α_{ij} , $(i, j) \in T$, to successfully fit the data. To find these models, two simultaneous test procedures are developed to eliminate models which do not fit the data, and an index of goodness of fit is proposed to compare models with different numbers $d(S)$ of elements in S.

The first test procedure uses the partition of $L^2(S)$ of Goodman [1970]. In this case, models H_k that $\alpha_{ij} = 0$ for $(i, j) \in S_k$, $k = 1, \cdots, q$, are considered such that $S_1 \subset \cdots \subset S_q$. It is assumed known that $\alpha_{ij} = 0$ for $(i, j) \in S_1$. For $k = 1, \cdots, q - 1$, the statistic $L^2(S_{k+1}) - L^2(S_k)$ is then the likelihood ratio chi-square for the null hypothesis H_{k+1} and the alternative H_k . As Haberman [1974] notes, if $H_{k'}$ is true, then the statistics $L^2(S_{k+1}) - L^2(S_k)$, $k = 1, \cdots, k' - 1$, are asymptotically independent with respective asymptotic chi-square distributions on $d(S_{k+1}) - d(S_k)$ degrees of freedom (D.F.). If a significance level γ is chosen and

$$\gamma' = 1 - (1 - \gamma)^{1/(q-1)},$$

then the probability that $L^2(S_{k+1}) - L^2(S_k)$, $k = 1, \cdots, k' - 1$, exceeds C_k , the upper γ' point for the chi-square distribution with $d(S_{k+1}) - d(S_k)$ D.F., is asymptotically no greater than γ. Thus a simultaneous test for these hypotheses is to reject all hypotheses H_k such that $k > k''$, where k'' is the smallest k such that

$$L^2(S_{k+1}) - L^2(S_k) > C_k . \tag{10}$$

If (10) does not hold for any $k < l$, then $k'' = l$. The asymptotic probability that a hypothesis is falsely rejected does not exceed γ. Given the results of this test, one may also reject any hypothesis that $\alpha_{ij} = 0$ for $(i, j) \in S$ for which $S_{k''+1}$ is contained in S.

The second test procedure is based on the direct estimation method of Goodman [1970]. In this technique, a set V is given such that for any set S of interest, $V \subset S$. A subset U of the pairs $\{(i, j): 1 \leq i \leq r, 1 \leq j \leq c\}$ is then chosen so that $U \cap (V \cup N) = \emptyset$. The ML estimates $\hat{\alpha}_{ij}$ are computed under the hypothesis that $\alpha_{ij} = 0$ for $(i, j) \in V$, together with the corresponding asymptotic standard deviations s_{ij} . If U has q elements, then significance level γ is assigned and the set X is found, where X is the set of indices in U such that the standardized value $\hat{\alpha}_{ij}/s_{ij}$ satisfies

$$|\hat{\alpha}_{ij}/s_{ij}| > Z \tag{11}$$

and Z is the upper $\gamma/\{2q\}$ point of the standardized normal distribution. The asymptotic probability that some $(i, j) \in X$ but $\alpha_{ij} \neq 0$ does not exceed γ. Given this result, one may reject any model that $\alpha_{ij} = 0$ for $(i, j) \in S$ if $X \not\subset S$. This procedure is particularly easily applied if V is empty, for then

$$\hat{\mu}_{ij} = \log n_{ij} \qquad i = 1, \cdots, r, \quad j = 1, \cdots, c, \tag{12}$$

$$\hat{\alpha}_{ij} = \sum_{k=1}^{r} x_k^{(i)} \sum_{l=1}^{c} y_l^{(j)} \log n_{kl} \qquad (i, j) \in U, \tag{13}$$

and

$$s_{ij}^2 = \sum_{k=1}^{r} \{x_k^{(i)}\}^2 \sum_{l=1}^{c} [\{y_l^{(j)}\}^2/n_{kl}] \qquad (i, j) \in U. \tag{14}$$

Each of these simultaneous test procedures may be used to eliminate models from consideration; however, they provide no criteria for choices among the remaining models. One possible approach is to note that if $\alpha_{ij} = 0$ for $(i, j) \in S$, then $L^2(S)/d(S)$ has an asymp-

totic distribution which has a mean of 1 and a variance of $2/d(S)$. Thus a possible index for comparison of models is $L^2(S)/d(S)$. Models are particularly attractive when $d(S)$ is large and this ratio is near 1.

5. EXAMPLES

To test the procedures considered in this paper, two tables are analyzed in this section. The first table (Table 1), which is found in Wing [1962], compares frequency of visits with length of stay of 132 long-term schizophrenic patients in two London mental hospitals. The three categories in each variable may be assigned the equal spaces scores $u_1 = v_1 = 1$, $u_2 = v_2 = 0$, and $u_3 = v_3 = -1$. If one makes inferences conditional on the observed sample size, the table may be regarded as a single multinomial sample of size 132. The simultaneous test procedures of section 4 may be applied to the class of sets S such that $(1, 1) \notin S$ and the hierarchy assumption holds that $(i, j) \in S$ implies $(i', j') \in S$ for $i' \geq i$ and $j' \geq j$. Since the table has relatively few observations and the number of possible models is fairly large, the significance level selected for this analysis should be relatively high. In this case, 20% will be used.

The partition of L^2 may be applied with $S_1 = \emptyset$, $S_2 = \{(3, 3)\}$, $S_3 = \{(2, 3), (3, 2), (3, 3)\}$, and $S_4 = \{(2, 2), (2, 3), (3, 2), (3, 3)\}$. Here S_1 corresponds to the unrestricted model, S_2 corresponds to the model that the quadratic by quadratic interaction is 0, S_3 corresponds to the model that the dependence between frequency of visits and length of stay may be described by a single linear by linear interaction, and S_4 corresponds to the hypothesis that frequency of visits and length of stay are independent. Given $\gamma = 0.20$, the critical points used in (21) are $C_1 = 3.3$, $C_2 = 5.3$, and $C_3 = 3.3$. As may be seen by examination of Table 2, the hypotheses of independence (S_4) and of linear by linear interaction (S_3) are rejected. Given that S_3 is rejected, it is only necessary to consider S_1, S_2, $\{(2, 3), (3, 3)\}$, $\{(3, 2), (3, 3)\}$, $\{(1, 3), (2, 3), (3, 3)\}$, and $\{(3, 1), (3, 2), (3, 3)\}$ as possible models.

Direct estimation may be used with V empty and U equal to the set of pairs (i, j) such that $1 \leq i \leq 3$, $1 \leq j \leq 3$, and i or j is not 1. In this case, the critical point Z in (11)

TABLE 1

VISITING BY LENGTH OF STAY IN HOSPITAL

Frequency of visiting	Length of stay in hospital		
	At least 2 years but less than 10 years	At least 10 years but less than 20 years	At least 20 years
Goes home, or visited regularly	43	16	3
Visited less than once a month. Does not go home	6	11	10
Never visited and never goes home	9	18	16

TABLE 2

CHI-SQUARE STATISTICS FOR TABLE 1

S	d(S)	$X^2(S)$	$L^2(S)$	$L^2(S)/d(S)$
{(2,2), (2,3), (3,2), (3,3)}	4	35.17	38.35	9.59
{(1,3), (2,3), (3,1) (3,2), (3,3)}	5	13.19	14.71	2.94
{(2,3), (3,1), (3,2), (3,3)}	4	10.80	11.88	2.97
{(1,3), (2,3), (3,2), (3,3)}	4	8.97	9.48	2.37
{(3,1), (3,2), (3,3)}	3	10.41	11.12	3.71
{(2,3), (3,2), (3,3)}	3	7.31	7.12	2.37
{(1,3), (2,3), (3,3)}	3	3.26	3.20	1.07
{(3,2), (3,3)}	2	6.52	6.46	3.23
{(2,3), (3,3)}	2	0.02	0.02	0.01
{(3,3)}	1	0.00	0.00	0.00

is 2.24. As may be seen in Table 3, the only models not rejected are those with S equal to S_1 , S_2 , {(2, 3), (3, 3)}, or {(1, 3), (2, 3), (3, 3)}. If the $L^2(S)/d(S)$ criterion is used, then the most attractive models are those with $S \subset \{(1, 3), (2, 3), (3, 3)\}$.

Of the simultaneous testing procedures considered here, the direct estimation procedure has been most successful in terms of rejecting models with large test statistics without ignoring models with small test statistics. To determine whether this gain has practical importance, it is helpful to consider what information the model $\alpha_{13} = \alpha_{23} = \alpha_{33} = 0$ provides concerning the data. As may be seen by comparison of Tables 1 and 4, the fit obtained by use of this model is quite good. To interpret the results, note that the model implies that $\underset{\sim}{\mu}$ satisfies

$$\mu_{ij} = a_i + b_i v_j \quad j = 1, \cdots, 3$$

for some a_i and b_i , $i = 1, \cdots, 3$. Hence, given the frequency of visits, μ_{ij} is a linear function of the length of stay. Given that $\alpha_{13} = \alpha_{23} = \alpha_{33} = 0$, the ML estimates of b_i may be computed from Table 4 by the relationship $\hat{b}_i = \hat{\mu}_{i1} - \hat{\mu}_{i2} = \log(\hat{m}_{i1}/\hat{m}_{i2})$, $1 \leqq i \leqq 3$. The respective slope estimates b_1 , b_2 , and b_3 are 1.178, -0.224, and -0.247. These estimates partially support Wing's conclusion that frequency of visits tends to be lower in longer term patients; however, it appears that the contrast is between frequently visited

TABLE 3

ESTIMATES OF α_{ij} FOR THE UNRESTRICTED MODEL FOR TABLE 1

Effect	Estimate	Standard Deviation	Standardized Value
α_{12}	-0.543	0.332	-1.637
α_{13}	-0.437	0.292	-1.493
α_{21}	-0.896	0.358	-2.499
α_{22}	1.619	0.364	4.447
α_{23}	0.036	0.289	0.125
α_{31}	0.038	0.326	0.118
α_{32}	0.898	0.365	2.460
α_{33}	-0.016	0.313	0.050

TABLE 4

MAXIMUM LIKELIHOOD ESTIMATES OF EXPECTED FREQUENCIES FOR DATA IN TABLE 1 FOR $\alpha_{13} = \alpha_{23} = \alpha_{33} = 0$

Frequency of visiting	Length of stay in hospital		
	At least 2 years but less than 10 years	At least 10 years but less than 20 years	At least 20 years
Goes home, or visited regularly	44.19	13.61	4.19
Visited less than once a month. Does not go home	7.07	8.85	11.07
Never visited and never goes home	10.98	14.05	17.98

patients and infrequently or never visited patients. These slope estimates do not progress in a linear fashion as one proceeds from the frequently visited category to the never visited category. More formally, it should be noted that $b_1 - b_3$, which is α_{22}, has a ML estimate of 1.424 with estimated standard deviation 0.069. The nonlinearity of the slopes is reflected in the estimate of $\alpha_{32} = (b_1 - 2b_2 + b_3)/3^{1/2}$ of 0.796. The estimated standard deviation of α_{32} is 0.101. Thus, a more precise notion of the relationship between the variables is obtained here than was obtained by the simple chi-square test for independence used by Wing [1962].

In the second example, the results shown in Table 5 are analyzed. In this table, the respondents in the Midtown mental health study described by Srole *et al.* [1962] are classified in terms of their mental health and in terms of their socioeconomic status. To test whether the prevalence of mental health problems tends to decrease as socioeconomic status increases, the mental health categories are assigned the scores 3, 1, -1, and -3 and the status categories are assigned the scores 5, 3, 1, -1, -3, and -5. The principal models of interest are the trivial model with $S_1 = \emptyset$, the model with only a linear by linear interaction with $S_2 = \{(i, j): i = 1, \cdots, 4, j = 1, \cdots, 6, i > 2 \text{ or } j > 2\}$, and the independence model with $S_3 = \{(i, j): i = 1, \cdots, 4, j = 1, \cdots, 6, i > 1 \text{ or } j > 1\}$. Other models satisfying the hierarchy principle may be of interest if the model $\alpha_{ij} = 0$ for $(i, j) \in S_2$ fails to fit the data.

TABLE 5

MENTAL HEALTH STATUS BY PARENTAL SOCIOECONOMIC STATUS

Mental health category	Parental socioeconomic status stratum					
	A	B	C	D	E	F
Well	64	57	57	72	36	21
Mild symptom formation	94	94	105	141	97	71
Moderate symptom formation	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

TABLE 6

CHI-SQUARE STATISTICS FOR TABLE 5

S	d(S)	$X^2(S)$	$L^2(S)$	$L^2(S)/d(S)$
S_2	14	9.73	9.90	0.71
S_3	15	45.99	47.43	3.16

The results of the analysis are summarized in Tables 6, 7, and 8. If one lets $\gamma = 0.20$ and uses the partition of L^2 with S_1 , S_2 , and S_3 , then the critical values are $C_1 = 21.0$ and $C_2 = 2.6$. Thus the independence model is rejected and the remaining models are accepted. This conclusion supports the analysis in the Midtown study. If direct estimation is used to examine all possible hierarchical models, then the critical point Z is 2.62. By this criterion, the only set including S_2 which can be considered is $S_2 \cup \{(1, 6)\}$. Since the model corresponding to this set provides little more insight than does the model corresponding to S_2 and since $L^2(S_2)/d(S_2) < 1$, there appears little reason to prefer any other model to the linear by linear interaction model.

Given the model that $\alpha_{ij} = 0$ for $(i, j) \in S_2$, the relationship between socioeconomic status and mental health category is completely described by the coefficient α_{22} . The ML estimate $\hat{\alpha}_{22}$ is 0.848 and has an estimated asymptotic standard deviation of 0.140. To interpret this coefficient, observe that under the proposed model, the log cross-product ratio

$$\log\left(\frac{m_{ij}m_{kl}}{m_{il}m_{kj}}\right) = \mu_{ij} - \mu_{il} - \mu_{kj} + \mu_{kl}$$

$$= \alpha_{22}(u_i - u_k)(v_j - v_l).$$

Thus, the interaction between categories i and k of the mental health variable and categories j and l of the socioeconomic status variable is a product of α_{22} , the distance $u_i - u_k$ between

TABLE 7

ESTIMATES OF α_{ij} FOR THE UNRESTRICTED MODEL FOR TABLE 5

Effect	Estimate	Standard Deviation	Standardized Value
α_{12}	0.197	0.138	1.423
α_{13}	-0.673	0.131	-5.125
α_{14}	-0.426	0.132	-3.224
α_{15}	0.378	0.128	2.947
α_{16}	0.291	0.118	2.463
α_{21}	-0.173	0.139	-1.240
α_{22}	0.931	0.150	6.195
α_{23}	-0.243	0.142	-1.710
α_{24}	0.021	0.142	0.149
α_{25}	-0.055	0.137	-0.400
α_{26}	0.094	0.125	0.756
α_{31}	-0.847	0.130	-6.523
α_{32}	0.053	0.138	0.385
α_{33}	-0.094	0.131	-0.717
α_{34}	-0.140	0.132	-1.059
α_{35}	0.036	0.128	0.282
α_{36}	0.031	0.118	0.259
α_{41}	-0.965	0.119	-8.077
α_{42}	0.237	0.125	1.889
α_{43}	0.057	0.120	-0.476
α_{44}	0.105	0.122	0.866
α_{45}	0.051	0.119	0.432
α_{46}	-0.010	0.111	-0.092

TABLE 8

ESTIMATED EXPECTED CELL FREQUENCIES FOR DATA IN TABLE 5 FOR $\alpha_{ij} = 0, 1 \leq i \leq 4, 1 \leq j \leq 6, i \text{ OR } j > 2$

Mental health category	Parental socioeconomic status stratum					
	A	B	C	D	E	F
Well	65.29	54.21	55.91	65.28	38.96	27.35
Mild symptom formation	104.42	94.94	107.20	137.04	89.56	68.84
Moderate symptom formation	50.15	49.92	61.72	86.39	61.81	52.02
Impaired	42.14	45.93	62.18	95.29	74.66	68.80

the mental health category scores, and the distance $v_i - v_l$ between the socioeconomic status scores. Again, much more information can be obtained here than can be obtained with a simple chi-square test of independence.

MODÈLES LOG-LINÉAIRES POUR DES TABLEAUX DE FRÉQUENCE DONT LES CLASSES SONT ORDONNÉES

RESUME

Des modèles log-linéaires sont présentés pour le cas de tableaux de fréquence dont les classes sont ordonnées. On donne des procédures pour le choix des modèles, la détermination des équations du maximum de vraisemblance, le calcul des estimateurs M. V., la détermination des variances asymptotiques et des tests.

REFERENCES

Armitage, P. [1955]. Tests for linear trends in proportions and frequencies. *Biometrics 11*, 375–86.

Bishop, Y. M. M. [1969]. Full contingency tables, logits, and split contingency tables. *Biometrics 25*, 383–400.

Cochran, W. G. and Cox, G. M. [1957]. *Experimental Design* (2nd Edn.). Wiley, New York.

Fisher, R. A. and Yates, F. [1963]. *Statistical Tables for Biological, Agricultural, and Medical Research* (6th Edn.). Hafner, New York.

Goodman, L. A. [1970]. The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Ass. 65*, 226–56.

Haberman, S. J. [1974]. *The Analysis of Frequency Data.* University of Chicago Press, Chicago (in press).

Halmos, P. [1958]. *Finite-dimensional Vector Spaces.* Van Nostrand, Princeton, New Jersey.

Kendall, M. G. and Stuart, A. [1966]. *The Advanced Theory of Statistics, Vol. 2.* Hafner, New York.

Mantel, N. [1963]. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J. Amer. Statist. Ass. 58*, 690–700.

Srole, L., Langner, T. S., Michael, S. T., Opler, M. K., and Rennie, T. A. C. [1962]. *Mental Health in the Metropolis: The Midtown Manhattan Study.* McGraw-Hill, New York.

Williams, O. D. and Grizzle, J. E. [1972]. Contingency tables having ordered response categories. *J. Amer. Statist. Ass. 67*, 55–63.

Wing, J. K. [1962]. Institutionalism in mental hospitals. *Brit. J. Soc. Clin. Psychol. 1*, 38–51.

Yates, F. [1948]. The analysis of contingency tables with groupings based on quantitative characters. *Biometrika 35*, 176–81.

APPENDIX: COMPUTATION OF THE MAXIMUM LIKELIHOOD ESTIMATE $\hat{\mu}$

To compute $\hat{\mathbf{u}}$, the result of Haberman [1974] is used that $\hat{\mathbf{u}}$ is the same under all four sampling methods considered, so it is sufficient to find $\hat{\mathbf{u}}$ by maximizing the kernel

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij}\mu_{ij} - e^{\mu_{ij}})$$

of the log likelihood obtained under sampling method (A) subject to the conditions that (1) holds and that $\alpha_{ij} = 0$ for $(i, j) \in S$. Since any $\boldsymbol{\mu}$ satisfying these conditions may be written as

$$\boldsymbol{\mu} = \sum_{(i,j)\in T}\alpha_{ij}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}, \tag{15}$$

where $T = \{(i, j): 1 \leq i \leq r, 1 \leq j \leq c, (i, j) \notin S\}$, differentiation of $L(\boldsymbol{\mu})$ yields the likelihood equations

$$\begin{aligned}(\mathbf{n}, \mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}) &= \sum_{k=1}^{r}\sum_{l=1}^{c} n_{kl}x_k^{(i)}y_l^{(j)} \\ &= (\hat{\mathbf{m}}, \mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}) \qquad ((i, j) \in T)\end{aligned} \tag{16}$$

where $\hat{\mathbf{m}} = \{\exp(\hat{\mu}_{ij})\}$. These equations hold if, and only if, for some b_{ij}, $(i, j) \in S$,

$$\mathbf{n} = \hat{\mathbf{m}} + \sum_{(i,j)\in S} b_{ij}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}, \tag{17}$$

since the vectors $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $i = 1, \cdots, r$, $j = 1, \cdots, c$, are an orthonormal basis of the space of $r \times c$ arrays (see Halmos [1958] page 174). In the special case in which no restraints are imposed, S is empty and $\hat{\mathbf{m}} = \mathbf{n}$. Since the second differential of $L(\boldsymbol{\mu})$ is negative definite, (16) and the constraints on $\hat{\boldsymbol{\mu}}$ uniquely determine the ML estimate.

In special cases, (16) or (17) may be used to find explicit expressions for $\hat{\mathbf{m}}$. The case $S = \emptyset$ has already been noted. If rows and columns are independent, then $S = \{(i, j): 2 \leq i \leq r, 2 \leq j \leq c\}$ and $\hat{m}_{ij} = n_{i+}n_{+j}/n_{++}$, where customary summation notation is used. In general, however, iterative solutions for these equations are necessary.

For the models considered in this paper, the Newton–Raphson algorithm is quite attractive since, as in logit analysis, it is not difficult to verify by computation of first and second differentials that the Newton–Raphson algorithm for maximization of the log likelihood may be described in terms of weighted least squares (see Haberman [1974]). To do so, suppose that an initial approximation $\boldsymbol{\mu}^{(0)}$ for $\hat{\boldsymbol{\mu}}$ is given and the Newton–Raphson algorithm is used to generate successive approximations $\boldsymbol{\mu}^{(\nu)}$ for $\hat{\boldsymbol{\mu}}$, $\nu \geq 1$. If $m_{ij}^{(\nu)} = \exp(\mu_{ij}^{(\nu)})$ is the approximation of $\hat{m}_{ij}$ corresponding to $\mu_{ij}^{(\nu)}$, then working log-means $\tilde{\mu}_{ij}^{(\nu)}$ may be defined by the equation

$$\tilde{\mu}_{ij}^{(\nu)} = \mu_{ij}^{(\nu)} + (n_{ij} - m_{ij}^{(\nu)})/m_{ij}^{(\nu)}. \tag{18}$$

Given the vectors $\tilde{\boldsymbol{\mu}}^{(\nu)} = \{\tilde{\mu}_{ij}^{(\nu)}\}$ and $\mathbf{m}^{(\nu)} = \{m_{ij}^{(\nu)}\}$, $\boldsymbol{\mu}^{(\nu+1)}$ is defined so that the weighted sum of squares

$$\sum_{i=1}^{r}\sum_{j=1}^{c} m_{ij}^{(\nu)}(\tilde{\mu}_{ij}^{(\nu)} - \mu_{ij}^{(\nu+1)})^2$$

is minimized subject to the constraint that for some $\alpha_{ij}^{(\nu+1)}$, $(i, j) \in T$,

$$\boldsymbol{\mu}^{(\nu+1)} = \sum_{(i,j)\in T}\alpha_{ij}^{(\nu+1)}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}. \tag{19}$$

Since $\{\log n_{ij}\}$ is the ML estimate of $\hat{\boldsymbol{\mu}}$ under the trivial model with $S = \emptyset$, this vector is a useful choice for $\boldsymbol{\mu}^{(0)}$. Given this choice, $\boldsymbol{\mu}^{(1)}$ minimizes the weighted sum of squares

$$\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\log n_{ij} - \mu_{ij}^{(1)})^2$$

subject to (19). As noted by Haberman [1974], $\mathbf{u}^{(1)}$ and $\hat{\mathbf{u}}$ have the same asymptotic distribution. Thus, convergence with initial value $\mathbf{u}^{(0)}$ is generally very rapid.

Minimization of a weighted sum of squares subject to linear constraints is a familiar statistical problem. It is well known that if T has f elements, $\boldsymbol{\alpha}^{(\nu+1)} = \{\alpha_{ij}{}^{(\nu+1)} : (i, j) \in T\}$, M is the rc by f matrix with columns $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $(i, j) \in T$, and $D(\mathbf{m})$ is the rc by rc diagonal matrix with diagonal elements m_{ij}, $1 \leq i \leq r$, $1 \leq j \leq c$, then

$$\mathbf{u}^{(\nu+1)} = M\boldsymbol{\alpha}^{(\nu+1)} \tag{20}$$

and

$$\boldsymbol{\alpha}^{(\nu+1)} = \{M'D(\mathbf{m}^{(\nu)})M\}^{-1}M'D(\mathbf{m}^{(\nu)})\tilde{\mathbf{u}}^{(\nu)}, \tag{21}$$

where M' is the transpose of M. If $(i, j) \in T - N$, then the asymptotic variance of $\hat{\alpha}_{ij}$, the ML estimate of α_{ij}, may be estimated by the (i, j)th diagonal element of $\{M'D(\hat{\mathbf{m}})M\}^{-1}$.

A less known alternate set of matrix equations is often more useful when the number of elements of S is small compared to rc. If S has d elements and M_0 is the rc by d matrix with columns $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $(i, j) \in S$, then

$$\mathbf{u}^{(\nu+1)} = \tilde{\mathbf{u}}^{(\nu)} - \{D(\mathbf{m}^{(\nu)})\}^{-1}M_0\mathbf{b}^{(\nu+1)}, \tag{22}$$

where $\{D(\mathbf{m}^{(\nu)})\}^{-1}$ is equal to $D(\{1/m_{ij}{}^{(\nu)}\})$ and

$$\mathbf{b}^{(\nu+1)} = [M_0'\{D(\mathbf{m}^{(\nu)})\}^{-1}M_0]^{-1}M_0'\tilde{\mathbf{u}}^{(\nu)}. \tag{23}$$

Use of (22) and (23) results in considerable reduction in computational labor when d is much less than f since in such cases inversion of the d by d matrix $M_0'\{D(\mathbf{m}^{(\nu)})\}^{-1}M_0$ is much easier than inversion of the f by f matrix $M'D(\mathbf{m}^{(\nu)})M$.

To show that (22) holds, note that since the vectors $\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}$, $i = 1, \cdots, r$, $j = 1, \cdots, c$, are an orthonormal basis of W, the space of r by c arrays, (20) and (21) are equivalent to the conditions

$$M_0'\mathbf{u}^{(\nu+1)} = \mathbf{0} \tag{24}$$

and

$$M'D(\mathbf{m}^{(\nu)})\mathbf{u}^{(\nu+1)} = M'D(\mathbf{m}^{(\nu)})\tilde{\mathbf{u}}^{(\nu)}. \tag{25}$$

Since

$$M_0'[\tilde{\mathbf{u}}^{(\nu)} - \{D(\mathbf{m}^{(\nu)})\}^{-1}M_0\mathbf{b}^{(\nu+1)}] = \mathbf{0} \tag{26}$$

and

$$M'D(\mathbf{m}^{(\nu)})[\tilde{\mathbf{u}}^{(\nu)} - \{D(\mathbf{m}^{(\nu)})\}^{-1}M_0\mathbf{b}^{(\nu+1)}] = M'D(\mathbf{m}^{(\nu)})\tilde{\mathbf{u}}^{(\nu)}, \tag{27}$$

(22) holds.

The coefficients $\mathbf{b}^{(\nu+1)}$ in (23) are approximations for the coefficients $\mathbf{b} = \{b_{ij} : (i, j) \in S\}$ in (17) as may be seen by noting that

$$M_0'\tilde{\mathbf{u}}^{(\nu)} = M_0'\{D(\mathbf{m}^{(\nu)})\}^{-1}(\mathbf{n} - \mathbf{m}^{(\nu)}) \tag{28}$$

and

$$\mathbf{n} - \hat{\mathbf{m}} = M_0\mathbf{b} \tag{29}$$

as $\mathbf{u}^{(\nu)}$ converges to $\hat{\mathbf{u}}$, $\mathbf{b}^{(\nu)}$ converges to $\mathbf{b}$.

If (22) and (23) are used, the coefficients $\hat{\boldsymbol{\alpha}}$ may be found by noting that $\hat{\boldsymbol{\alpha}} = M'\hat{\boldsymbol{\mu}}$. Since (20), (21), (22), and (23) imply that

$$M[M'D(\hat{\mathbf{m}})M]^{-1}M'D(\hat{\mathbf{m}}) = I - \{D(\hat{\mathbf{m}})\}^{-1}M_0[M_0'\{D(\hat{\mathbf{m}})\}^{-1}M_0]^{-1}M_0 , \tag{30}$$

where I is the identity matrix, it follows that the estimated asymptotic variance of $\hat{\alpha}_{ij}$, $(i, j) \in T - N$, is the (i, j)th diagonal element of

$$M'\{D(\hat{\mathbf{m}})\}^{-1}M - [M'\{D(\hat{\mathbf{m}})\}^{-1}M_0][M_0'\{D(\hat{\mathbf{m}})\}^{-1}M_0]^{-1}[M'\{D(\hat{\mathbf{m}})\}^{-1}M_0]',$$

which is equal to

$$(\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}, \{D(\hat{\mathbf{m}})\}^{-1}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}) - (M_0'\{D(\hat{\mathbf{m}})\}^{-1}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)},$$
$$[M_0'\{D(\hat{\mathbf{m}})\}^{-1}M_0]^{-1}M_0'\{D(\hat{\mathbf{m}})\}^{-1}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}).$$

If S is empty, then the estimated asymptotic variance of $\hat{\alpha}_{ij}$ is

$$(\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}, \{D(\hat{\mathbf{m}})\}^{-1}\mathbf{x}^{(i)} \otimes \mathbf{y}^{(j)}) = \sum_{k=1}^{r} \{x_k^{(i)}\}^2 \sum_{l=1}^{c} \{y_l^{(j)}\}^2/n_{kl} . \tag{31}$$

Received August 1973, *Revised March* 1974

Key Words: Log-linear models; Ordered classifications; Model selection.