ON LOG-LINEAR MODELS FOR MULTIPLE-RECORD SYSTEMS

M. Nabil El-Khorazaty, Gary G. Koch and Alcinda L. Lewis
University of North Carolina at Chapel Hill
Peter B. Imrey, University of Illinois

1. Introduction

Estimating the size of a population of humans, animals, or events relating to a community of such units is a major methodological problem shared by researchers in many disciplines. Where counts of organisms are involved, we may be concerned with such quantities as the total number of residents, victims of a congenital anomaly, criminals or victims of crimes, drug abusers, parties involved in automotive accidents, or of fish in a lake, deer in a forest, or bacteria on a microscope slide. Populations of events frequently studied are those of births, deaths, migrations, marriages, separations, divorces, and diagnoses of cancer or other diseases within a chosen time interval. Frequently, the ultimate aim is not only estimation of magnitude, as shown by a total, but rather of change over time as shown by a rate of population growth or the incidence rate of an event such as disease attack.

In principle, most data required to achieve reasonable aims for the study of human populations can be supplied by census (which covers the whole population) and civil registration systems (continuous recording of events of interest), supplemented by periodic sample surveys during intercensal periods. However, these traditional statistical systems are not adequate in many countries of the world and, for obvious reasons, are unsuitable for use with animal populations. Inadequacies of the traditional sources have led to the development of statistical procedures designed to combine the information available from multiple sources or protocols for detecting the same individuals or events, when each protocol is known to be insufficiently sensitive if used by itself. For human populations these procedures, generally grouped under the term Multiple-Record Systems (MRS), attempt to estimate the number of persons or events while adjusting for the individual fallibilities of census, survey or vital registration systems. Analogous "Capture-Mark-Recapture" (CMR) techniques attempt to compensate for the virtual impossibility of locating and distinguishing, in a short time, even the majority of members of any interesting animal population. In practice, CMR is distinguished by the use of i) detection protocols which are not coterminous in time, and ii) labeling and subsequent identification procedures which are unavailable for human studies.

The Multiple-Record System (MRS) involves data collection from two or more sources of information (called recording systems) which cover the same sample or sub-sample of areas and the same time period. The special case of two sources (Dual-Record System (DRS)) has been used widely in the last 30 years to adjust for omissions in the recording of vital events and to estimate population growth rates. In this regard, Chandrasekar and Deming (1949) present a theoretical framework for estimating the total number of events under the following Assumptions (1)-(3).

1. No coverage errors with respect to the scope of area and/or time period in which individuals or events are recorded (i.e., each system only records individuals or events that pertain to the target area and/or time period under study);
2. Independence of recording systems (i.e., the probability that both systems detect a randomly chosen individual or event is the product of the probabilities of detecting a randomly chosen element for each system individually);
3. No misclassification errors with respect to determining exactly which systems have detected an individual or event (i.e., a perfect matching rule exists for linking information from the two systems to determine correctly the number of individuals or events detected by both).

Chakraborty (1963) and Das Gupta (1964) extend this approach to situations involving three or more sources of information.

A fruitful approach to the study of MRS and CMR data is to view the involved set of detection protocols (recording systems for MRS, capture or observation methods or times for CMR) as a probabilistic process, or channel, with an input and resulting output. Input to the processor is a single element of the population to be studied, while the resulting output is a response pattern which delineates exactly which of the various detection protocols (if any) have recorded, or captured, that element. The aggregate result of passing every population element through the processor may be arranged, for d detection protocols, in a 2^d contingency table with each dimension, or marginal, describing the success or failure of a single protocol in capturing the elements of the target population. In a typical MRS or CMR application, we see this contingency table absent the single cell containing those elements missed by all detection protocols. It is this "incomplete contingency table" which must be used to generate estimates of population size. Sometimes the population under study may be partitioned into a set of subpopulations according to such demographic variables as geographic location, urbanization, or sex, so that we see such an incomplete table with its missing cell for each subpopulation.

Fienberg (1972), Bishop, Fienberg and Holland (1975), El-Khorazaty (1975) and Koch, El-Khorazaty and Lewis (1977) advocate fitting log-linear models to the observed cells of the above tables, and using these models to obtain refined estimates of the missing cell(s) and, hence, of the population total. Such refined estimates may be obtained by

i) controlling for statistical dependence of specific types among the actions of the various detection protocols;
ii) accounting for or modeling the effects of subpopulations or their defining factors

on the probabilistic properties of the detection protocols.

This generalization of the Chandrasekar-Deming approach for the DRS and the Peterson-Lincoln approach for the CMR allows great latitude in choosing an estimation procedure realistically adapted to the properties of actual recording systems.

In this paper we give a matrix formulation of the general log-linear model applicable to data obtained from the operation of a multiple-record system on a stratified population. The matrix formulation yields explicit matrix product expressions for the true and estimated asymptotic covariance matrices of efficient estimators for the log-linear model parameter vector, as well as corresponding results for the asymptotic covariances of fitted detection probabilities in the several strata, and of the stratum-specific inflation factors used to estimate the stratum sizes.

2. Matrix Formulation of the Log-Linear Model

2.1. Notation

Let $i = 1,2,\ldots,s$ index a set of sub-populations (or strata). Let $g = 1,2,\ldots,d$ index a set of recording systems, and $j_g = 1,2$ represent the presence $(j_g = 1)$ or absence $(j_g = 2)$ of the attribute corresponding to registration by the g-th recording system. Let the vector subscript $\underset{\sim}{j}' = (j_1,j_2,\ldots,j_d)$ index the multivariate response profiles for simultaneous recording status with respect to the d recording systems. Let $n_{i;j_1j_2\cdots j_d}$ denote the number of elements from the i-th sub-population with recording status $(j_1,j_2,\ldots,j_d)$. Let

$$p_{i;j_1j_2\cdots j_d} = n_{i;j_1j_2\cdots j_d}/n_i \qquad (2.1.1)$$

where n_i is the total number of elements of sub-population i $(i = 1,2,\ldots,s)$ recorded by any of the record systems. Let

$$\overline{\pi}_{i;j_1j_2\cdots j_d} = \frac{\pi_{i;j_1j_2\cdots j_d}}{1 - \pi_{i;22\ldots2}} \qquad (2.1.2)$$

where $\pi_{i;j_1j_2\cdots j_d}$ is the probability that a population element has recording status $\underset{\sim}{j}$ for sub-population i. Thus, $\overline{\pi}_{i;j_1j_2\cdots j_d}$ denotes the conditional probability that a population element has recording status $\underset{\sim}{j}$ for sub-population i, given that it is observed.

2.2. General Procedure for Estimating Population Size

(a) Formulate a rank $t \le s(2^d-2)$ log-linear model for the $\pi_{i;j_1j_2\cdots j_d}$;

(b) Estimate the parameters of the model from the unrestricted maximum likelihood estimators $p_{i;j_1j_2\cdots j_d}$ of $\overline{\pi}_{i;j_1j_2\cdots j_d}$ provided by the observed incomplete contingency table;

(c) Obtain estimates $\hat{\pi}_{i;j_1j_2\cdots j_d}$ of $\pi_{i;j_1j_2\cdots j_d}$ from the fitted model;

(d) Estimate N_i, the size of stratum i, by $\dfrac{n_i}{1-\hat{\pi}_{i;22\ldots2}}$.

When $s = 1$, $d = 2$ and the model chosen in (a) contains only main effects corresponding to the two record systems (which are thus assumed to operate independently in the sense of no association of detection by the two systems), the above procedure yields the classical Chandrasekar-Deming estimate. When $s = 1$ and $d > 2$, the choice of a model with no interaction terms, but with main effects corresponding to each system, yields the extension of their estimator derived by Chakraborty (1963).

2.3. Representation of the Model

For any specific stratum i, a general log-linear model for the corresponding $\pi_{i;j_1j_2\cdots j_d}$ may be written as

$$\underset{\sim}{\pi}_i = \underset{\sim}{\pi}_i(\underset{\sim}{\beta}) = \underset{\sim\sim}{\exp}(\underset{\sim}{X}_i\underset{\sim}{\beta})/\underset{\sim}{1}_r'\underset{\sim\sim}{\exp}(\underset{\sim}{X}_i\underset{\sim}{\beta}) \qquad (2.3.1)$$

where $\underset{\sim}{\pi}_i$ is the vector of the $\pi_{i;\,j_1j_2\cdots j_d}$ arranged in lexicographic order, $\underset{\sim}{X}_i$ is a known "design matrix" specifying the structure of the log-linear model, $\underset{\sim}{\beta}$ is the (unknown) vector of $t \le (r-2)$ model parameters, $\underset{\sim}{1}_k$ is a k-vector of units, $\underset{\sim\sim}{\exp}$ is the elementwise exponential operator, and $r = 2^d$. $\underset{\sim}{X}_i$ is assumed to be of full rank, with columns jointly linearly independent of the vector $\underset{\sim}{1}_r$ representing the underlying linear restriction that $\sum_j \pi_{i;j} = 1$. Fienberg (1972) and Bishop, Fienberg and Holland (1975) further restrict to the class of hierarchical analysis of variance models, due to the ease of obtaining the maximum likelihood estimate $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$, under the conditional multinomial likelihood for the n_{ij}, through the computational technique of "iterative proportional fitting". These models correspond to design matrices $\underset{\sim}{X}_i$ for which the set of columns of $\underset{\sim}{X}_i$, $[\underset{\sim}{X}_i]_c$, can be written as

$$[\underset{\sim}{X}_i]_c = \bigcup_{k=1}^{K} [\underset{\sim}{X}_{ik}]_c \quad \text{for some } K,$$

where each $\underset{\sim}{X}_{ik}$ is the usual design matrix corresponding to a complete (or saturated) factorial model involving as factors some subset of the d record systems. Results of this section apply to general $\underset{\sim}{X}_i$; we adopt the conditional multinomial likelihood, but address directly the problem of obtaining estimates only in Section 4.

For $s > 1$, we may generalize the above formulation to a model for $\underset{\sim}{\pi}' = (\underset{\sim}{\pi}_1',\underset{\sim}{\pi}_2',\ldots,\underset{\sim}{\pi}_s')$ as

$$\underset{\sim}{\pi} = \underset{\sim}{\pi}(\underset{\sim}{\beta}) = \underset{\sim}{D}_\eta^{-1}\{\underset{\sim\sim}{\exp}\,\underset{\sim}{X}\underset{\sim}{\beta}\} . \qquad (2.3.2)$$

Here the vector $\underset{\sim}{\beta}$ of unknown model parameters is of dimension $t \le s(r-2)$, and underlies the joint detection probabilities for all strata through the composite design matrix

$$\underset{rs\times t}{\underset{\sim}{X}'} = (\underset{\sim}{X}_1',\underset{\sim}{X}_2',\ldots,\underset{\sim}{X}_s') .$$

For general $\underset{\sim}{y}$, $\underset{\sim}{D}_y$ is the diagonal matrix with $\underset{\sim}{y}$ on the principal diagonal, and

$$\underset{\sim}{\eta} = [\underset{\sim}{1}_{rr} \otimes \underset{\sim}{I}_s]\ \underset{\sim\sim}{\exp}(\underset{\sim}{X}\underset{\sim}{\beta})$$

with $\underset{\sim}{1}_{vw}$, $\underset{\sim}{I}_v$ and $\otimes$ representing a $v \times w$ matrix of units, the $v \times v$ identity and Kronecker multiplication respectively. Each of the $\underset{\sim}{X}_i$ is assumed of full rank, with columns jointly linearly independent of $\underset{\sim}{1}_r$. Otherwise the $\underset{\sim}{X}_i$ may be of essentially free form and vary considerably from stratum to stratum. In particular, some

columns of $\underset{\sim}{X}_i$ may be $\underset{\sim}{0}$, indicating that certain parameters of β apply only to certain strata. Clearly, a model with small t is desirable if realistic.

3. Determination of Covariance Structure

As noted previously, we assume for the generated cell counts a multinomial distribution conditional on the totals n_i of elements detected in stratum i by any of the record systems. Thus, the joint likelihood ϕ may be written as

$$\phi = \prod_{i=1}^{s}\left[n_i! \prod_{\underset{\sim}{j}\neq(2,2,\ldots,2)} \pi_{i;\underset{\sim}{j}}^{n_{i;\underset{\sim}{j}}}/n_{i;\underset{\sim}{j}}!\right], \quad (3.1)$$

with the s constraints that

$$\sum_{\underset{\sim}{j}\neq(2,2,\ldots,2)} \pi_{i;\underset{\sim}{j}} = 1 \quad \text{for all} \quad i = 1,2,\ldots,s .$$

Incorporating the log-linear model to this likelihood and expressing the result in matrix terms yields

$$\phi = \frac{\prod_{i=1}^{s} n_i!}{\prod_{i=1}^{s}\prod_{\underset{\sim}{j}\neq(2,2,\ldots,2)} n_{i;\underset{\sim}{j}}!} \frac{\exp(\underset{\sim}{n}_0'\underset{\sim}{X}_0\beta)}{\prod_{i=1}^{s}\{\underset{\sim}{1}'_{(r-1)}\exp(\underset{\sim}{X}_{i0}\beta)\}^{n_i}} \quad (3.2)$$

where $\underset{\sim}{n}_0' = (\underset{\sim}{n}_{10}', \underset{\sim}{n}_{20}', \ldots, \underset{\sim}{n}_{s0}')$, $\underset{\sim}{n}_{i0}$ is the vector of observed cell counts from the i-th stratum incomplete table (lexicographic order), $\underset{\sim}{X}_{i0}$ is the matrix consisting of the first $(r-1)$ rows of $\underset{\sim}{X}_i$, and $\underset{\sim}{X}_0' = (\underset{\sim}{X}_{10}', \underset{\sim}{X}_{20}', \ldots, \underset{\sim}{X}_{s0}')$.

The asymptotic covariance matrix of any asymptotically efficient (such as maximum likelihood or minimum Neyman chi-square) estimate $\hat{\beta}$ of β is available as the negative inverse of Fisher's Information Matrix. Using matrix differentiation methods similar to those of Forthofer and Koch (1973), we obtain

$$\frac{d}{d\beta}[\log\phi] = -\underset{\sim}{n}_0'\underset{\sim}{X} - [\underset{\sim}{1}_{(r-1)} \otimes \underset{\sim}{n}_\cdot]'\underset{\sim}{D}_{\underset{\sim}{\Pi}_0(\beta)}\underset{\sim}{X}_0 , \quad (3.3)$$

where $\underset{\sim}{n}_\cdot' = (n_1, n_2, \ldots, n_s)$ and $\underset{\sim}{\Pi}_0(\beta)$ is defined analogously to $\underset{\sim}{n}_0$ from the $\pi_{i;\underset{\sim}{j}}$. Hence, the asymptotic covariance $\underset{\sim}{V}_{\hat{\beta}}(\underset{\sim}{\Pi}_0)$ of $\hat{\beta}$ is obtained as

$$\underset{\sim}{V}_{\hat{\beta}}(\underset{\sim}{\Pi}_0) = \left\{\frac{-d^2}{d\beta d\beta'}[\log\phi]\right\}^{-1}$$

$$= \left\{\sum_{i=1}^{s} n_i \underset{\sim}{X}_{i0}'[\underset{\sim}{D}_{\underset{\sim}{\Pi}_0} - \underset{\sim}{\Pi}_0\underset{\sim}{\Pi}_0']\underset{\sim}{X}_{i0}\right\}^{-1}, \quad (3.4)$$

following simplifications. Since $\hat{\underset{\sim}{\Pi}}_0 = \underset{\sim}{\Pi}_0(\hat{\beta})$ is consistent for $\underset{\sim}{\Pi}_0 = \underset{\sim}{\Pi}_0(\beta)$, a consistent estimator for the covariance matrix $\underset{\sim}{V}_{\hat{\beta}}(\underset{\sim}{\Pi}_0)$ is

$$\underset{\sim}{V}_{\hat{\beta}} = \underset{\sim}{V}_{\hat{\beta}}(\hat{\underset{\sim}{\Pi}}_0) = \left\{\sum_{i=1}^{s} n_i \underset{\sim}{X}_{i0}'[\underset{\sim}{D}_{\hat{\underset{\sim}{\Pi}}_0} - \hat{\underset{\sim}{\Pi}}_0\hat{\underset{\sim}{\Pi}}_0']\underset{\sim}{X}_{i0}\right\}^{-1} . \quad (3.5)$$

The asymptotic covariance matrix for the estmator $\hat{\underset{\sim}{\Pi}}_0$ of the vector of conditional probabilities of the response profiles for the various strata, and the estimators $\hat{\gamma}_i = \gamma_i(\hat{\beta})$ of the stratum-specific ratios

$$\gamma_i = \frac{\pi_{i;22\ldots2}}{1-\pi_{i;22\ldots2}} = \frac{\pi_{i;22\ldots2}}{[\underset{\sim}{1}'_{(r-1)}, 0]\underset{\sim}{\pi}_i} ,$$

is obtained by use of the well-known δ-method as based on the first-order Taylor series approximations for these estimators. In this regard, the compound function notation used in Forthofer and Koch (1973) is used to express $\hat{\underset{\sim}{\Pi}}_0$ and the $\hat{\gamma}_i$ in the form

$$\hat{\underset{\sim}{\theta}}' = (\hat{\underset{\sim}{\Pi}}_{10}, \hat{\gamma}_1, \hat{\underset{\sim}{\Pi}}_{20}, \hat{\gamma}_2, \ldots, \hat{\underset{\sim}{\Pi}}_{s0}, \hat{\gamma}_s)$$

as $\hat{\underset{\sim}{\theta}} = \hat{\underset{\sim}{\theta}}(\hat{\beta}) = \underset{\sim}{\exp}[\underset{\sim}{A}_3\underset{\sim}{\log}\{\underset{\sim}{A}_2\underset{\sim}{\exp}(\underset{\sim}{A}_1\hat{\beta})\}]$ where $\underset{\sim}{A}_1 = \underset{\sim}{X}_0$,

$$\underset{\sim}{A}_2 = \begin{bmatrix} \underset{\sim}{I}_r \\ \underset{\sim}{1}'_{(r-1)}, \; 0 \end{bmatrix} \otimes \underset{\sim}{I}_s ,$$

$$\underset{\sim}{A}_3 = [\underset{\sim}{I}_r, \; -\underset{\sim}{1}_r] \otimes \underset{\sim}{I}_s ,$$

and $\underset{\sim}{\log}$ is the elementwise logarithmic operator. As a result, a consistent estimator for the corresponding asymptotic covariance matrix can be determined as the matrix product

$$\hat{\underset{\sim}{V}}_{\hat{\underset{\sim}{\theta}}} = \underset{\sim}{D}_{\underset{\sim}{\chi}_3}\underset{\sim}{A}_3\underset{\sim}{D}^{-1}_{\underset{\sim}{a}_2}\underset{\sim}{A}_2\underset{\sim}{D}_{\underset{\sim}{\chi}_1}\underset{\sim}{A}_1[\underset{\sim}{V}_{\hat{\underset{\sim}{\theta}}}(\hat{\underset{\sim}{\Pi}}_0)]\underset{\sim}{A}_1'\underset{\sim}{D}_{\underset{\sim}{\chi}_1}\underset{\sim}{A}_2'\underset{\sim}{D}^{-1}_{\underset{\sim}{a}_2}\underset{\sim}{A}_3'\underset{\sim}{D}_{\underset{\sim}{\chi}_3} \quad (3.6)$$

where $\underset{\sim}{\chi}_1 = \underset{\sim}{\exp}(\underset{\sim}{A}_1\hat{\beta})$, $\underset{\sim}{a}_2 = \underset{\sim}{A}_2\underset{\sim}{\chi}_1$, $\underset{\sim}{\chi}_3 = \underset{\sim}{\exp}\{\underset{\sim}{A}_3\underset{\sim}{\log}(\underset{\sim}{a}_2)\}$.

The approach ultimately leads as described in Section 2.2, to the estimators

$$\hat{n}_{i;22\ldots2} = n_i\hat{\gamma}_i$$

and

$$\hat{N}_i = n_i(1+\hat{\gamma}_i)$$

for the missing cell and total size of stratum i, and

$$\hat{N} = \sum_{i=1}^{s} n_i(1+\hat{\gamma}_i)$$

for the population size. Since the n_i are random variables assumed to have independent binomial distributions with parameters N_i and $(1-\pi_{i;22\ldots2})$, the methods indicated in Darroch (1958) and Fienberg (1972) can be used in conjunction with the above results to produce estimators for the asymptotic variances of the $\hat{n}_{i;22\ldots2}$, $\hat{N}_i$ and $\hat{N}$. In particular, these quantities reduce to

$$V_{\hat{n}_{i,22\ldots2}} = n_i^2 V_{\hat{\gamma}_i} + \{\hat{n}^3_{i;22\ldots2}/n_i\hat{N}_i\}, \quad (3.7)$$

$$V_{\hat{N}_i} = n_i^2 V_{\hat{\gamma}_i} + \{\hat{n}_{i;22\ldots2}\hat{N}_i/n_i\} . \quad (3.8)$$

$$V_{\hat{N}} = \sum_{i=1}^{s}\sum_{i'=1}^{s} n_i n_{i'} V_{\hat{\gamma}_i,\hat{\gamma}_{i'}} + \sum_{i=1}^{s}\{\hat{n}_{i;22\ldots2}\hat{N}_i/n_i\} \quad (3.9)$$

where $V_{\hat{\gamma}_i}$ and $V_{\hat{\gamma}_i,\hat{\gamma}_i}$ are estimates for the variance of $\hat{\gamma}_i$ and covariance of $\hat{\gamma}_i$ and $\hat{\gamma}_{i'}$, from (3.6).

For the case $s = 1$, the estimators for the asymptotic variances of $\hat{n}_{i;22\ldots2}$ and $\hat{N}_i$ are essentially the same as those given by Fienberg (1972) but avoid iterative computations required in general by his approach even after estimation of $\hat{\gamma}_i$, $\hat{n}_{i;22\ldots2}$ and $\hat{N}_i$.

4. Strategies for Fitting Log-Linear Models

In this section, we describe three strategies available for estimating parameters and fitted joint detection probabilities for the models of Section 2. Associated statistics for evaluating adequacy of fit are also referenced.

The most general method involves a slight modification of the approach of Grizzle and Williams (1972) for fitting log-linear models to complete contingency tables, which they developed as an application of the general methodology

described by Grizzle, Starmer and Koch (1969). Weighted least-squares (WLS) computational algorithms are applied to fit the postulated log-linear model to the observed vector $\underset{\sim}{\log}\, \underset{\sim}{p}$, where $\underset{\sim}{p}$ contains the various $p_{i;j_1j_2\cdots j_d}$, the unrestricted maximum likelihood estimate of $\pi_{i;j_1j_2\cdots j_d}$. The covariance matrix used is obtained by substituting the $p_{i;j_1j_2\cdots j_d}$ for $\pi_{i;j_1j_2\cdots j_d}$ in the asymptotic covariance matrix of $\underset{\sim}{\log}\, \underset{\sim}{p}$, determined by applying the δ-method for deriving the covariances of transformed random variables (see Grizzle, Starmer and Koch (1969)). Thus, $\underset{\sim}{\log}\, \underset{\sim}{p}$ is expanded in a Taylor series about $\underset{\sim}{\log}\, \underset{\sim}{\pi}_0$, and the covariance matrix of the linear term extracted. This method yields a direct estimate $\bar{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ without iteration; the fitted joint detection probabilities are obtained by substituting $\bar{\underset{\sim}{\beta}}$ into the model equations. The estimator $\bar{\underset{\sim}{\beta}}$ is a member of the class of procedures based on minimizing Neyman's (1949) modified chi-square criterion subject to a linearized hypothesis. As such, it is a Best Asymptotically Normal (BAN) estimate of $\underset{\sim}{\beta}$. For moderate to large samples in practice, $\bar{\underset{\sim}{\beta}}$ tends to be close to the estimate $\tilde{\underset{\sim}{\beta}}$ which maximizes the conditional likelihood based on the n_i, in the sense that individual components of $\bar{\underset{\sim}{\beta}}$ and $\tilde{\underset{\sim}{\beta}}$ tend to differ by less than the estimated standard deviation of either. The usual weighted least-squares algorithms produce test statistics, both for fit of the model and additional linear parametric restrictions, which belong to the class of test criteria defined by Wald (1943). Thus, in terms of asymptotic distribution and power, they are equivalent to the corresponding likelihood ratio tests based on the conditional likelihood. All computations for this approach may be executed using a general computer program for the analysis of categorical data, GENCAT (Landis, Stanish, Freeman and Koch (1976)), available from the University of Michigan.

When the postulated model is hierarchical, in terms of the entire set of dimensions involving both the different recording systems and the stratification variables, then the maximum likelihood estimators and likelihood ratio tests (based on the conditional likelihood) may be easily obtained. In this situation

$$[X^*] = \bigcup_{m=1}^{M} [X^*_{(m)}]$$

where $X^*_{(m)}$ is the design matrix of a factorial model involving a subset of the dimensions determined by the recording systems and/or stratification variables. If M is minimal, the observed marginal tables generating the $X^*_{(m)}$, $m=1,2,\ldots,M$, form a set of minimal sufficient statistics for the parameters of the model X^* (Birch (1963), Bishop, Fienberg and Holland (1975)). The sufficient statistics not only generate the fitted table which maximizes the underlying likelihood, but in fact are reproduced by it, as the maximum likelihood estimates of joint detection probabilities are the unique set of probabilities which both satisfy the model structure and generate marginal expected counts identical to the set of minimal sufficient statistics (Birch (1963)). This result is expressed in more general terms by expression (3.3).

For some models, the fitted joint detection probabilities may be calculated explicitly and directly from the sufficient marginal tables. Generally, however, it is necessary to use a modification of the technique of "iterative proportional fitting" (IPF, or "raking") of Deming and Stephan (1940), which converges correctly in all cases. If $C_{(m)}$ is the observed marginal table generated by $X^*_{(m)}$, the technique is executed as follows:

i) form Table T, of the same dimensions as the observed incomplete table, with zeros in cells representing unobserved elements and units in all other cells.

ii) collapse T to form the marginal array $C^{(1)}_{(1)}$ generated by $X^*_{(1)}$ from T; form $T^{(1)}_{(1)}$ by inflating each cell of T by the ratio of its marginal category frequency in $C_{(1)}$ to that in $C^{(1)}_{(1)}$.

iii) for $m = 2,\ldots,M$ form $C^1_{(m)}$ from $T^1_{(m-1)}$ using $X^*_{(m)}$; form $T^1_{(m)}$ from $T^1_{(m-1)}$ using $C^1_{(m)}$.

iv) for $\nu \leq 2$, cycle through ii) to iii) substituting $T^{(\nu-1)}_{(m)}$ for T, $C^{(\nu)}_{(m)}$ for $C^{(1)}_{(m)}$, and $T^{(\nu)}_{(m)}$ for $T^{(1)}_{(m)}$; continue until $T^{(\nu)}_{(m)}$, and $T^{(\nu-1)}_{(m)}$ are sufficiently close.

The elements of $T^{(\nu)}_{(m)}$ are then divided by the appropriate stratum sizes n_i to yield the joint detection probabilities for each stratum. The estimated parameter vector $\tilde{\underset{\sim}{\beta}}$ is obtained by substituting these into the model equations and solving, if desired. Estimated population size for each stratum, and likelihood ratio tests of fit associated with the model, or with comparisons of alternate models, may be calculated using the fitted probabilities without ever explicitly obtaining the model parameters.

The estimates obtained by IPF may be preferable to those given by the WLS procedure when some observed stratum counts are modest, inasmuch as the asymptotic theory for the maximum likelihood estimators depends on the expected counts in cells of the marginal tables $C_{(m)}$, $m = 1,2,\ldots,M$, where as that underlying the WLS approach depends on expected counts in individual cells of the incomplete table. All computations for the IPF-MLE analysis may be executed using a computer program for fitting log-linear models to contingency tables, ECTA, available from the University of Chicago.

When modest observed counts make the use of WLS unattractive and the proposed model is not hierarchical, estimates may often be derived by applying a Functional Asymptotic Regression Methodology (FARM) approach. This capitalizes on the observation that non hierarchical models can be written as hierarchical models with linear restrictions on the parameters. Thus, we attempt to find an unsaturated hierarchical model from which the

non hierarchical model of interest may be derived through the imposition of linear restrictions. IPF is applied to derive an initial estimate of β whose sampling variability derives from the expected counts corresponding to margins of the observed data table rather than interior cells. WLS algorithms are then applied to this preliminary estimate of β, using its estimated asymptotic covariance matrix under the hierarchical model (as determined by the δ-method), to introduce the linear restrictions which reduce the hierarchical model to the more parsimonious non hierarchical model of interest. The appropriate likelihood ratio test is used to assess fit of the initial hierarachical model, and a conditional WLS test used to evaluate adequacy of the subsequent reduction. Further reductions of the non hierarchical model may be evaluated by the application of WLS to the fitted parameters.

The FARM procedure is somewhat simpler to implement computationally than to describe conceptually. Once an initial hierarchical model is chosen, fitted joint detection probabilities are obtained under this model by IPF or WLS. The FARM estimate of β is then obtained by applying the WLS computational algorithms to the vector of these estimated probabilities instead of the observed proporation vector p. As a result, the FARM analysis may be performed, when IPF is used in the first stage, by simple execution in sequence of the computer programs ECTA and GENCAT described previously.

The literature describing development of log-linear model theory and the fitting strategies described here is vast, and no attempt has been made in this paper to adequately credit the contributors. The application of log-linear model theory to the MRS-CMR problem is due to Fienberg (1972), and a full exposition of the IPF-MLE approach appears in Bishop, Fienberg and Holland (1975). The WLS approach was adapted and described by Koch, El-Khorazaty and Lewis (1977), while FARM procedures are due to Koch, Imrey, Freeman, and Tolley (1977), and are applied to the MRS problem by El-Khorazaty, Imrey, Koch and Lewis (1977).

ACKNOWLEDGEMENTS

This research was in part supported by the U.S. Bureau of the Census through Joint Statistical Agreement JSA 76-93. The authors would like to thank Jackie O'Neal for her conscientious typing of this manuscript.

REFERENCES

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society*, Series B, 25, 220-33.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Mass.

Chakraborty, P.N. (1963). On a method of estimating birth and death rates from several agencies. *Calcutta Statistical Association, Bulletin 12*, 106-12.

Chandrasekar, C. and Deming, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-15.

Darroch, J.N. (1958). The multiple-recapture census, I. Estimation of a closed population. *Biometrika*, 45, 343-59.

Das Gupta, P. (1964). On the estimation of the total number of events and the probabilities of detecting an event from information supplied by several agencies. *Calcutta Statistical Association, Bulletin 13*, 89-100.

Deming, W.E. and Stephan, F.F. (1940). On a least-square adjustment of a sampled frequency table when the marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-44.

El-Khorazaty, M.N. (1975). Methodological strategies for the analysis of categorical data from multiple-record systems. University of North Carolina Institute of Statistics Mimeo Series No. 1019, Chapel Hill, N.C.

El-Khorazaty, M.N., Imrey, P.B., Koch, G.G. and Lewis, A.L. (1977). Applications of functional asymptotic regression methodology to multiple-record system data. In preparation.

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

Forthofer, R.N. and Koch G.G. (1973). An analysis for compounded functions of categorical data. *Biometrics*, 29, 143-57.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.

Grizzle, J.E. and Williams, O.D. (1972). Log-linear models and tests of independence for contingency tables. *Biometrics*, 28, 137-56.

Koch, G.G., El-Khorazaty, M.N. and Lewis, A.L. (1977). The asymptotic covariance structure of log-linear model estimated parameters for the multiple recapture census. To appear in *Communications in Statistics A*.

Koch, G.G., Freeman, D.H., Jr., Imrey, P.B. and Tolley, H.D. (1977). The asymptotic covariance structure of estimated parameters from contingency table log-linear models. To appear in *Biometrics*.

Landis, J.R., Stanish, W.M., Freeman D.H., Jr., and Koch, G.G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least-squares to generate Wald statistics. To appear in *Computer Programs in Biomedicine*.

Neyman, J. (1949). Contributions to the theory of the chi-square test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, 230-73. University of California Press, Berkeley, Calif.

Wald, A. (1943). Tests of statistical hypotheses concerning general parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-82.