AN ASYMPTOTIC COVARIANCE STRUCTURE FOR TESTING HYPOTHESES ON RAKED CONTINGENCY TABLES FROM COMPLEX SAMPLE SURVEYS

Daniel H. Freeman, Jr., Yale University and Gary G. Koch, U. of North Carolina

0. Introduction

Recently, there have been substantial breakthroughs in the analysis of cross-classified frequency counts. One of the major algorithms for this type of analysis is variously known as iterative proportional fitting (IPF), raking, or the Deming and Stephan algorithm. The original version of the algorithm was discussed by Deming [1943]. Subsequently, there have been a number of papers and books on the techniques which utilize the algorithm (e.g. Bishop et al. [1975]). The aspect of the algorithm to be discussed in this paper is its application to complex survey data. It is assumed that such a random sample has a known probability structure.

Social and life scientists have been applying the algorithm to complex survey data for several years. Two examples are papers by Frederick J. Scheurin [1973] and Robert M. Hauser et al. [1975]. Scheurin states,

> "...when the sample size is large relative to the number of cells then substantively insignificant effects can become statistically significant. It also turns out to be quite difficult to make even approximate significance statements when data come from complex multistage samples..." (p.164).

This was stated in the context of a discussion concerning the use of log-linear models for purpose of generating hypotheses from a Current Population Survey data set. The focus was on poverty statistics. In the second example, Hauser et al. adjusted the "Occupational Changes in a Generation" data set downward by factor of 0.62 to reflect the efficiency of the survey design relative to simple random sampling (Hauser et al. [1975]: 282). However, they point out there should be additional adjustments for non-sampling error and simultaneous inference. Thus, the application of log-linear models to survey data requires dealing with two issues:

i. the relatively large size of the samples;
ii. the complexity of the survey design.

It is true that the size of the sample often leads to the statistical significance of largely uninterpretable interaction effects. However, this is not necessarily the case. For example, Freeman et al. [1977] discusses the fitting of relatively simple models to physician visit data from the National Health Interview Survey. This was a survey of about 40,000 households or 120,000 individuals. Moreover, the survey design was indirectly incorporated into that study. It is this issue which must be considered prior to deciding that the formal hypothesis testing as unnecessary because of the sample size.

The survey design is frequently too complex for the use of direct estimates of variance. However, for linear sample statistics techniques such as jack-knifing and pseudo-replication may be used to generate valid estimated covariance matrices. For "raked tables" the sample estimates are generally non-linear. Causey [1973] pointed out that Taylor series estimates are feasible for such tables. The tables are raked so as to minimize the "discrimination information." This is discussed for the simple survey situation in a number of places including Gokhale and Kullback [1976].

This paper shows that the problem is in fact a direct application of the "Functional Asymptotic Regression Methodology," Koch et al. [1975]. The key assumption is that the central covariance matrices are estimated either directly or indirectly by some method which accounts for the survey design. Previously, Koch et al. examined the problem where it could be assumed that the data were based on independent simple random samples. As noted in Freeman et al. [1977] the violation of this assumption in complex surveys can result is substantial reductions in the power of the test statistics. The discussion is in three parts. First, a general survey notation is presented. The raking model and its covariance matrix estimates are discussed. Lastly, an example is given.

1. Notation

Consider a set of d attributes. Let $j_g = 1,2,...,L_g$ index the response categories for the g-th attribute where $g = 1,2,...,d$. Let $\underset{\sim}{j} = (j_1,j_2,...,j_d)$ denote the vector response profile. Let

$$N_{\underset{\sim}{j}\ell} = \begin{cases} 1 & \text{if element } \ell \text{ from population under study is classified as having response profile } \underset{\sim}{j}=(j_1,...,j_d) \\ 0 & \text{if otherwise} \end{cases} \qquad (1.1)$$

where $\ell = 1,2,...,N$ with N being the total number of elements in the population. Let

$$U_\ell = \begin{cases} 1 & \text{if element } \ell \text{ from population is in sample} \\ 0 & \text{if otherwise} \end{cases} \qquad (1.2)$$

The $\{U_\ell\}$ characterize the sample design including the nature of any clustering, stratification, and/or multistage selection. Let $\phi_\ell = E\{U_\ell\}$ denote probability of selection for element ℓ from population. Let $n = \sum_{\ell=1}^{N} \phi_\ell$ denote the sample size.

The multivariate relationships among the d attributes can be summarized in terms of the d-dimensional contingency table of weighted frequencies

$$\hat{N}_{\underset{\sim}{j}} = \hat{N}_{j_1 j_2 \ldots j_d} = \sum_{\ell=1}^{N} \frac{1}{\phi_\ell} U_\ell N_{j_1 j_2 \ldots j_d, \ell}. \tag{1.3}$$

In this framework, let $p_{\underset{\sim}{j}} = (\hat{N}_{\underset{\sim}{j}}/N)$ denote the corresponding relative frequency or weighted observed sample proportion. The $p_{\underset{\sim}{j}}$ are unbiased estimators for the parameters

$$\begin{aligned}\pi_{\underset{\sim}{j}} = \pi_{j_1 j_2 \ldots j_d} &= \frac{1}{N} E\{\hat{N}_{j_1 j_2 \ldots j_d}\} \\ &= \frac{1}{N} \sum_{\ell=1}^{N} N_{j_1 j_2 \ldots j_d, \ell}\end{aligned} \tag{1.4}$$

which reflect the average distribution of the respective response profiles in the population.

Let $\hat{\underset{\sim}{N}}$, $\underset{\sim}{p}$, and $\underset{\sim}{\pi}$ denote the vectors

$$\hat{\underset{\sim}{N}} = \begin{bmatrix} \hat{N}_{11\ldots1} \\ \ldots \\ \hat{N}_{L_1 L_2 \ldots L_d} \end{bmatrix}, \quad \underset{\sim}{p} = \begin{bmatrix} p_{11\ldots1} \\ \ldots \\ p_{L_1 L_2 \ldots L_d} \end{bmatrix}$$

$$\underset{\sim}{\pi} = \begin{bmatrix} \pi_{11\ldots1} \\ \ldots \\ \pi_{L_1 L_2 \ldots L_d} \end{bmatrix} \tag{1.5}$$

2. Methods of Adjustment

A sample from a specific population may be regarded as yielding two types of information:

A. Estimators for marginal distributions of certain subsets of attributes. This type of information is called "allocation structure."

B. Estimators for higher order measures of association and/or interaction which reflect the relationships across the marginal subsets in (A). This type of information is called "association structure."

With these considerations in mind, the observed sample can be adjusted to provide estimators for other target populations of interest if the following assumptions hold:

i. the target population has KNOWN "allocation structure" via census or other sample survey data

ii. the target population has the SAME "association structure" as the sampled population.

Examples of such target populations include:

a. various local (county or state) subdivisions of a nationally sampled population

b. other local, national, or international target populations which may partially overlap a sampled local population.

More specifically, let $\underset{\sim}{\pi}$ denote the parameter vector which characterizes the distribution of the response profiles for the sampled population, and let $\underset{\sim}{p}$ denote its corresponding estimator. Let $\underset{\sim}{\pi}_T$ denote the parameter vector for the target population. Let $\underset{\sim}{A}_T$ denote a matrix of coefficients whose columns generate the pertinent marginal distributions comprising the known "allocation structure," and let $\underset{\sim}{\xi}_T$ denote their corresponding known values. Thus, assumption (i) means that $\underset{\sim}{\pi}_T$ satisfies

$$\underset{\sim}{A}_T' \underset{\sim}{\pi}_T = \underset{\sim}{\xi}_T \tag{2.1}$$

where without loss of generality, $\underset{\sim}{A}_T'$ will be regarded as having full rank by deletion of unnecessary rows. The matrix $\underset{\sim}{A}_T'$ also reflects the fact that the elements of $\underset{\sim}{\pi}_T$ satisfy the constraint

$$\underset{\sim}{1}' \underset{\sim}{\pi}_T = 1 \tag{2.2}$$

where $\underset{\sim}{1}'$ is a row vector of 1's. Given the formulation (2.1) of "allocation structure," attention will be directed at the asymptotic covariance structure of the estimator for the parameter vector $\underset{\sim}{\pi}_T$ which is obtained by applying assumption (ii) with respect to an appropriate definition of "association structure."

2.1. Adjustment with respect to complete "association structure"

Let $\underset{\sim}{K}$ denote an ortho-complement matrix to $\underset{\sim}{A}_T$. Then assumption (ii) means that $\underset{\sim}{\pi}_T$ satisfies

$$\underset{\sim}{K}'\{\underset{\sim}{\log}_e(\underset{\sim}{\pi}_T)\} = \underset{\sim}{K}'\{\underset{\sim}{\log}_e(\underset{\sim}{\pi})\} \tag{2.3}$$

where, in this context, "association structure" is formulated in terms of log-linear contrast functions. If $\underset{\sim}{p}$ denotes the sample estimator of $\underset{\sim}{\pi}$ defined by (1.5), then (2.1) and (2.3) imply that the marginal adjustment (raking) estimator $\hat{\underset{\sim}{\pi}}_T$ of $\underset{\sim}{\pi}_T$ is characterized by the equations

$$\underset{\sim}{A}_T' \hat{\underset{\sim}{\pi}}_T = \underset{\sim}{\xi}_T \tag{2.4}$$

$$K'\{\log_e(\hat{\pi}_T)\} = K'\{\log_e(p)\}. \quad (2.5)$$

Within this framework, the estimator $\hat{\pi}_T$ may be determined (provided both assumptions (i) and (ii) are true so that a solution π_T to (2.1) and (2.3) almost always exists if the sample size n is sufficiently large) by applying the Deming-Stephan Iterative Proportional Fitting (IPF) algorithm to adjust an initial estimator which satisfies (2.5) to conform successively with each of the respective marginal configurations which comprise the "allocation structure" equations (2.4) since such operations preserve the agreement of successive solutions with the "association structure" equations (2.5).

The asymptotic covariance matrix of the estimators $\hat{\pi}_T$ is obtained by the well-known δ-method (based on the first order Taylor series) with the required first derivative matrix being determined by implicit techniques. In this regard, if both sides of (2.4) - (2.5) are differentiated with respect to p, it follows that

$$\begin{bmatrix} A_T' \\ K'D_{\hat{\pi}_T}^{-1} \end{bmatrix} \begin{bmatrix} \dfrac{d\hat{\pi}_T}{dp} \end{bmatrix} = \begin{bmatrix} 0 \\ K'D_p^{-1} \end{bmatrix} \quad (2.6)$$

where D_y denotes a diagonal matrix with the elements of y on the diagonal. Since

$$\begin{bmatrix} A_T' \\ K'D_{\hat{\pi}_T}^{-1} \end{bmatrix}^{-1} = \{D_{\hat{\pi}_T}A_T[A_T'D_{\hat{\pi}_T}A_T]^{-1}, K[K'D_{\hat{\pi}_T}K]^{-1}\} \quad (2.7)$$

the equations (2.6) may be solved to yield

$$\left[\dfrac{d\hat{\pi}_T}{dp} \,\middle|\, p=\pi \right] = K[K'D_{\pi_T}^{-1}K]^{-1}K'D_{\pi}^{-1} \quad (2.8)$$

Thus, if $V(\pi)$ denotes the covariance matrix of the estimator p, then the asymptotic covariance matrix for $\hat{\pi}_T$ is given by

$$V_{\hat{\pi}_T}(\pi) = K[K'D_{\pi_T}^{-1}K]^{-1}K'D_{\pi}[V(\pi)]D_{\pi}^{-1}K[K'D_{\pi_T}^{-1}K]^{-1}K'. \quad (2.9)$$

For univariate problems, a similar result is found in Causey (1972).

If the sample design is simple random sampling (with replacement), then

$$V(\pi) = \frac{1}{n}\{D_{\pi} - \pi\pi'\}. \quad (2.10)$$

Thus, for this special case, the covariance matrix (2.9) may be simplified to

$$V_{\hat{\pi}_T}(\pi) = K[K'D_{\pi_T}^{-1}K]^{-1}K'D_{\pi}^{-1}K[K'D_{\pi_T}^{-1}K]^{-1}K'/n \quad (2.11)$$

Moreover, if the target population is identical to the sampled population (which is the case when the "allocation structure" of the population under study is known a priori as is the case with samples from registration systems like licensed drivers), then π_T and π are identical so that (2.11) may be further simplified to

$$V_{\hat{\pi}_T}(\pi) = K[K'D_{\pi}^{-1}K]^{-1}K'/n \quad (2.12)$$

An analogous simplification could also be applied to the more general result (2.9) for this situation.

Reasonable estimators for the covariance matrix $V_{\hat{\pi}_T}(\pi)$ may be constructed by replacing $V(\pi)$ by an appropriate consistent estimator V_p which is obtained by either direct or replication methods and replacing π and π_T by p and $\hat{\pi}_T$ respectively. The resulting estimated covariance matrix $V_{\hat{\pi}_T}$ may be used in conjunction with $\hat{\pi}_T$ to test various hypotheses by weighted least squares methods. In this regard, an appropriate test statistic for the hypothesis

$$H_0: \; C\pi_T = 0, \quad (2.13)$$

where C is assumed to be a full rank matrix, is the Wald statistic

$$Q_C = \hat{\pi}_T' C' [CV_{\hat{\pi}_T}C']^{-1} C\hat{\pi}_T \quad (2.14)$$

which has approximately a chi-square distribution with D.F. = Rank(C) in large samples.

2.2. Adjustment with respect to reduced "association structure"

For certain situations, it may be possible to assume that the vector π is characterized by a log-linear model

$$\pi = \pi(\beta) = \{\exp(X\beta)\}/\{1'[\exp(X\beta)]\}, \quad (2.15)$$

where X is a known full rank design matrix whose columns represent a basis for the main effects and interactions which constitute the model and β is an unknown parameter vector. When the model (2.15) holds, a reasonable estimator for π may be obtained by solving the equations

$$X'[\pi(\hat{\beta})] = X'p \ . \tag{2.16}$$

If the matrix X has an hierarchical structure which includes with any given interaction variable all other interaction variables of the same type and all corresponding lower order interactions, the equations (2.16) may be solved by applying the Deming-Stephan IPF algorithm to adjust an initial estimator

$$\hat{\pi}_0 = \prod_{g=1}^{d} L_g^{-1}\{1\} \tag{2.17}$$

which trivially satisfies the model (2.15) to conform successively with each of the respective marginal configurations which are associated with the equations (2.16).

The asymptotic covariance matrix of the estimator $\hat{\beta}$ of β is obtained by the δ-method with the first derivative matrix being determined by implicit techniques. In this regard, if both sides of (2.16) are differentiated with respect to p, it follows that

$$X' \frac{d}{dp} \frac{\exp(X\beta)}{1'\{\exp(X\beta)\}} = X'$$

$$\tag{2.18}$$

$$X'[D_{\hat{\pi}} - \hat{\pi}\hat{\pi}']X \ \frac{d\beta}{dp} = X'$$

where $\hat{\pi} = \{\exp(X\hat{\beta})/1'[\exp(X\hat{\beta})]\}$. The equations (2.18) may be solved to yield

$$\left.\frac{d\hat{\beta}}{dp}\right|_{p=\pi} = \{X'[D_\pi - \pi\pi']X\}^{-1}X' . \tag{2.19}$$

Thus, the asymptotic covariance matrix $V_{\hat{\beta}}(\pi)$ for $\hat{\beta}$ is given by

$$V_{\hat{\beta}}(\pi) = \{X'[D_\pi - \pi\pi']X\}^{-1}X'V(\pi)X\{X'[D_\pi - \pi\pi']X\}^{-1} \tag{2.20}$$

If the sample design is simple random sampling, then (2.10) may be used to simplify (2.20) to

$$V_{\hat{\beta}}(\pi) = \frac{1}{n}\{X'[D_\pi - \pi\pi']X\}^{-1} . \tag{2.21}$$

Moreover, it can be noted that $\hat{\beta}$ represents the maximum likelihood estimator of β for this situation.

A reasonable estimator for $V_\beta(\pi)$ may be constructed by replacing $V(\pi)$ by an appropriate estimator V_p as described previously and replacing π by $\hat{\pi}$. The resulting estimated covariance matrix V_β together with $\hat{\beta}$ provide a framework for further analysis by weighted least squares methods.

The results (2.15) - (2.21) may be applied to the marginal adjustment (raking) situation by noting that (2.15) implies that the "association structure" equations (2.3) may be written as

$$\begin{bmatrix} K' \\ X_C' \end{bmatrix} \{\log_e(\pi_T)\} = \begin{bmatrix} K' \\ X_C' \end{bmatrix} \{\log_e(\pi)\} = \begin{bmatrix} K'X\beta \\ 0 \end{bmatrix} \tag{2.22}$$

where X_C is an ortho-complement matrix to $[A_T, X]$ and K is an ortho-complement matrix to $[A_T, X_C]$. By proceeding as described in (2.4) - (2.9), it can be verified that the asymptotic covariance matrix for $\hat{\pi}_T$ under the model (2.15) is given by

$$V_{\pi_T}(\beta) = H[V_\beta(\pi)]H' \tag{2.23}$$

where

$$H = [K, X_C] \begin{bmatrix} K'D_{\pi_T}^{-1}K \ , & K'D_{\pi_T}^{-1}X_C \\ X'D_{\pi_T}^{-1}K \ , & X_C'D_{\pi_T}^{-1}X_C \end{bmatrix}^{-1} \begin{bmatrix} K'X \\ 0 \end{bmatrix}$$

Further extensions of these results may be undertaken by allowing ξ_T to be replaced by an estimator $\hat{\xi}_T$ which is either independent or correlated with p.

3. Example

The data in Table 1 have been used by Ireland and Kullback [1968; pp. 707-713] to illustrate the application of IPF for the adjustment of a contingency table to a known marginal "allocation structure." They are being reanalyzed here to indicate the reduction in variance which is achieved by using such "raking" procedures to estimate the cell probabilities π.

These data originally come from a study undertaken by Roberts et al. [1939; Biometrika, 31, pp. 56-66]. The experimental design involves $n = 3734$ mice from a single population, each of which is classified with respect to the presence or absence of the attributes A, B, and D. The "allocation structure" of interest is defined in terms of the hypothesis that the probability of the presence (or absence) of each separate attribute is (1/2). Thus, with respect to the matrix notation in (2.1), it follows that

$$\underset{\sim}{A}_T' = \begin{bmatrix} 1&1&1&1&1&1&1&1 \\ 1&1&1&1&0&0&0&0 \\ 1&1&0&0&1&1&0&0 \\ 1&0&1&0&1&0&1&0 \end{bmatrix}, \quad \underset{\sim}{\xi}_T = \begin{bmatrix} 1.0 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} \qquad (3.1)$$

The "association structure" which is to be preserved in the sense of (2.3) corresponds to the log-linear functions

$$\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{K}'[\underset{\sim}{\log}_e(\underset{\sim}{\pi})] \qquad (3.2)$$

where

$$\underset{\sim}{K}' = \begin{bmatrix} 1&1&-1&-1&-1&-1&1&1 \\ 1&-1&1&-1&-1&1&-1&1 \\ 1&-1&-1&1&1&-1&-1&1 \\ 1&-1&-1&1&-1&1&1&-1 \end{bmatrix}, \qquad (3.3)$$

which pertain to the first and second order interactions among the three attributes.

By using IPF to adjust the observed frequencies in Table 1 to the "allocation structure" specified by (3.1), Ireland and Kullback obtain the "raked" predicted cell frequencies shown in Table 2. The corresponding predicted proportions $\hat{\underset{\sim}{\pi}}_T$ are also given there together with their respective standard errors based on (2.12). Thus, by comparing these results with their counterparts in Table 1, it can be noted that the predicted proportions $\hat{\underset{\sim}{\pi}}_T$ are very similar to the original observed proportions, but have substantially smaller estimated standard errors.

Finally, since the "allocation structure" (3.1) corresponds to an hypothesis rather than a priori known constraints, Ireland and Kullback indicate that its acceptability for these data is supported by a nonsignificant ($\alpha = .25$) Minimum Discrimination Information Chi-Square Statistic for goodness of fit $Q_{MDI}(\hat{\underset{\sim}{\pi}}_T|\underset{\sim}{p}) = 3.42$ with D.F. = 3.

ACKNOWLEDGMENTS

This research was supported in part by National Institutes of Health, Institute of General Medical Sciences through Grant GM-70004-05 and the GM-00047 and by the U.S. Bureau of the Census through Joint Statistical Agreement JSA 75-2. The authors would like to thank Stephen E. Fienberg and Frederick J. Scheurin for several comments which were helpful during the initial stages of this research. Additional insights were gained as a part of general discussions of statistical methodology during the Social Science Research Council 1975 Summer Workshop on Crime and Criminal Justice Statistics. Finally, we would like to thank Alcinda Lewis and Jane Beth Markley for computational assistance, and Joann Raccio, Karen McKee, and Rebecca Wesson for their cheeerful and conscientious typing of the manuscript.

REFERENCES

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis, MIT Press, Cambridge.

Causey, B.D. (1972). Sensitivity of raked contingency table totals to changes in problem conditions, Annals of Mathematical Statistics 43: 656-658.

Deming, W.E. (1943). Chapter VII: Adjusting sample frequencies to expected marginal totals, in Statistical Adjustment of Data, Dover Publications, Inc., New York: 96-127.

Freeman, D.H., Freeman, J.L., Brock, D.B., and Koch, G.G. (1977). Strategies in the multivariate analysis of data from complex surveys II: An application to the U.S. National Health Interview Survey, International Statistical Review, accepted for publication.

Gokhale, D.V. and Kullback, S. (1976). An information analysis of multinomial experiments and contingency tables: The general linear hypothesis, presentation to Washington Statistical Society.

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

Hauser, R.M., Koffel, J.N., Travis, H.P., and Dickinson (1975). Temporal change in occupational mobility: Evidence for men in the United States, American Sociological Review 40: 279-297.

Ireland, C.T. and Kullback, S. (1968). Minimum discrimination information estimation, Biometrics 24: 707-713.

Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975), Strategies in the multivariate analysis of data from complex surveys, International Statistical Review 43: 59-78.

Koch, G.G., Freeman, D.H., and Tolley, H.D. (1975). The asymptotic covariance structure of estimated parameters from contingency table log-linear models, Institute of Statistics Mimeo Series No. 1046, University of North Carolina.

Scheurin, F.L. (1973). Ransacking CPS tabulations: applications of the log-linear model to poverty statistics, Annals of Economic and Social Measurement 2: 159-182.

1. TABULATION OF MICE ACCORDING TO ATTRIBUTES A, B, AND D

	Response profile for attributes A vs B vs D							
Attribute A	Y	Y	Y	Y	N	N	N	N
Attribute B	Y	Y	N	N	Y	Y	N	N
Attribute D	Y	N	Y	N	Y	N	Y	N
Overall group observed cell frequency	475	460	462	509	467	440	494	427
Observed proportions	0.1272	0.1232	0.1237	0.1363	0.1251	0.1178	0.1323	0.1144
Estimated s.e.	0.0055	0.0054	0.0054	0.0056	0.0054	0.0053	0.0055	0.0052

Y denotes presence of the attribute; N denotes absence.

2. "RAKED" PREDICTED CONTINGENCY TABLE FOR ATTRIBUTES A, B, AND D

	Response profile for attributes A vs B vs D							
Attribute A	Y	Y	Y	Y	N	N	N	N
Attribute B	Y	Y	N	N	Y	Y	N	N
Attribute D	Y	N	Y	N	Y	N	Y	N
Overall group "raked" predicted cell frequency	463.3	464.5	438.7	500.5	475.4	463.8	489.6	438.2
Predicted Proportions	0.1241	0.1244	0.1175	0.1340	0.1273	0.1242	0.1311	0.1174
Estimated s.e.	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041

Y denotes presence of the attribute; N denotes absence.