CROSS-CLASSIFIED DATA FROM COMPLEX SAMPLE SURVEYS: THE EFFECTS OF COMPLEX ADJUSTMENT PROCEDURES ON VARIANCES

Daniel H. Freeman, Jr.* and Gary G. Koch
University of North Carolina at Chapel Hill

1. Introduction

The United States Bureau of the Census response error model (hereafter, CRE) has been found to be useful for the elucidation of non-sampling errors by a number of investigators (Hansen et. al. (1961), Hansen et. al. (1964), and Koch et. al. (1975)). The CRE extends in a straightforward manner to complex cross-classifications of survey data (Freeman (1975)). This paper will utilize the CRE model to investigate the effects of complex post-sampling data adjustments on the estimated variances.

The major portion of the discussion is devoted to the effect of "raking the data" to known margins. This procedure, proposed by Deming and Stephan (1940), is also known as iterative proportional fitting (denoted IPF). IPF is also used to obtain maximum likelihood estimates for fitting log-linear models to data from a Poisson or multinomial distribution (Fienberg (1970)). An approximation to the variance matrix generated by the procedure is given in Section 3 along with a statistic for the evaluation of the variance reduction achieved by the procedure. These results are then applied to the special case of simple ratio adjustments for representativeness and non-response. These simple ratio adjustments have been investigated empirically by Brock et. al. (1975).

While IPF is a post-sampling adjustment aimed at variance reduction, it is also of interest to fit other types of functions to survey data. Examples of these are functions which are directed at the underlying process which generated the responses of the surveyed populations. The multiple logit and the bivariate Weibull are two examples of such functions and are discussed in Section 5. The important point that is brought out is that these functions are within the scope of the CRE and the method of analysis is completely consistent with that used for the IPF discussion.

2. The IPF Post-sampling Adjustment Procedure

One approach to the minimization of non-sampling errors is to adjust the survey estimates of the frequency counts in cross-classified data to correspond to an independently determined set of population parameters of margins. The IPF adjustment uses these margins directly in the adjustment process. Thus the data set is forced to reflect pre-determined criteria of representativeness, an intuitively appealing property. However, in the context of simple random sampling IPF has several additional properties. Fienberg (1970) showed that IPF will converge to a set of domain or cell estimates under fairly general conditions. These conditions are that an underlying multinomial is a correct model and that there are observations in all of the fitted cells. The latter condition is true for many large surveys as long as the cross-classification is not too complex. The importance of the former condition is now known. A related property is that underlying measures of association are preserved by the procedure (Mosteller (1968)). A third property, to be exploited here, is that IPF maximized the likelihood equation of the multinomial distribution (Fienberg (1970)).

Some notation is required for the simplicity of discussion. Let the population consist of N individuals, denoted i, n of which are selected for the sample. The sampling procedure may be characterized by the indicator function, u_i:

$$u_i = \begin{cases} 1 & \text{if individual } i \text{ is selected,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For simplicity it is assumed that there are no complex measurement effects. Implicitly there is a known structure on the expected response vector, $\underset{\sim}{Y}_i$. This means that the category of response, j, reflects the joint response components for $\underset{\sim}{Y}_i$ which may be arrayed as an s-dimensional contingency table. The k-th variable may contain up to J_k levels, k = 1,...,s.

For example, consider a set of 8 binary responses. Then $\underset{\sim}{Y}_i$ is a vector with 2^8 components and,

$$\underset{\sim}{Y}_{ij} = Y_{i(j_1,\ldots,j_8)} = \begin{cases} 1 & \text{if individual } i \text{ is classified in category } (j_1,\ldots,j_8) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, $j_k = 1,2$ for k = 1,2,...,8.

The corresponding estimates for the population distribution are given by,

$$\hat{\underset{\sim}{N}} = \sum_{i=1}^{N} \frac{u_i}{\phi_i} \underset{\sim}{Y}_i \quad \text{and} \tag{3}$$

$$\hat{\underset{\sim}{N}}' = (\hat{N}_{11\ldots1}, \hat{N}_{11\ldots2}, \hat{N}_{22\ldots1}, \hat{N}_{22\ldots2})$$

where $E\{u_i\} = \phi_i$ is the selection probability of the i-th individual and,

$$\hat{N}_{j_1,j_2,\ldots,j_8} = \sum_{i=1}^{N} \frac{u_i}{\phi_o} Y_{i(j_1,j_2,\ldots,j_8)}. \tag{4}$$

The components of various margins may also be expressed as vectors. For example,

$$\hat{N}_{+\ldots j_k,\ldots j_{k''}\ldots+} = \sum_{i=1}^{N} \frac{u_i}{\phi_i} \sum\ldots\sum Y_{i(j_1,j_2,\ldots,j_8)} \tag{5}$$

*now at Yale University

where the sum is over all $\underset{\sim}{j}$ with $j_{k'} = j_{k'}$, $j_{k}'' = j_{k}''$. Then

$$\hat{N}_{+\ldots k'\ldots k''\ldots+} = (\hat{N}_{+\ldots 1\ldots 1\ldots+}, \hat{N}_{+\ldots 1\ldots 2\ldots+}, \ldots, \hat{N}_{+\ldots J_{k'}\ldots J_{k''}\ldots+}).$$

For the general situation suppose the (s-2,s-1), (s-2,s) and (s-1,s) pairwise margins may be taken as given. Then the IPF adjustment is given by,

$$N^{*(I,1)}_{j_1,j_2,\ldots,j_s} = N^{*(I-1,3)}_{j_1,j_2,\ldots,j_s} \frac{N_{+\ldots+,j_{s-2},j_{s-1},+}}{N^{*(I-1,3)}_{+\ldots+,j_{s-2},j_{s-1},+}};$$

$$N^{*(I,2)}_{j_1,j_2,\ldots,j_s} = N^{*(I,1)}_{j_1,j_2,\ldots,j_s} \frac{N_{+\ldots+,j_{s-2},+,j_s}}{N^{*(I,1)}_{+\ldots+,j_{s-2},+,j_s}}; \tag{6}$$

$$N^{*(I,3)}_{j_1,j_2,\ldots,j_s} = N^{*(I,2)}_{j_1,j_2,\ldots,j_s} \frac{N_{+\ldots+,j_{s-1},j_s}}{N^{*(I,2)}_{+\ldots+,j_{s-1},j_s}}$$

where $N^{*(0,3)}_{j_1,j_2,\ldots,j_s} = \hat{N}_{j_1,j_2,\ldots,j_s}$ and I is the superscript indicating the cycle of the adjustment. This procedure converges in most cases of interest to a set of estimates $\underset{\sim}{N}^*$, where

$$\begin{aligned} N^*_{+\ldots s-2,s-1,+} &= N_{+\ldots s-2,s-1,+}, \\ N^*_{+\ldots s-2,+s} &= N_{+\ldots s-2,+,s-1}, \\ N^*_{+\ldots+,s-1,s} &= N_{+\ldots+,s-1,s}, \end{aligned} \tag{7}$$

the fixed margins. These estimates maximize the multinomial likelihood equation

$$\begin{aligned} L(\underset{\sim}{N} | N_{+\ldots s-2,s-1,+}, N_{+\ldots s-2,+,s-1}, N_{+\ldots+,s-1,s}) &= \ln(\text{constant}) \\ + \sum_{\text{all } \underset{\sim}{j}} \cdots \sum \hat{N}_{j_1,j_2,\ldots,j_s} &\ln\{\pi_{j_1,j_2,\ldots,j_s}\} \\ -N(\sum_{\text{all } \underset{\sim}{j}} \cdots \sum \pi_{j_1,j_2,\ldots,j_s} &- 1), \end{aligned} \tag{8}$$

where $\pi_{j_1,j_2,\ldots,j_s} = N_{j_1,j_2,\ldots,j_s}/N$ is to be estimated subject to (7). Thus, a set of estimates $\underset{\sim}{P}^*$ are obtained from IPF, $\underset{\sim}{\pi} \hat{=} \underset{\sim}{P}^* = \frac{1}{N}\underset{\sim}{N}^*$, where "$\hat{=}$" means is estimated by.

There are two reasons for being interested in the characterization of $\underset{\sim}{P}^*$ given by (8). The first (Koch and Tolley (1975)) is that they correspond to a set of smoothed means. The corresponding equations (8) are a useful modeling tool which are computationally simple, intuitively suggestive, and have desirable large sample properties. The second reason is that (8) may well have meaning in the context of super population theory and that the only difficulty with $\underset{\sim}{N}^*$ is in obtaining estimates of the variance matrix which reflect the complex sampling plan characterized by the u_i. As will be shown, if consistent estimates of the variance matrix $\hat{\underset{\sim}{N}}$ exist, then corresponding estimates of $\underset{\sim}{N}^*$ exist. This point follows easily from the theory of implicit functions.

The constraints in (7) imply an underlying set of parameters for the multiplicative associations in the cross-classification (Fienberg (1970)). That is to say, there exists an implicit function $\underset{\sim}{F}(\beta)$ where the $\underset{\sim}{\beta}$ are a unknown set of parameters, such that

$$\begin{aligned} &\underset{\sim}{\pi} = \underset{\sim}{F}(\underset{\sim}{\beta}), \ \underset{\sim}{\beta}' = (\beta_1, \beta_2, \ldots, \beta_u), \\ &u \leq (\prod_{k=1}^{s} J_k) - 1. \end{aligned} \tag{9}$$

Thus, $\underset{\sim}{F}(\beta)$ may be substituted for π in (8) and (8) maximized under (7) by IPF. Thus, IPF solves,

$$\left[\frac{\partial}{\partial \underset{\sim}{\beta}} (\ln \underset{\sim}{F}(\underset{\sim}{\beta}))\right]' \left[\underset{\sim}{F}(\underset{\sim}{\beta}) - \hat{\underset{\sim}{p}}\right] = 0 \tag{10}$$

where $\hat{\underset{\sim}{p}} = \frac{1}{N}\hat{\underset{\sim}{N}}$. It follows that $\underset{\sim}{\beta}$ is implicitly defined in terms of $\hat{\underset{\sim}{p}}$, $\hat{\underset{\sim}{\beta}} = \underset{\sim}{H}(\hat{\underset{\sim}{p}})$. For regular functions $\underset{\sim}{F}(\beta)$, $\underset{\sim}{H}(\hat{\underset{\sim}{p}})$ is differentiable about some value of $\underset{\sim}{\pi}$, say $\underset{\sim}{\pi}_0$, and therefore can be expanded in a Taylor series about $\underset{\sim}{\pi}_0$,

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{H}(\hat{\underset{\sim}{p}}) \hat{=} \underset{\sim}{h}(\hat{\underset{\sim}{p}}) = \underset{\sim}{H}(\underset{\sim}{\pi}_0) + \left[\frac{\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}}\Bigg|_{\hat{\underset{\sim}{p}}=\underset{\sim}{\pi}_0}\right](\hat{\underset{\sim}{p}} - \underset{\sim}{\pi}_0). \tag{11}$$

From (3) it follows that $E\{\hat{\underset{\sim}{p}}\} = \underset{\sim}{\pi}_0$, so $\underset{\sim}{h}(\hat{\underset{\sim}{p}})$ is a consistent estimate of $\underset{\sim}{\beta}$ when $\underset{\sim}{\pi}_0$ obtains and the IPF model is appropriate. This permits the definition of the linearized variance of $\hat{\underset{\sim}{\beta}}$ to be,

$$\underset{\sim}{V}(\hat{\underset{\sim}{\beta}}) = \left[\frac{\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}}\Bigg|_{\hat{\underset{\sim}{p}}=\underset{\sim}{\pi}_0}\right] \underset{\sim}{V}(\hat{\underset{\sim}{p}}) \left[\frac{\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}}\Bigg|_{\hat{\underset{\sim}{p}}=\underset{\sim}{\pi}_0}\right]' \qquad (12)$$

where $\underset{\sim}{V}(\hat{\underset{\sim}{p}})$ may be any valid and consistent estimate of the covariance matrix of $\hat{\underset{\sim}{p}}$ which reflects the underlying sample design. Such estimates may be obtained by the method of Balanced Repeated Replication (BRR) as discussed by McCarthy (1966), Koch et. al. (1975), and Brock et. al. (1975). Once $\underset{\sim}{V}(\hat{\underset{\sim}{\beta}})$ is in hand another similar Taylor series expansion of $\underset{\sim}{F}(\underset{\sim}{\beta})$ yields $\underset{\sim}{V}(\underset{\sim}{p}^*)$ since (9) implies $\underset{\sim}{p}^* = \underset{\sim}{F}(\hat{\underset{\sim}{\beta}})$. Thus $\underset{\sim}{V}(\underset{\sim}{p}^*)$ is sufficient for a consistent estimate of $\underset{\sim}{V}(\underset{\sim}{p}^*)$ as was to be shown. The only difficult term is $(\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})/\partial \hat{\underset{\sim}{p}})$ which can be found by differentiating (10) by $\hat{\underset{\sim}{p}}$.

3. The Variance Structure of the IPF Model

The equation for $\underset{\sim}{V}(\hat{\underset{\sim}{\beta}})$ shows the importance of the form which $\underset{\sim}{F}(\underset{\sim}{\beta})$ takes. Specifically, for the model shown in (8) which IPF maximized, $\underset{\sim}{F}(\underset{\sim}{\beta})$ takes a log-linear form,

$$\underset{\sim}{\pi} = \exp\{\underset{\sim}{X}\ \underset{\sim}{\beta}\}\ , \qquad (13)$$

where $\underset{\sim}{X}$ is a $((\prod_{j_k=1}^{s} J_k) \times u)$ design matrix which corresponds to a complete factorial design (Fienberg (1970b)), with additive constraints. For example, if there are three binomial responses, then

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix}, \text{ and } \underset{\sim}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} \qquad (14)$$

If β_1 through β_8 are estimated, this is equivalent to not adjusting the data. If the total population size is known then β_1 is fixed. If the first order margins are known then β_2, β_3, and β_4 are fixed. If the second order margins are known then β_5, β_6, and β_7 are fixed. If the third order margin is known then the sample need not have been drawn. These parameters correspond to various levels of multiplicative association. The usual level of adjustment is done in the reverse of the order of model reduction in the linear models context. That is, the highest order terms are the last to be eliminated.

Returning to the $\underset{\sim}{V}(\hat{\underset{\sim}{\beta}})$ implied by (13) the solution for $[\partial \underset{\sim}{H}(\underset{\sim}{p})/\partial \underset{\sim}{p}]$ follows the method of proof given in Koch and Tolley (1975).

Let,

$$F^{(1)}_{i,j}(\hat{\underset{\sim}{\beta}}) = \left[\frac{\partial}{\partial \beta_j}(F_i(\underset{\sim}{\beta}))\right]\Bigg|_{\underset{\sim}{\beta}=\hat{\underset{\sim}{\beta}}} \qquad (15)$$

where $i = 1,\ldots,r = \prod_{k=1}^{s} J_k$, $j = 1,\ldots,u$, and

$$F^{(2)}_{i,j}(\hat{\underset{\sim}{\beta}}) = \left[\frac{\partial^2}{\partial \beta_j^2}(F_i(\underset{\sim}{\beta}))\right]\Bigg|_{\underset{\sim}{\beta}=\hat{\underset{\sim}{\beta}}}. \qquad (16)$$

Noting that $\sum_{i=1}^{r} F_i(\underset{\sim}{\beta}) = 1$ implies $\sum_{i=1}^{r} F^{(1)}_{i,j}(\underset{\sim}{\beta}) = \sum_{i=1}^{r} F^{(2)}_{i,j}(\underset{\sim}{\beta}) = 0$. Then

$$\left[\frac{\partial H_j(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}_i}\right]\Bigg|_{\hat{\underset{\sim}{p}}_i=\underset{\sim}{\pi}_i} = \frac{F^{(1)}_{i,j}(\underset{\sim}{\beta})/F_i(\underset{\sim}{\beta})}{\sum_{i=1}^{r}[F^{(1)}_{i,j}(\underset{\sim}{\beta})]^2/F_i(\underset{\sim}{\beta})} \qquad (17)$$

for $i = 1,\ldots,r$; $j = 1,\ldots,u$.

Substituting (15) and (16) into (13) yields:

$$F^{(1)}_{i,j}(\hat{\underset{\sim}{\beta}})/F_i(\hat{\underset{\sim}{\beta}}) = x_{ij}\ ,$$
$$[F^{(1)}_{i,j}(\hat{\underset{\sim}{\beta}})]^2/F_i(\hat{\underset{\sim}{\beta}}) = x_{ij}p^*_j\ . \qquad (18)$$

These in turn yield

$$\frac{\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}}\Bigg|_{\hat{\underset{\sim}{p}}=\underset{\sim}{p}^*}$$

$$= \begin{bmatrix} \frac{x_{11}}{\sum_{i=1}^{r} x_{i1}^2 p_i^*} & \frac{x_{21}}{\sum_{i=1}^{r} x_{i1}^2 p_i^*} & \cdots & \frac{x_{r1}}{\sum_{i=1}^{r} x_{i1}^2 p_i^*} \\ \frac{x_{12}}{\sum_{i=1}^{r} x_{i2}^2 p_i^*} & \frac{x_{22}}{\sum_{i=1}^{r} x_{i2}^2 p_i^*} & \cdots & \frac{x_{r2}}{\sum_{i=1}^{r} x_{i2}^2 p_i^*} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{1u}}{\sum_{i=1}^{r} x_{iu}^2 p_i^*} & \frac{x_{2u}}{\sum_{i=1}^{r} x_{iu}^2 p_i^*} & \cdots & \frac{x_{ru}}{\sum_{i=1}^{r} x_{iu}^2 p_i^*} \end{bmatrix} = G(p^*) \qquad (19)$$

When $J_k = 2$, $k = 1,\ldots,s$ then $x_{ij}^2 = 1$ all i,j so

$$\left[\frac{\partial \underset{\sim}{H}(\hat{\underset{\sim}{p}})}{\partial \hat{\underset{\sim}{p}}}\right]\Bigg|_{\hat{\underset{\sim}{p}}=\underset{\sim}{p}^*} = \underset{\sim}{X}' \qquad (20)$$

So returning to the original problem of approximating $\underset{\sim}{V}(\underset{\sim}{p}^*)$:

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{H}(\hat{\underset{\sim}{p}}) , \qquad (21)$$

$$\underset{\sim}{V}(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{G}(\underset{\sim}{p}^*)\ \underset{\sim}{V}(\hat{\underset{\sim}{p}})[\underset{\sim}{G}(\underset{\sim}{p}^*)]' , \qquad (22)$$

$$\underset{\sim}{p}^* = \underset{\sim\sim}{\exp}\{\underset{\sim}{X}\ \hat{\underset{\sim}{\beta}}\} , \qquad (23)$$

and

$$\underset{\sim}{V}(\underset{\sim}{p}^*) = \underset{\sim}{D}_{\underset{\sim}{p}^*}\ \underset{\sim}{G}'\underset{\sim}{G}\ \underset{\sim}{V}(\hat{\underset{\sim}{p}})\ \underset{\sim}{G}'\underset{\sim}{G}\ \underset{\sim}{D}_{\underset{\sim}{p}^*} , \qquad (24)$$

where $\underset{\sim}{G} = \underset{\sim}{G}(\underset{\sim}{p}^*)$ is defined in (19) and $\underset{\sim}{D}_{\underset{\sim}{p}^*}$ is a diagonal matrix with the IPF estimates on the main diagonal. As noted in Section 2, $\underset{\sim}{V}(\hat{\underset{\sim}{p}})$ may be obtained from BRR or any other method which yields consistent estimates.

Given these estimates of the variance matrix of $\underset{\sim}{p}^*$ it is desirable to find an overall measure of the variance reduction achieved by IPF. One approach is suggested by the use of weighted least squares as discussed in Koch et. al. (1975). Let $\hat{\underset{\sim}{p}}_1$ be the vector of sample estimates with the r-th element deleted and $\underset{\sim}{V}_1(\hat{\underset{\sim}{p}}_1)$ be the corresponding covariance matrix. Similarly define $\underset{\sim}{p}^*_u$ and $\underset{\sim}{V}_u(\underset{\sim}{p}^*_u)$ as the vector of adjusted estimates with the last r - u elements deleted and the corresponding covariance matrix. Then if $\underset{\sim}{\ell}_1 = (\hat{\underset{\sim}{p}}_1 - \underset{\sim}{1}\ b_1)$ where b_1 is estimated by weighted least squares and $\underset{\sim}{1}$ is a vector of (r - 1) 1's, then

$$Q_1 = \underset{\sim}{\ell}_1'\ \underset{\sim}{V}_1^{-1}(\underset{\sim}{p}_1)\ \underset{\sim}{\ell}_1 \qquad (25)$$

is a measure of the total variation in the unadjusted cross-classification. Similarly,

$$Q_u = \underset{\sim}{\ell}_u'\ \underset{\sim}{V}_u^{-1}(\underset{\sim}{p}^*_u)\ \underset{\sim}{\ell}_u \qquad (26)$$

is a measure of the total variation in the adjusted cross-classification. For large data sets in the super-population context Q_1 and Q_u have χ^2 distributions with r - 1 and u - 1 degrees of freedom respectively. Moreover, these form valid indices of the overall variance reduction achieved by IPF.

4. Application to Poststratification

Some sample surveys are conducted in such a way that for single margins involving several variables complete information about the target population is available. For example, in a survey to determine health characteristics, the demographic characteristics for each of the survey strata may be independently determined. Thus, the ratio of the known variables to their estimated values may be used to inflate the strictly estimated survey variables. This process is known as poststratification. Frequently, this inflation takes the form of a simple multiplicative adjustment.

The poststratification procedure is in fact a simplified form of the IPF procedure discussed in Sections 2 and 3. What is done is to array the demographic variables as a single variable with a number of levels equal to the product of the number of levels in each of the known variables. Thus, the poststratification corresponds to a single iteration of the IPF algorithm. It immediately follows from the discussion in Section 3 that the Q statistic is increased to the extent of the elimination of sampling variance and measurement error. There is a corresponding reduction in the standard errors. A cautionary note should be observed in that poststratification, as with the IPF procedure, will induce bias in the estimate to the extent of error in the known variables.

The important point is that when poststratification, or for that matter IPF, is contemplated the Q statistics and their corresponding ratios may be computed to estimate the relative improvement in precision. Thus, assessments may be made and weighed against the possibility of inducing bias in the estimation process.

5. Complex Postsampling Models

Subsequent to the selection and analysis of a sample survey, substantive analysts often become aware of the possibility that the responses were in fact generated by an underlying stochastic process. These processes usually can be formalized into some type of statistical function. Two such functions are the multiple logit and the Weibull distributions. Each of these distributions have found wide acceptance in the biological and social sciences because of their flexibility and interpretability. The fitting of these distributions for survey data may easily be implemented in the framework of the CRE model as will be demonstrated. Here the multiple logit is considered in the context of multiple binomials and the Weibull is considered in the context of a bivariate distribution for a cross-sectional representative sample.

Consider the situation where the survey responses are classified by two levels for each of three factors. This is again a 2^3 design on the response space. Let π_{ijk} denote the probability of a response at joint levels (i,j,k) where i = 1,2, j = 1,2, k = 1,2; and

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\pi_{ijk} = 1 . \qquad (27)$$

Then the multiple logit model is:

$$\ln \pi_{ijk} = x^{(1)}_{ijk}\beta_1 + x^{(2)}_{ijk}\beta_2 + x^{(3)}_{ijk}\beta_3 + x^{(4)}_{ijk}\beta_4 + x^{(5)}_{ijk}\beta_5 + x^{(6)}_{ijk}\beta_6 - \ln D \quad (28)$$

where

$$x^{(1)}_{ijk} = 1 \quad k = 2,\ 0 \quad \text{otherwise};$$

$$x^{(2)}_{ijk} = 1 \quad j = 2,\ 0 \quad \text{otherwise};$$

$$x^{(3)}_{ijk} = 1 \quad i = 2,\ 0 \quad \text{otherwise};$$

$$x^{(4)}_{ijk} = x^{(2)}_{ijk}\, x^{(3)}_{ijk}$$

$$x^{(5)}_{ijk} = x^{(2)}_{ijk}\, x^{(4)}_{ijk}$$

$$x^{(6)}_{ijk} = x^{(3)}_{ijk}\, x^{(4)}_{ijk}$$

and

$$D = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \exp\{\sum_{r=1}^{6} x^{(r)}_{ijk}\beta_r\}.$$

Implicit in the model is the constraint (27) and the eighth degree of freedom is reserved for a goodness of fit test. The formulation of Sections 2 and 3 in terms of models which correspond to maximum likelihood functions may also be used. The essential point is the use of maximum likelihood search procedures to estimate the β_r.

The estimation of standard errors again requires the evaluation of

$$\frac{\partial}{\partial \underset{\sim}{\beta}} \{\ln[F(\underset{\sim}{\beta})]\} \quad (29)$$

where,

$$\ln F_{ijk}(\underset{\sim}{\beta}) = \sum_{r=1}^{6} x^{(r)}_{ijk}\beta_r - \ln D. \quad (30)$$

In matrix notation the solutions may be written as in (10).

The covariance matrix of the parameters may now be obtained from (17) and (12). In particular, after some algebra

$$\left.\frac{\delta H(p)}{\delta p}\right|_{p=\pi} = \begin{bmatrix} \frac{-1}{\pi_{\cdot\cdot 1}} & \frac{1}{\pi_{\cdot\cdot 2}} & \frac{-1}{\pi_{\cdot\cdot 1}} & \frac{1}{\pi_{\cdot\cdot 2}} & \frac{-1}{\pi_{\cdot\cdot 1}} & \frac{1}{\pi_{\cdot\cdot 2}} & \frac{-1}{\pi_{\cdot\cdot 1}} & \frac{1}{\pi_{\cdot\cdot 2}} \\ \frac{-1}{\pi_{\cdot 1\cdot}} & \frac{-1}{\pi_{\cdot 1\cdot}} & \frac{1}{\pi_{\cdot 2\cdot}} & \frac{1}{\pi_{\cdot 2\cdot}} & \frac{-1}{\pi_{\cdot 1\cdot}} & \frac{-1}{\pi_{\cdot 1\cdot}} & \frac{1}{\pi_{\cdot 2\cdot}} & \frac{1}{\pi_{\cdot 2\cdot}} \\ \frac{-1}{\pi_{1\cdot\cdot}} & \frac{-1}{\pi_{1\cdot\cdot}} & \frac{-1}{\pi_{1\cdot\cdot}} & \frac{-1}{\pi_{1\cdot\cdot}} & \frac{1}{\pi_{2\cdot\cdot}} & \frac{1}{\pi_{2\cdot\cdot}} & \frac{1}{\pi_{2\cdot\cdot}} & \frac{1}{\pi_{2\cdot\cdot}} \\ \frac{-1}{1-\pi_{\cdot 22}} & \frac{-1}{1-\pi_{\cdot 22}} & \frac{-1}{1-\pi_{\cdot 22}} & \frac{1}{\pi_{\cdot 22}} & \frac{-1}{1-\pi_{\cdot 22}} & \frac{-1}{1-\pi_{\cdot 22}} & \frac{-1}{1-\pi_{\cdot 22}} & \frac{1}{\pi_{\cdot 22}} \\ \frac{-1}{1-\pi_{2\cdot 2}} & \frac{-1}{1-\pi_{2\cdot 2}} & \frac{-1}{1-\pi_{2\cdot 2}} & \frac{-1}{1-\pi_{2\cdot 2}} & \frac{-1}{1-\pi_{2\cdot 2}} & \frac{1}{\pi_{2\cdot 2}} & \frac{-1}{1-\pi_{2\cdot 2}} & \frac{1}{\pi_{2\cdot 2}} \\ \frac{-1}{1-\pi_{22\cdot}} & \frac{-1}{1-\pi_{22\cdot}} & \frac{-1}{1-\pi_{22\cdot}} & \frac{-1}{1-\pi_{22\cdot}} & \frac{-1}{1-\pi_{22\cdot}} & \frac{-1}{1-\pi_{22\cdot}} & \frac{1}{\pi_{22\cdot}} & \frac{1}{\pi_{22\cdot}} \end{bmatrix} \quad (31)$$

Thus, as with the IPF procedure, $\hat{\underset{\sim}{\beta}}$ and $V(\hat{\underset{\sim}{\beta}})$ may be obtained by substituting $\hat{\underset{\sim}{p}}$ for $\underset{\sim}{\pi}$ in (28). This in turn yields the estimates of $\underset{\sim}{p}^*$ and $\underset{\sim}{V}(\underset{\sim}{p}^*)$ under the model, where $\underset{\sim}{p}^*$ refers to the logit model proportions.

An asymptotically equivalent procedure is that of weighted least squares. The formulation is as follows. Let

$$\underset{\sim}{K}\ \underset{\sim\sim}{\ln}\ \underset{\sim}{A}\ \underset{\sim}{\pi} \triangleq \underset{\sim}{X}\ \underset{\sim}{\beta} \quad (32)$$

where

$$\underset{\sim}{A} = \underset{\sim 8}{I}\ ,\ \underset{\sim}{K} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \underset{\sim}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}$$

Then $\underset{\sim}{\beta}$ may be estimated directly from weighted least squares procedures and corresponding estimates of $\underset{\sim}{\pi}$ obtained by inverting (32). Again, the only critical aspect is in obtaining $\underset{\sim}{V}(\hat{\underset{\sim}{p}})$ from some procedure such as BRR. Thus, it is apparent that functions with a simple exponential form, such as multiple logits, may be fitted to complex survey data. The bivariate Weibull distribution represents a more complex function.

The Weibull distribution has been of interest to statisticians since its introduction to the literature, partly because of its usefulness in data which are thought to reflect either increasing or decreasing hazard functions. If t represents the time to the occurrence of an event of interest (e.g., a death or the detection of a tumor), then the Weibull cumulative distribution function may be written as:

$$G(t|\mu,\delta,w) = 1 - \exp\{-\mu(t-w)^{\delta}\} \quad \text{for } t \geq w \text{ and } \mu,\delta \geq 0. \quad (33)$$

As demonstrated in Freeman, Freeman, and Koch (1974) such models may easily be fitted in the usual weighted least squares framework. This may be seen by examining $\theta(t)$ which denotes the log-log of the survival rate. Specifically let

$$\theta(t) = \ln[-\ln\{1 - G(t|\mu,\delta,w)\}] = \ln(\mu) + \delta \ln(t-w), \quad (34)$$

where $\ln(\mu)$ and δ are to be estimated and w is to be fixed by the experimental situation. With this reparameterization $\theta(t)$ can easily be fitted, both for univariate and bivariate data.

Typically, the Weibull distribution is fitted to data which is assumed to arise from a simple random sample. However, the CRE model and the variance estimation procedure of BRR permit the fitting of this distribution to data arising from a complex sample survey. The only change would be instead of estimating $\widehat{Var}(\hat{p})$ based on the multinomial model assumption, $V(\hat{p})$ would be based on the Balanced Repeated Replication procedure.

6. Summary and Conclusions

This paper was concerned with various post-sampling adjustments to complex survey data and the fitting of complex functions to such data. The IPF (iterative proportional fitting) procedure was investigated and an approximation to the variance structure induced by the procedure was developed. Further, a statistic, Q, was suggested for evaluating the overall variance reduction achieved by the procedure. This statistic is particularly useful since it is a natural by-product of the inference structure discussed in Koch et. al. (1975).

The discussion of IPF was followed by a discussion of the procedure for fitting more complex distributions to data. The multiple logit and bivariate Weibull distributions were examined in this context. It was pointed out that once the distribution of interest had been reparameterized into the linear model framework, the only change from the simple random sampling inference structure was in terms of estimating the variance matrix. This was shown to be easily accomplished when BRR is used. Thus, it follows that the response error model in conjunction with the BRR permits the implementation of the weighted least squares methodology for inference in the post-sampling analysis of complex sample survey data.

ACKNOWLEDGMENTS

This research was in part supported by the National Institutes of Health (Grants GM-70004 and HD-00371) and by the U. S. Bureau of the Census through Joint Statistical Agreements JSA 74-2 and JSA 75-2.

REFERENCES

Brock, D.B., Freeman, D.H., Freeman, J.L. and Koch, G.G. (1975). An application of categorical data analysis to the National Health Interview Survey, Proceedings of the 1975 Social Statistics Section of the ASA.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, Ann. Math. Stat. 11, 427-444.

Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables, Ann. Math. Stat. 41, 907-918.

Freeman, D.H. (1975). The regression analysis of data from complex sample surveys: an empirical investigation of covariance matrix estimation, Institute of Stat. Mimeo Series No. 1020, Chapel Hill, N. C.

Freeman, D.H., Freeman, J.L., and Koch, G.G. (1974). A modified χ^2 approach for fitting Weibull models to synthetic life tables, Institute of Stat. Mimeo Series No. 958, Chapel Hill, N. C.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance, in Contributions to Statistics, Pergamon Press Ltd., London, 111-136.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys, Int. Stat. Rev. 43, 55-74.

Koch, G.G. And Tolley, H.D. (1975). A generalized modified χ^2 analysis of categorical bacterial survival data from a complex dilution experiment, Biometrics 31, 59-92.

McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys, Vital and Health Statistics, P.H.S. Pub. No. 1000, Ser. 2, No. 14, National Center for Health Statistics, Rockville, Md.

Mosteller, F. (1968). Association and estimation in contingency tables, J.A.S.A. 63, 1-28.