

## **Analysis of Sample Survey Data Involving Categorical Response Variables: Methods and Software**

**J.N.K. RAO, S. KUMAR, and G. ROBERTS<sup>1</sup>**

### **ABSTRACT**

During the past 10 years or so, rapid progress has been made in the development of statistical methods of analysing survey data that take account of the complexity of survey design. This progress has been particularly evident in the analysis of cross-classified count data. Developments in this area have included weighted least squares estimation of generalized linear models and associated Wald tests of goodness of fit and subhypotheses, corrections to standard chi-squared or likelihood ratio tests under loglinear models or logistic regression models involving a binary response variable, and jackknifed chisquared tests. This paper illustrates the use of various extensions of these methods on data from complex surveys. The method of Scott, Rao and Thomas (1989) for weighted regression involving singular covariance matrices is applied to data from the Canada Health Survey (1978-79). Methods for logistic regression models are extended to Box-Cox models involving power transformations of cell odds ratios, and their use is illustrated on data from the Canadian Labour Force Survey. Methods for testing equality of parameters in two logistic regression models, corresponding to two time points, are applied to data from the Canadian Labour Force Survey. Finally, a general class of polytomous response models is studied, and corrected chi-squared tests are applied to data from the Canada Health Survey (1978-79). Software to implement these methods using the SAS facilities on a main frame computer is briefly described.

**KEY WORDS:** Corrections to chi-squared tests; Logistic regression; Power transformations; Wald tests; Weighted least squares.

### **1. INTRODUCTION**

Standard statistical methods, based on the assumption of independent identically distributed observations, are being used extensively by researchers in the social and health sciences, and in other subject matter areas. These methods have also been implemented in standard statistical packages, including SPSSX, BMDP, SAS and GLIM. In practice, however, much data are obtained from complex sample surveys involving clustering and stratification, so that the application of standard methods to these data without some adjustment for survey design can lead to erroneous inferences. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the complexity of the sample design is ignored in the analysis of data. Moreover, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, *e.g.*, residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods, and emphasized the need for new methods that take proper account of the complexity of survey design. During the past 10 years or so, rapid progress has been made in the development of such methods, particularly for analysing cross-classified count data. This paper will focus on the analysis of

---

<sup>1</sup> J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario; S. Kumar and G. Roberts, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario.

count data, but it should be noted that important results on other types of analyses have also been obtained: Regression analysis (Fuller 1975; Nathan and Holt 1980; Pfefferman and Nathan 1981; Scott and Holt 1982), principal component analysis (Skinner, Holmes and Smith 1986), factor analysis (Fuller 1986), logistic regression involving continuous covariates (Binder 1983).

Rao and Scott (1984) have made a systematic study of the impact of survey design on standard Pearson chi-squared or likelihood ratio tests for multiway tables of counts, under hierarchical log-linear models. They have also obtained simple first order corrections to standard tests which can be computed from published tables that include "design effects" for cell estimates and marginal totals, thus facilitating secondary analyses from published reports (see also Gross 1984; Bedrick 1983; Rao and Scott 1987). These first order corrections take account of the design in the sense that the actual type I error rates of tests based on the corrected statistics are closer to nominal levels, compared to the standard tests which could have greatly inflated type I error rates. More accurate second order corrections, based on the Satterthwaite approximation to a weighted sum of independent  $\chi^2$  variables, were also developed by Rao and Scott (1984), but these tests require the knowledge of a full estimated covariance matrix of cell estimates. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975), and the jackknifed chi-squared tests (Fay 1985), all requiring either the full estimated covariance matrix or access to cluster-level data. Fay (1985) and Thomas and Rao (1987) have shown that the Wald statistic, although asymptotically correct, can become highly unstable as the number of cells in the multiway table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level. On the other hand, Fay's jackknife tests and the Rao-Scott corrections have performed well under quite general conditions. In some cases, the instability in the Wald statistic may be remedied by collapsing the table according to eigenvectors associated with the nonnegligible eigenvalues of the estimated covariance matrix adjusted for singularities caused by linear constraints on the probabilities, as proposed by Singh (1985); see also Singh and Kumar (1986).

Roberts, Rao and Kumar (1987) assumed a logistic regression model for the cell (domain) proportions associated with a binary response variable, and obtained first order corrections to standard chi-squared and likelihood ratio tests of goodness-of-fit and nested hypotheses. Simple upper bounds to first order corrections, depending only on the design effects of cell response proportions, were also obtained to facilitate secondary analyses from published tables. Scott (1986) proposed an alternative method which uses standard tests on transformed data derived from the original data and the cell design effects. Roberts, Rao and Kumar (1987) also provided second order corrections to standard tests, but these require access to a full estimated covariance matrix of cell response proportions. Diagnostics for detecting outliers and influential points were developed as well, again taking the survey design into account.

The primary purpose of this paper is to present various extensions of the previous methods and illustrate their use on data from large-scale surveys, including the Canada Health Survey (1978-1979) and the Canadian Labour Force Survey. It is assumed, throughout the paper, that the user has access to a full estimated covariance matrix of cell estimates. In Section 2, weighted least squares (WLS) estimators of the parameters of generalized linear models having singular covariance matrices, caused by linear constraints on the probabilities (or proportions), are presented. Associated Wald tests of goodness-of-fit and of subhypotheses are also provided. A smoothed version of the WLS estimators, and associated Wald tests of subhypotheses are given as well. These methods should be used only when the number of cells in a table is small and/or the number of sample clusters in the survey design is relatively large.

The methods for logistic regression models are extended, in Section 3, to Box-Cox models involving power transformations of cell odds ratios. These models, which include the logistic regression model as a special case, could provide significantly better fits than the logistic regression models, as demonstrated by Guerrero and Johnson (1982) in the context of binomial proportions.

Methods for testing equality of parameters in two logit models, corresponding to two different time periods, are given in Section 4. If the hypothesis of equality is accepted, one could obtain "smoothed" estimates of cell proportions for the current period that are more efficient than the corresponding smoothed estimates based only on the current period data.

Section 5 gives an extension of the type of results obtained for logistic regression models to a general class of polytomous response models. The special case of McCullagh's (1980) ordered response model is studied in detail.

Finally, an account of the software for implementing the above methodology is given in Section 6.

## 2. WEIGHTED LEAST SQUARES ESTIMATORS AND WALD TESTS

The approach of Koch, Freeman and Freeman (1975) is designed to estimate the parameters of generalized linear models of the form  $g^*(p) = X^*\beta^*$ , using a sample estimate,  $\hat{p}$ , of the population cell probabilities denoted by a  $T$ -vector  $p$ , and a consistent estimate of  $\text{cov}(\hat{p}) = V_p$  (say). In this method, the asymptotic covariance matrix of the  $u$ -vector  $g^*(p)$  is assumed to be nonsingular ( $u < T$ ); however, many models, including the traditional loglinear model, are of the form  $g(p) = X\beta$ , where  $g(p)$  is a  $T$ -vector with a singular asymptotic covariance matrix, and  $X$  is a  $T \times r$  full rank matrix of known constants. It is possible to reduce the latter models to the nonsingular form  $g^*(p) = X^*\beta^*$ , as done by Grizzle and Williams (1972) for the loglinear model, but Scott, Rao and Thomas (1989) have developed the following unified approach for singular models, by appealing to the optimal theory for linear models having singular covariance matrices.

The cell probabilities  $p$  and  $\hat{p}$  are subject to linear constraints of the form  $K'p = \pi$  and  $K'\hat{p} = \pi$ , where  $K$  is a  $T \times L$  full rank matrix of known constants and  $\pi$  is an  $L$ -vector of known constants  $\pi_i$  ( $L < T$ ). As a result, the covariance matrix of  $\hat{p}$  will be singular. For example, in the case of stratified sampling with complex sample designs within strata, we can write  $K = I_L \otimes 1_m$ ,  $\pi_i = n_i/n$  ( $i = 1, \dots, L$ ) and  $p = (p_{11} \dots p_{1m}; \dots; p_{L1} \dots p_{Lm})'$  with  $p_{ij} = (n_i/n)\tilde{p}_{ij}$ , where  $\tilde{p}_{ij}$  is the  $j$ -th category probability within the  $i$ -th stratum ( $\sum_j \tilde{p}_{ij} = 1$ ;  $i = 1, \dots, L$ ;  $j = 1, \dots, m$ ),  $n_i$  is the sample size from the  $i$ -th stratum,  $\sum n_i = n$ ,  $1_m$  is a  $m$ -vector of 1's,  $I_L$  is the identity matrix of order  $L$  and  $\otimes$  denotes the Kronecker product.

Assume that  $X\beta$  can be written as  $X_0\beta_0 + X_1\beta_1$ , where  $X_0$  is a  $T \times L$  matrix such that  $K'H^{-1}X_0$  is nonsingular and where  $H = (\partial g/\partial p)'$  is the  $T \times T$  matrix of partial derivatives of  $g(p)$ . In particular,  $X_0$  can be taken as  $K$  if the constraint matrix  $K$  is included in  $X$ , as frequently assumed. Since restrictions on  $p$  imply constraints on the parameters  $\beta$ ,  $\beta_0$  can be determined exactly from the constraints, for a given  $\beta_1$ .

### Weighted least squares estimators

The model may be written as

$$\hat{g} = g(\hat{p}) = X\beta + \delta \quad (2.1)$$

where  $\delta$  is the error vector with  $P \lim \delta = 0$ , and  $\hat{g}$  has a singular asymptotic covariance matrix  $V_g = HV_pH'$  which is consistently estimated as  $\hat{V}_g = \hat{H}\hat{V}_p\hat{H}'$ , assuming that  $\hat{V}_p$  is a consistent estimator of  $V_p$ . Here  $\hat{H} = H(\hat{p})$ . Scott, Rao and Thomas (1989) derived an asymptotically best linear unbiased estimator (ABLUE) of  $\beta_1$  as

$$\hat{\beta}_1 = (\tilde{X}'_1\hat{M}\tilde{X}_1)^{-1}\tilde{X}'_1\hat{M}\hat{g}, \quad (2.2)$$

where

$$\hat{M} = (\hat{V}_g + X_0X'_0)^{-1} \quad (2.3)$$

is a nonsingular generalized inverse of  $\hat{V}_g$ , and

$$\tilde{X}_1 = [I - X_0X'_0\hat{M}]X_1. \quad (2.4)$$

A consistent estimator of the asymptotic covariance matrix of  $\hat{\beta}_1$  is given by

$$\text{est cov}(\hat{\beta}_1) = (\tilde{X}'_1\hat{M}\tilde{X}_1)^{-1}. \quad (2.5)$$

### Wald tests

Letting  $\hat{\beta} = (X'\hat{M}X)^{-1}X'\hat{M}\hat{g} = (\hat{\beta}_0', \hat{\beta}_1')'$ , a Wald test of goodness of fit of the model (2.1) is given by

$$W = (\hat{g} - X\hat{\beta})'\hat{M}(\hat{g} - X\hat{\beta}) \quad (2.6)$$

which is distributed asymptotically as a  $\chi^2$  variable with  $T - r$  degrees of freedom (d.f.). The model is considered tenable at the  $\alpha$ -level if  $W > \chi^2_{T-r}(\alpha)$ , the upper  $\alpha$ -point of  $\chi^2$  with  $T - r$  d.f..

Given the model (2.1), tests of linear hypotheses on the model parameters  $\beta_1$  can also be obtained. A Wald test of the linear hypothesis  $C_1\beta_1 = c_1$  is given by

$$W_1 = (C_1\hat{\beta}_1 - c_1)'[C_1 \text{est cov}(\hat{\beta}_1)C'_1]^{-1}(C_1\hat{\beta}_1 - c_1) \quad (2.7)$$

which is distributed asymptotically as a  $\chi^2$  variable with  $h$  d.f., where  $C_1$  is a  $h \times (r - L)$  full rank matrix of known constants ( $h < r - L$ ), and  $c_1$  is a  $h$ -vector of known constants. The hypothesis is rejected at the  $\alpha$ -level if  $W_1 > \chi^2_h(\alpha)$ , the upper  $\alpha$ -point of  $\chi^2$  with  $h$  d.f. Note that  $\beta_0$  should not be included in the linear hypothesis since it is fixed by the design constraints  $K'p = K'g^{-1}(X\beta) = \pi$ .

### Smoothed version of ABLUE and associated Wald tests

We can also obtain a smoothed version of ABLUE of  $\beta_1$ , say  $\beta_1^*$ , using iteration, as follows:

$$\check{\beta}_{t+1} = \check{\beta}_t + (X'M_tX)^{-1}X'M_tH_t(\hat{p} - p_t), \quad t = 0, 1, 2, \dots \quad (2.8)$$

with starting values  $M_0 = \hat{M}$ ,  $\check{\beta}_0 = (X'\hat{M}X)^{-1}X'\hat{M}\hat{g} = \hat{\beta}$ ,  $H_0 = H(\hat{\beta})$  and  $p_0 = p(\hat{\beta})$ . Further,  $M_t = (\hat{V}_{gt} + X_0X'_0)^{-1}$  with  $\hat{V}_{gt} = H_t\hat{V}_pH'_t$ ,  $H_t = H(\check{\beta}_t)$  and  $p_t = p(\check{\beta}_t)$ ,  $t \geq 1$ . At convergence, we get  $\beta^* = (\beta_0^*, \beta_1^*)'$  as the solution of the following equations:

$$X'M(\beta)H(\beta)(\hat{p} - p(\beta)) = 0. \quad (2.9)$$

Equations (2.9) reduce to quasilielihood equations (McCullagh 1983) when  $V_p$  is proportional to  $V(p)$ , a known function of  $p$ . Here, the dependence on  $\beta$  is made explicit by writing  $p = p(\beta)$ ,  $H = H(\beta)$  and  $M = V_g + X_0 X_0' = M(\beta)$ . The smoothed estimate  $\beta^*$  also satisfies the constraints  $K'p = K'g^{-1}(X\beta) = \pi$ , unlike  $\hat{\beta}$ . The asymptotic covariance matrices of  $\beta_1^*$  and  $\hat{\beta}_1$  are identical, but  $\beta_1^*$  might perform better in small samples.

Given the model (2.1), an alternate Wald test of the hypothesis  $C_1\beta_1 = c_1$  is given by

$$W_1^* = (C_1\beta_1^* - c_1)' [C_1 \text{ est cov}(\beta_1^*) C_1']^{-1} (C_1\beta_1^* - c_1) \quad (2.10)$$

which is distributed asymptotically as a  $\chi^2$  with  $h$  d.f., where

$$\text{est cov}(\beta_1^*) = (X_1^{*'} M^* X_1^*)^{-1}, \quad (2.11)$$

and  $X_1^* = [I - X_0 X_0' M^*] X_1$ ,  $M^* = (V_g^* + X_0 X_0')^{-1}$  with  $V_g^* = H^* \hat{V}_p H^{*'} and  $H^* = H(\beta^*)$ .$

### Example

The previous results were applied to a two-way table from the Canada Health Survey (1978-79). This survey was designed to provide reliable information on the health of Canadians. The information collected was made up of an interview component for the whole sample and a physical measures component for a subsample. A complex multistage design involving stratification and clustering was employed, and the estimates of cell totals or proportions were subjected to post-stratification on age-sex, to improve their efficiency. The reader is referred to Hidioglou and Rao (1987) for a description of the survey and the procedures used for estimating cell counts, proportions, and their estimated variances and covariances. For the physical measures component, a collapsed stratum technique for variance estimation was employed since a single primary sampling unit was selected in some of the strata.

Table 1 gives the estimated proportions,  $\hat{p}_{ij}$ , derived from the physical measures component in a cross-classification of fitness level (recommended = 1, minimal acceptable = 2, below acceptable or screened out = 3) and type of cigarette smoker (regular = 1, occasional = 2, never = 3). The estimated covariance matrix of the  $\hat{p}_{ij}$ ,  $\hat{V}_p$ , can be obtained from the authors.

Since both the variables in Table 1 are ordinal, we considered the following loglinear model with linear  $\times$  linear interaction:

$$\log p_{ij} = \bar{u} + u_{1(i)} + u_{2(j)} + \gamma(v_i - \bar{v})(w_j - \bar{w}), \quad i = 1, 2, 3 \quad j = 1, 2, 3 \quad (2.12)$$

**Table 1**  
Estimated Cell Proportions in a  $3 \times 3$  Table (Canada Level):  
Type of Cigarette Smoker  $\times$  Fitness Level (Sample Size  $n = 2505$ )  
Ages 15-64

Type of cigarette smokers	Fitness Level		
	1	2	3
1	0.22005	0.14951	0.16998
2	0.02301	0.00962	0.01146
3	0.20329	0.09933	0.11374

subject to side constraints  $\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0$ , where  $v_i$  and  $w_j$  are known scores with means  $\bar{v}$  and  $\bar{w}$  respectively. For simplicity, equidistant scores were taken:  $u_i = 1, 2, 3$ ;  $v_j = 1, 2, 3$ . The model (2.12) is of the form  $g(p) = X_0\beta_0 + X_1\beta_1$  with  $g_{ij}(p) = \log p_{ij}$ ,  $X_0 = K = 1_9$ , a  $9 \times 1$  vector of 1's,  $\beta_0 = \tilde{u}$ ,  $\beta_1 = (u_{1(1)}, u_{1(2)}, u_{2(1)}, u_{2(2)}, \gamma)'$ , and

$$X_1' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Noting that  $\hat{H} = \text{diag}(\hat{p}_{ij}^{-1}, i = 1, 2, 3; j = 1, 2, 3)$ , the Wald test of goodness-of-fit of the model (2.12) can be computed from (2.6), using the proportions  $\hat{p}_{ij}$  in Table 1 and the estimated covariance matrix,  $\hat{V}_p$ . We obtain

$$W = 3.59$$

which is not significant at the 5% level compared to  $\chi^2_{T-r}(0.05) = \chi^2_3(0.05) = 7.81$  (note that  $T = 9, r = 6$ ). The Wald statistic  $W$  is likely to be stable in this example since the number of cells  $T (= 9)$  is small relative to the number of sample clusters ( $= 50$ ).

We can also conduct a test of independence, *i.e.*  $\gamma = 0$ , given the model (2.12), using  $W_1$  given by (2.7) or  $W_1^*$ , based on the smoothed estimates  $\beta_1^*$ , given by (2.10). Noting that  $C_1 = (0, \dots, 0, 1)$ ,  $c_1 = 0$ , we obtain

$$W_1 = 8.23, \quad W_1^* = 8.75,$$

both larger than  $\chi^2_1(0.01) = 6.63$ , the upper 1% point of  $\chi^2$  with 1 d.f. The nested hypothesis of independence is therefore not tenable.

Accepting the model (2.12), we obtain the following values of weighted least squares estimates,  $\hat{\beta}_1$ , and smoothed estimates,  $\beta^*$ :

$$\hat{\beta}_1 = (0.912, -1.550, 0.339, -0.255, -0.086)'$$

$$\beta_0^* = -2.665, \quad \beta_1^* = (0.917, -1.568, 0.344, -0.262, 0.087)'.$$

The estimate  $\beta^*$  can also be used to produce smoothed estimates of the  $p_{ij}$ ,  $p_{ij}^* = p_{ij}(\beta^*)$ , which satisfy the constraint  $\sum \sum p_{ij}(\beta^*) = 1$ .

### 3. BOX-COX TRANSFORMATION MODELS

Logistic regression models are extensively used for the analysis of variation in the estimated proportions associated with a binary response variable. Suppose that the population of interest is partitioned into  $I$  cells according to the levels of one or more factors. Let  $P_i$  be the population response proportion in the  $i$ -th cell. Then a logistic regression model for the proportions  $P_i = F_i(\beta)$  is given by

$$\log\{F_i/(1 - F_i)\} = x_i'\beta, \quad i = 1, \dots, I, \quad (3.1)$$

where  $x_i = (x_{1i}, \dots, x_{si})'$  is an  $s$ -vector of known constants derived from the factor levels with  $x_{1i} = 1$ , and  $\beta$  is an  $s$ -vector of unknown parameters.

Guerrero and Johnson (1982) extended the applicability of logistic regression models by introducing an additional parameter,  $\lambda$ , through a Box-Cox power transformation of the odds ratios  $F_i/(1 - F_i)$ . Their model is given by

$$v_i(\lambda) = \{F_i/(1 - F_i)\}^{(\lambda)} = x_i' \beta, \quad i = 1, \dots, I, \quad (3.2)$$

where  $\beta$  and  $x_i$  are as in (3.1) and

$$\{F_i/(1 - F_i)\}^{(\lambda)} = \begin{cases} \log\{F_i/(1 - F_i)\} & \text{if } \lambda = 0 \\ \lambda^{-1}[\{F_i/(1 - F_i)\}^\lambda - 1] & \text{if } \lambda \neq 0. \end{cases}$$

The model (3.2) includes as a special case ( $\lambda = 0$ ) the logistic regression model (3.1). Guerrero and Johnson (1982) applied this model to data from the National Survey of Household Income and Expenditures in Mexico to explain the variation in female participation in the Mexican labour force. They found that a value of  $\lambda = -6.63$  provided a significantly better fit than the logit model ( $\lambda = 0$ ), the values of the standard chi-squared statistic being 4.8 (7 d.f.) and 12.8 (8 d.f.) respectively. However, they applied standard methods for binomial proportions, ignoring the survey design.

### Pseudo MLE

In this section, the methods of Roberts, Rao and Kumar (1987) for the logistic regression model are extended to the power transformation model (3.2). Due to difficulties in obtaining appropriate likelihood functions for general sample designs, we use "pseudo" maximum likelihood estimates,  $\hat{\beta}$  and  $\hat{\lambda}$ , obtained from the product binomial likelihood equations for  $\beta$  and  $\lambda$  by replacing the simple response proportion  $r_i/n_i$  with the corresponding survey estimate  $\hat{P}_i$  of  $P_i$ , and  $n_i/n$  with the corresponding survey estimate  $\hat{W}_i$  of the domain proportion  $W_i$ . Here  $r_i$  is the number of "successes" in a sample of size  $n_i$  from the  $i$ -th cell, and  $n = \sum n_i$ . See Guerrero and Johnson (1982), for the product binomial likelihood equations. The pseudo maximum likelihood estimates (m.l.e.),  $\hat{\theta}' = (\hat{\beta}', \hat{\lambda})$ , can be obtained iteratively by a quasi-Newton procedure, as in Guerrero and Johnson (1982). The fitted response proportions are given by  $\hat{F} = F_i(\hat{\theta})$ .

Let  $\hat{V}_p$  be the estimated covariance matrix of the survey estimates  $\hat{P} = (\hat{P}_1, \dots, \hat{P}_I)'$ , and let

$$B = D(\hat{F})^{-1} D(1 - \hat{F})^{-1} (\partial F / \partial \hat{\theta})'. \quad (3.3)$$

Here  $D(\hat{F}) = \text{diag}(\hat{F}_i, i = 1, \dots, I)$ ,  $D(1 - \hat{F}) = \text{diag}(1 - \hat{F}_i, i = 1, \dots, I)$  and  $(\partial F / \partial \hat{\theta})'$  is the  $I \times (s + 1)$  matrix of partial derivatives  $\partial F_i / \partial \beta_j$  and  $\partial F_i / \partial \lambda$  evaluated at  $\hat{\theta}$ :

$$\partial F_i / \partial \beta_j = x_{ji} F_i^2 (1/Q_i)^{1+1/\lambda}$$

$$\partial F_i / \partial \lambda = F_i^2 (Q_i \log Q_i - Q_i + 1) \lambda^{-2} (1/Q_i)^{1+1/\lambda}, \quad (3.4)$$

where  $Q_i = 1 + \lambda \sum_j x_{ji} \beta_j$ . The estimated asymptotic covariance matrix of  $\hat{\theta}$ , taking account of the survey design, is then given by (see Roberts 1985)

$$\text{est cov}(\hat{\theta}) = (B' \hat{\Delta} B)^{-1} (B' D(\hat{W}) \hat{V}_p D(\hat{W}) B) (B' \hat{\Delta} B)^{-1}, \quad (3.5)$$

where  $\hat{\Delta} = \text{diag}(\hat{W}_i \hat{F}_i (1 - \hat{F}_i); i = 1, \dots, I)$  and  $D(\hat{W}) = \text{diag}(\hat{W}_i, i = 1, \dots, I)$ .

It is also of interest to find the standard errors of the residuals  $\hat{R}_i = \hat{P}_i - \hat{F}_i$  since the standardized residuals  $\hat{R}_i / \text{s.e.}(\hat{R}_i)$  can be used to detect any outlying cell proportions. The estimated asymptotic covariance matrix of the vector of residuals  $\hat{R} = (\hat{R}_1, \dots, \hat{R}_I)'$  is given by

$$\text{est cov}(\hat{R}) = A \text{ est cov}(\hat{\theta}) A' = \hat{V}_R, \quad (3.6)$$

where

$$A = I - D(\hat{F}) D(1 - \hat{F}) B (B' \hat{\Delta} B)^{-1} B' D(\hat{W}).$$

The square root of the diagonal elements,  $\hat{V}_{ii,R}$ , of (3.6) provide the estimated standard errors of the  $\hat{R}_i, i = 1, \dots, I$ .

### Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of goodness-of-fit of the model (3.2) are given by

$$X^2 = n \sum_{i=1}^I (\hat{P}_i - \hat{F}_i)^2 \hat{W}_i / \{\hat{F}_i (1 - \hat{F}_i)\} \quad (3.7)$$

and

$$G^2 = 2n \sum_{i=1}^I \hat{W}_i [\hat{P}_i \log(\hat{P}_i / \hat{F}_i) + (1 - \hat{P}_i) \log\{(1 - \hat{P}_i) / (1 - \hat{F}_i)\}], \quad (3.8)$$

respectively, where the term in  $[\ ]$  of (3.8) equals  $-\log(1 - \hat{F}_i)$  at  $\hat{P}_i = 0$  and  $-\log \hat{F}_i$  at  $\hat{P}_i = 1$ .

Under product binomial sampling, it is well-known that both  $X^2$  and  $G^2$  are asymptotically identically distributed as a  $\chi^2$  variable with  $I - s - 1$  d.f., but for general sample designs this result is no longer valid. In fact,  $X^2$  (or  $G^2$ ) is asymptotically distributed as a weighted sum,  $\sum \delta_k W_k$ , of independent  $\chi^2$  variables,  $W_k$ , each with 1 d.f., where the weights  $\delta_k$  ( $k = 1, \dots, I - s - 1$ ) can be interpreted as “generalized design effects” (see Roberts 1985). Under product binomial sampling,  $\delta_k = 1$  for all  $k$ , and  $\sum \delta_k W_k$  reduces to  $\chi^2$  with  $I - s - 1$  d.f.

A first-order correction to  $X^2$  (or  $G^2$ ) is obtained by treating  $X_c^2 = X^2 / \hat{\delta}$  or  $G_c^2 = G^2 / \hat{\delta}$  as  $\chi^2$  with  $I - s - 1$  d.f., where

$$(I - s - 1) \hat{\delta} = \sum \delta_k = n \sum_{i=1}^I \hat{V}_{ii,R} \hat{W}_i / \{\hat{F}_i (1 - \hat{F}_i)\} \quad (3.9)$$

and  $\hat{V}_{ii,R}$  is the estimated variance of the  $i$ -th residual  $\hat{R}_i$ .

A more accurate, second order correction to  $X^2$  (or  $G^2$ ), based on the Satterthwaite approximation to  $\sum \delta_k W_k$ , is obtained by treating

$$X_S^2 = \frac{X_c^2}{1 + \hat{a}^2} \text{ or } G_S^2 = \frac{G_c^2}{1 + \hat{a}^2} \text{ as } \chi^2 \text{ with } (I - s - 1) / (1 + \hat{a}^2) \text{ d.f.} \quad (3.10)$$

Here  $\hat{a}^2 = \sum (\hat{\delta}_k - \hat{\delta})^2 / \{ (I - s - 1) \hat{\delta}^2 \}$  is the squared coefficient of variation of the  $\hat{\delta}_i$  which can be computed, without evaluating the individual weights  $\hat{\delta}_i$ , from (3.9) and from

$$\sum \hat{\delta}_k^2 = \sum_{i=1}^I \sum_{l=1}^I \hat{V}_{il,R}^2 (n\hat{W}_i) (n\hat{W}_l) / \{ \hat{f}_i \hat{f}_l (1 - \hat{f}_i) (1 - \hat{f}_l) \}, \quad (3.11)$$

where  $\hat{V}_{il,R}$  is the  $(i,l)$ -th element of  $\hat{V}_R$  given by (3.6).

Nested hypotheses, given the model (3.2), can also be tested by correcting the standard tests for nested hypotheses, but we omit this topic for simplicity (see Roberts 1985 and Kumar and Rao 1985 for details). It is simpler, however, to use Wald tests based on the estimates  $\hat{\beta}$  and the associated estimated asymptotic covariance matrix.

### Example

The previous method was applied to data from the monthly Canadian Labour Force Survey (October, 1980). The Labour Force Survey design employs multi-stage cluster sampling with two stages in the self-representing urban areas and three or four stages in the non-self-representing areas in each province. A detailed description of the sample design and associated estimation procedures for the Labour Force Survey is given in Statistics Canada (1977).

The sample from the Labour Force Survey, for the present example, consisted of males aged 15-64 who were in the labour force and not full-time students. Two factors, age and education, were chosen to explain the unemployment rates via a Box-Cox transformation model. Age-group levels were formed by dividing the interval  $[15, 64]$  into ten groups with the  $j$ -th age group being the interval  $[10 + 5j, 14 + 5j]$  for  $j = 1, \dots, 10$  and then using the mid-point of each interval,  $A_j = 12 + 5j$ , as the value of age for all persons in that age group. Similarly, the levels of education,  $E_k$ , were formed by assigning to each person a value based on the median years of school resulting in the following six levels: 7, 10, 12, 13, 14 and 16. The resultant age by education cross-classification provides a two-way table of  $I = 60$  survey estimates,  $\hat{P}_{jk}$ , of employment rates  $P_{jk}$ . The estimated covariance matrix  $\hat{V}_P$  was based on more than 450 sample clusters.

We considered the following transformation model for  $P_{jk} = F_{jk}(\theta)$  involving linear and quadratic age effects and linear education effect:

$$\begin{aligned} v_{jk}(\lambda) &= \{F_{jk} / (1 - F_{jk})\}^{(\lambda)} \\ &= \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \quad j = 1, \dots, 10, \quad k = 1, \dots, 6. \end{aligned} \quad (3.12)$$

Table 2 contains the pseudo m.l.e. of  $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \lambda)'$  and associated standard errors, and the test statistics  $X^2$ ,  $G^2$ ,  $X_S^2$  and  $G_S^2$  for testing the goodness-of-fit of the model (3.12). The corresponding values under the logistic regression model ( $\lambda = 0$ ) are also given for comparison.

It is clear from Table 2 that the value of  $X^2$  (or  $G^2$ ) is essentially equal to the corresponding value under the logistic regression model. Thus in the present example the transformation model provides no improvement in the fit over the logistic regression model. This is also clear from the value of  $\hat{\lambda}$  ( $= 0.016$ ) which is not significantly different from  $\lambda = 0$  when compared to its standard error ( $= 0.085$ ). The estimates of regression coefficients are essentially equal under the two models, but the standard errors of the  $\hat{\beta}_i$  under the Box-Cox model are much larger than the corresponding standard errors under the logistic regression model, due to the large standard error associated with  $\hat{\lambda}$  and the fact that the  $\hat{\beta}_i$  depend on  $\hat{\lambda}$ .

**Table 2**  
Pseudo MLE of the Parameters ( $\hat{\beta}', \lambda$ ), their Standard Errors and  
Test Statistics Under the Transformation Model and under  
the Corresponding Logistic Regression Model ( $\lambda = 0$ )

	Transformation Model		Logistic Regression Model	
	estimate	s.e.	estimate	s.e.
$\hat{\beta}_0$	-3.28	0.975	-3.10	0.247
$\hat{\beta}_1$	0.219	0.0468	0.211	0.013
$\hat{\beta}_2$	-0.00227	0.00049	-0.00218	0.00017
$\hat{\beta}_3$	0.1579	0.0385	0.1509	0.0115
$\hat{\lambda}$	0.016	0.085	—	—
Test Statistics				
	value	d.f.	value	d.f.
$X^2$	99.6	55	99.8	56
$G_2$	102.6	56	102.5	56
$X_S^2$	40.7	39.2	23.4	24.2
$G_S^2$	42.0	39.2	23.9	24.2
$X_S^2(0.05)$	54.6	55	47.7	56
$G_S^2(0.05)$	56.4	55	48.9	56

If the survey design is ignored and the value of  $X^2$  (or  $G^2$ ) is referred to  $\chi_{0.05}^2(55) = 73.3$ , the upper 5% point of  $\chi^2$  with  $I - s - 1 = 55$  d.f., we would reject the model (3.12). On the other hand, the value of  $X_S^2$  (or  $G_S^2$ ) when adjusted to refer to  $\chi_{0.05}^2(55)$ , denoted as  $X_S^2(0.05)$  (or  $G_S^2(0.05)$ ) in Table 2, is not significant at the 5% level, indicating that the model provides a good fit to the data,  $\hat{P}_{jk}$ .

Box and Cox (1982) and Hinkley and Runger (1984) argued that statistical inference about  $\beta$  should proceed with the scale determined by the estimate  $\hat{\lambda}$  regarded as fixed. Thus, the estimated covariance matrix of  $\hat{\beta}$  is determined from (3.5) by replacing  $\partial F / \partial \hat{\theta}$  by  $\partial F / \partial \hat{\beta}$  in the expression for  $B$  (equation (3.3)). For our example, this argument would suggest that we can take  $\hat{\lambda} = 0$  and use the estimates of  $\beta$  and associated standard errors (or estimated covariance matrix) under the logistic regression model, given in Table 2.

#### 4. TESTING EQUALITY OF LOGISTIC REGRESSION MODELS

Structural changes between two time periods may be detected through tests of equality of parameters in the corresponding models. Such tests for standard linear regression models have been developed extensively in the econometric literature (see *e.g.*, Amemiya 1985, Sec. 1.5.3).

In this section, corrected chi-squared and likelihood ratio tests of equality of parameters in two logistic regression models, corresponding to two specified time periods, are obtained. If the hypothesis of equality is tenable, then “smoothed” (*i.e.*, fitted) estimates of cell proportions for the current period can be obtained by combining the data for the two periods.

These estimates are more efficient than the corresponding smoothed estimate based only on the current period data. The methodology is applied to data from the October 1980 and October 1981 Canadian Labour Force Survey, to study year-to-year structural changes. Note that the data for October 1980 has already been used, in Section 3, to illustrate the fitting of Box-Cox power transformation models, and it was found that a logistic regression model involving linear and quadratic age effects and linear education effect provides a good fit to the data.

Let  $P_{ti}$  be the population response proportion in the  $i$ -th cell for the period  $t$  ( $= 1, 2$ ). Then a logistic regression model for the proportions  $P_{ti} = F_i(\beta_t) = F_{ti}$  is given by

$$\log\{F_{ti}/(1 - F_{ti})\} = x_i'\beta_t, \quad i = 1, \dots, I; t = 1, 2 \quad (4.1)$$

where  $x_i$  is an  $s$ -vector of known constants derived from the factor levels, as in (3.1), and  $\beta_t$  is an  $s$ -vector of unknown parameters for period  $t$ . We are interested in testing the composite hypothesis  $\beta_1 = \beta_2 (= \beta)$  to study structural changes between the two time periods. If the hypothesis is accepted, "smoothed" estimates of the proportions  $P_{2i}$  for the current period ( $t = 2$ ) can be obtained as  $F_i(\hat{\beta})$  where  $\hat{\beta}$  is the pseudo m.l.e. of the common parameter  $\beta$ .

### Pseudo MLE

Let  $\hat{P}_{1i}$  and  $\hat{P}_{2i}$  ( $i = 1, \dots, I$ ) be the survey estimates based on sample sizes  $n_1$  and  $n_2$  respectively. Extending the notation in Section 3, "pseudo" maximum likelihood estimates,  $\hat{\beta}_t$ , are obtained from the product binomial likelihood equations for  $\beta_t$  by replacing the simple response proportions  $r_{ti}/n_{ti}$  with the corresponding survey estimates  $\hat{P}_{ti}$  of  $P_{ti}$  and  $n_{ti}/n_t$  with the corresponding survey estimates  $\hat{W}_{ti}$  of the domain proportions  $W_{ti}$ , thus yielding

$$X'D(\hat{W}_t)\hat{F}_t = X'D(\hat{W}_t)\hat{P}_t, \quad t = 1, 2 \quad (4.2)$$

where  $\hat{F}_t = F(\hat{\beta}_t)$  is the vector of fitted response proportions for period  $t$ ,  $D(\hat{W}_t) = \text{diag}(\hat{W}_{ti}, i = 1, \dots, I)$ , and  $X' = (x_1, \dots, x_I)$ . The estimates  $\hat{\beta}_t$  are obtained iteratively by a quasi-Newton procedure.

Under the hypothesis  $\beta_1 = \beta_2 (= \beta)$ , the pseudo maximum likelihood estimates,  $\hat{\beta}$ , are obtained by iteration from the following pseudo likelihood equations:

$$X'D(\hat{W}_c)\hat{F} = (n_1/n)X'D(\hat{W}_1)\hat{P}_1 + (n_2/n)X'D(\hat{W}_2)\hat{P}_2, \quad (4.3)$$

where  $D(\hat{W}_c) = (n_1/n)D(\hat{W}_1) + (n_2/n)D(\hat{W}_2)$ ,  $\hat{F} = F(\hat{\beta})$  is the vector of fitted response proportions or smoothed estimates of cell proportions for the current period, and  $n_1 + n_2 = n$ .

Let  $\hat{V}_P$  be the estimated covariance matrix of  $(\hat{P}_1', \hat{P}_2')'$  partitioned as

$$\hat{V}_P = \begin{bmatrix} \hat{V}_{11P} & \hat{V}_{12P} \\ \hat{V}_{21P} & \hat{V}_{22P} \end{bmatrix}.$$

Then the estimated covariance matrix of smoothed estimates  $\hat{F}$  is given by

$$\text{est cov}(\hat{F}) = B\hat{V}_PB', \quad (4.4)$$

where

$$B = D(\hat{W}_c)^{-1} \hat{\Delta} X (X' \hat{\Delta} X)^{-1} X' [(n_1/n) D(\hat{W}_1), (n_2/n) D(\hat{W}_2)] \quad (4.5)$$

and

$$\hat{\Delta} = \text{diag}(\hat{W}_c \hat{F}_i (1 - \hat{F}_i)), i = 1, \dots, I.$$

If the residuals are defined as  $\hat{R}_t = \hat{F}_t - \hat{F}$ , then the estimated covariance matrix of  $(\hat{R}'_1, \hat{R}'_2)'$  is given by

$$\hat{V}_R = \begin{bmatrix} \hat{V}_{11R} & \hat{V}_{12R} \\ \hat{V}_{21R} & \hat{V}_{22R} \end{bmatrix} = A \hat{V}_P A'. \quad (4.6)$$

Here

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with

$$A_{t1} = D(\hat{W}_c)^{-1} \hat{\Delta} X \left[ (X' \hat{\Delta}_t X)^{-1} X' D(\hat{W}_t) - \frac{n_t}{n} (X' \hat{\Delta} X)^{-1} X' D(\hat{W}_t) \right],$$

and

$$A_{t2} = -D(\hat{W}_c)^{-1} \hat{\Delta} X (X' \hat{\Delta} X)^{-1} X' \left\{ D(\hat{W}) - \frac{n_t}{n} D(\hat{W}_t) \right\}, t = 1, 2,$$

where

$$\hat{\Delta}_t = \text{diag}(\hat{W}_{ti} \hat{F}_i (1 - \hat{F}_i)), i = 1, \dots, I.$$

### Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of the nested hypothesis  $\beta_1 = \beta_2$ , given the model (4.1), are given by

$$X^2 = X_1^2 + X_2^2 \quad (4.8)$$

and

$$G^2 = G_1^2 + G_2^2, \quad (4.9)$$

where

$$X_t^2 = n_t \sum_{i=1}^I (\hat{F}_{ti} - \hat{F}_i)^2 \hat{W}_{ti} / \{\hat{F}_i (1 - \hat{F}_i)\}, t = 1, 2 \quad (4.10)$$

and

$$G_t^2 = 2n_t \sum_{i=1}^I \hat{W}_{ti} \left[ \hat{F}_{ti} \log(\hat{F}_{ti} / \hat{F}_i) + (1 - \hat{F}_{ti}) \log\{(1 - \hat{F}_{ti}) / (1 - \hat{F}_i)\} \right], t = 1, 2. \quad (4.11)$$

A first order correction to  $X^2$  (or  $G^2$ ) is obtained by treating  $X_c^2 = X^2/\hat{\delta}$ , or  $G_c^2 = G^2/\hat{\delta}$ , as  $\chi^2$  with  $s$  d.f., where

$$s\hat{\delta} = n_1 \sum_{i=1}^I \hat{V}_{11R}(ii) \hat{W}_{1i} / \{\hat{F}_i(1 - \hat{F}_i)\} + n_2 \sum_{i=1}^I \hat{V}_{22R}(ii) \hat{W}_{2i} / \{\hat{F}_i(1 - \hat{F}_i)\} \quad (4.12)$$

and  $\hat{V}_{\mu R}(ij)$  is the  $(i,j)$ th element of  $\hat{V}_{\mu R}$ . A more accurate, second order correction to  $X^2$  (or  $G^2$ ), based on the Satterthwaite approximation, is obtained by treating

$$X_S^2 = \frac{X_c^2}{1 + \hat{a}^2} \quad \text{or} \quad G_S^2 = \frac{G_c^2}{1 + \hat{a}^2} \quad \text{as } \chi^2 \quad \text{with } s/(1 + \hat{a}^2) \text{ d.f.} \quad (4.13)$$

Here  $\hat{a}^2 = (\sum_{k=1}^s \hat{\delta}_k^2 - s\hat{\delta}^2)/s\hat{\delta}^2$  which can be computed from (4.12) and the following formula for  $\sum \hat{\delta}_k^2$ :

$$\begin{aligned} \sum_{k=1}^s \hat{\delta}_k^2 &= n_1^2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{11R}(ij) \hat{W}_{1i} \hat{W}_{1j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i)(1 - \hat{F}_j)} \\ &\quad + n_2^2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{22R}(ij) \hat{W}_{2i} \hat{W}_{2j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i)(1 - \hat{F}_j)} \\ &\quad + 2n_1 n_2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{12R}(ij) \hat{W}_{1i} \hat{W}_{2j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i)(1 - \hat{F}_j)}, \end{aligned} \quad (4.14)$$

where  $\hat{V}_{12R}(ij)$  is the  $(i,j)$ -th element of  $\hat{V}_{12R}$ .

### Example

The previous method was applied to data from the October 1980 and October 1981 Canadian Labour Survey, to study year-to-year structural changes.

The logistic regression model involving linear and quadratic age effects and linear education effect provided a good fit to data from both periods with the following estimates of  $\beta_t$ :

$$\hat{\beta}_1: \{-3.08, 0.211, -0.00218, 0.1505\}$$

$$\hat{\beta}_2: \{-3.05, 0.179, -0.00169, 0.1707\},$$

where  $\log\{\hat{F}_{ijk}/(1 - \hat{F}_{ijk})\} = \hat{\beta}_{t0} + \hat{\beta}_{t1}A_j + \hat{\beta}_{t2}A_j^2 + \hat{\beta}_{t3}E_k, j = 1, \dots, 10; k = 1, \dots, 6$  and  $\hat{F}_{ijk}$  is the fitted employment rate in the  $(j,k)$ -th cell for period  $t$ . One cell was omitted in the fitting since the domain sample size  $n_{2i}$  is zero for the current period.

Turning to the test of the hypothesis  $\beta_1 = \beta_2$ , given the logistic regression models, we obtained the following values of  $X^2$ ,  $G^2$ ,  $X_c^2$ ,  $G_c^2$  and  $X_S^2$ ,  $G_S^2$ :

$$X^2 = 42.1 \quad X_c^2 = 24.6 \quad X_S^2 = 24.4$$

$$G^2 = 42.2 \quad G_c^2 = 24.6 \quad G_S^2 = 24.4.$$

Also  $s/(1 + \hat{a}^2) = 4/(1.0089) = 3.965 \doteq 4$ . By referring  $X_S^2$  or  $G_S^2$  to  $\chi_{0.05}^2(4) = 9.49$ , the upper 5% point of  $\chi^2$  with 4 d.f., we reject the hypothesis  $\beta_1 = \beta_2$  at the 5% level, indicating significant year-to-year structural changes for the month of October. The data for the two time periods, therefore, should not be pooled to get smoothed estimates of unemployment rates,  $1 - \hat{F}_{jk}$ , for the current period.

## 5. POLYTOMOUS RESPONSE MODELS

A variety of models has been suggested in the literature when the response variable is polytomous. The variety of models reflects, in part, the different scales of measurement possible for polytomous response variables, unlike binary response variables. In the main, there are nominal responses where any permutation of the response categories is equally valid, and ordinal responses where there is a natural ordering of the response categories.

Suppose that the population of interest is partitioned into  $I$  cells (or domains) according to the levels of one or more factors. Let  $P_{j(i)}$  be the population proportion in the  $i^{\text{th}}$  cell having the  $j^{\text{th}}$  response ( $j = 1, \dots, J + 1$ ) so that  $\sum_{j=1}^{J+1} P_j(i) = 1$  ( $i = 1, \dots, I$ ). Then a general polytomous response model for the proportions  $P_j(i)$  is given by

$$P_j(i) = F_{ij}(\theta), \quad i = 1, \dots, I; \quad j = 1, \dots, J, \quad (5.1)$$

where  $\theta$  is an  $r$ -vector of unknown parameters ( $r \leq IJ$ ) and  $F_{ij}(\theta)$  is a function of known form. In the nominal case, Haberman (1982) and others proposed the following model: the “multinomial logits”  $\log P_j(i) - \sum_{j'=1}^{J+1} \log P_{j'}(i) (J + 1)^{-1}$  are assumed to be unknown linear functions of  $x_i$ , the  $s$ -vector of known constants derived from the factor levels, i.e.,

$$F_{ij}(\theta) = \exp(x_i' \beta_j) / \sum_{k=1}^{J+1} \exp(x_i' \beta_k), \quad i = 1, \dots, I; \quad j = 1, \dots, J + 1 \quad (5.2)$$

with  $\sum \beta_k = 0$ . Because of the latter constraint on the  $\beta_k$ , (5.2) may be expressed as

$$F_{ij}(\theta) = \exp(x_i' \beta_j) / \left[ \sum_{k=1}^J \exp(x_i' \beta_k) + \prod_{k=1}^J \exp(-x_i' \beta_k) \right],$$

$$i = 1, \dots, I; \quad j = 1, \dots, J. \quad (5.3)$$

Note that (5.3) reduces to the usual logistic regression model in the special case of binary response.

In the ordinal case, a simple model which also has the feature of being invariant under the grouping of response categories is given by (McCullagh 1980)

$$\log\{C_{j(i)}/(1 - C_{j(i)})\} = \nu_j - x_i' \beta, \quad j = 1, \dots, J; \quad i = 1, \dots, I \quad (5.4)$$

where  $C_{j(i)} = \sum_{k=1}^j P_{k(i)}$  denotes the  $j^{\text{th}}$  cumulative probability in the  $i^{\text{th}}$  domain, and  $\theta' = (\nu_1, \dots, \nu_J, \beta')$ . To express (5.4) in the form (5.1), we note that  $P_i = L^{-1}C_i$ , where  $P_i = (P_{1(i)}, \dots, P_{J(i)})'$ ,  $C_i = (C_{1(i)}, \dots, C_{J(i)})'$  and  $L^{-1}$  is a  $J \times J$  nonsingular matrix with 1 in the diagonal,  $-1$  in the  $(i + 1, i)^{\text{th}}$  position ( $i < J$ ) and 0 elsewhere.

### Pseudo MLE

As before, we use pseudo m.l.e.,  $\hat{\theta}$  obtained from the product multinomial likelihood equations for  $\theta$  by replacing the simple response proportions  $n_{ij}/n_i$  with the corresponding survey estimates  $\hat{P}_{j(i)}$ , and  $n_i/n$  with the corresponding survey estimate  $\hat{W}_i$  of the domain proportion  $W_i$ . Here  $n_{ij}$  is the number of units with the  $j^{\text{th}}$  response in a sample of size  $n_i$  from the  $i^{\text{th}}$  domain and  $n = \sum n_i$ . The fitted response proportions are then given by  $\hat{F} = F(\hat{\theta}) = (\hat{F}_1', \dots, \hat{F}_I')'$ , where  $\hat{F}_i = (\hat{F}_{i1}, \dots, \hat{F}_{iJ})'$  and  $\hat{F}_{ij} = F_{ij}(\hat{\theta})$ .

Let  $\hat{V}_P$  be the estimated covariance matrix of the survey estimates  $\hat{P} = (\hat{P}_{1(1)}, \dots, \hat{P}_{J(1)}, \dots, \hat{P}_{1(I)}, \dots, \hat{P}_{J(I)})'$ , and  $\hat{M} = (\partial F / \partial \hat{\theta})'$ , the  $IJ \times r$  matrix of partial derivatives  $\partial F_{ij} / \partial \theta_k$  calculated at  $\hat{\theta}$ . Also, let  $\hat{Q}_i = \text{diag}(\hat{F}_i) - \hat{F}_i \hat{F}_i'$  and  $\hat{Q} = \text{diag}(\hat{Q}_i, i = 1, \dots, I)$ . The expressions for the partial derivatives  $\partial F_{ij} / \partial \theta_k$  for the models (5.3) and (5.4) are given in Roberts (1985). The estimated asymptotic covariance matrix of  $\hat{\theta}$ , taking account of the survey design, is then given by (see Roberts 1985).

$$\text{est cov}(\hat{\theta}) = (\hat{M}' \hat{\nabla} \hat{M})^{-1} (\hat{M}' \hat{\nabla} \hat{V}_P \hat{\nabla}' \hat{M}) (\hat{M}' \hat{\nabla} \hat{M})^{-1}, \quad (5.5)$$

where  $\hat{\nabla} = (D(\hat{W}) \otimes I) \hat{Q}^{-1}$  and  $D(\hat{W}) = \text{diag}(\hat{W}_i, i = 1, \dots, I)$ . In the special case of product multinomial sampling,  $\hat{V}_P = \hat{\nabla}^{-1}/n$  and (5.5) reduces to  $(\hat{M}' \hat{\nabla} \hat{M})^{-1}/n$ .

The vector of residuals,  $\hat{R} = \hat{P} - \hat{F}$ , is also of interest, since it may be useful in detecting model deviations. The estimated asymptotic covariance matrix of  $\hat{R}$  is given by

$$\text{est cov}(\hat{R}) = \hat{G} \hat{V}_P \hat{G}' \quad (5.6)$$

where  $\hat{G} = I - \hat{M}(\hat{M}' \hat{\nabla} \hat{M})^{-1} \hat{M}' \hat{\nabla}$ .

### Corrections to standard tests

For simplicity, we consider only the Pearson chi-squared test of goodness-of-fit of the model (5.1). It is given by

$$X^2 = n \sum_{i=1}^I \hat{W}_i \sum_{j=1}^{J+1} (\hat{P}_{j(i)} - \hat{F}_{ij})^2 / \hat{F}_{ij}. \quad (5.7)$$

Under independent multinomial sampling in each of the domains, it is well-known that  $X^2$  is asymptotically distributed as a  $\chi^2$  variable with  $IJ - r$  d.f.

To test the nested hypothesis  $\theta_2 = 0$ , given the model (5.1), let  $\hat{\theta}_1$  be the pseudo m.l.e. of  $\theta_1$  and  $\hat{F}$  be the corresponding vector of fitted response proportions, where  $\theta' = (\theta_1', \theta_2')$ ,  $\theta_1$  is  $q \times 1$  and  $\theta_2$  is  $u \times 1$  ( $q + u = r$ ). The Pearson chi-squared test of the nested hypothesis is then given by

$$X^2(2|1) = n \sum_{i=1}^I \hat{W}_i \sum_{j=1}^{J+1} (\hat{F}_{ij} - \hat{\hat{F}}_{ij})^2 / \hat{\hat{F}}_{ij} \quad (5.8)$$

which is asymptotically distributed as  $\chi^2$  with  $u$  d.f. under independent multinomial sampling in each of the domains. However, for a general sample design,  $X^2$  and  $X^2(2|1)$  are both asymptotically distributed as weighted sums of independent  $\chi^2$  variables, each with 1 d.f., where the weights can be interpreted as "generalized design effects" of particular linear transformations of  $\hat{P}$  (Roberts 1985).

A first-order correction to  $X^2(2|1)$  is obtained by treating

$$X_c^2(2|1) = X^2(2|1)/\hat{\delta} \cdot (2|1) \text{ as } \chi^2 \text{ with } u \text{ d.f.}, \quad (5.9)$$

where  $\hat{\delta} \cdot (2|1)$  is obtained by replacing  $\theta'$  by  $(\hat{\theta}'_1, 0')$  and  $V_P$  by  $\hat{V}_P$  in the following definition for  $\delta \cdot (2|1)$ :

$$u\delta \cdot (2|1) = \sum_{i=1}^u \delta_i(2|1) = \text{tr } D(2|1). \quad (5.10)$$

Here,  $\text{tr}$  denotes the trace operator and  $D(2|1)$  is a generalized design effects matrix given by

$$D(2|1) = (H'_2 \nabla H_2)^{-1} (H'_2 \nabla V_P \nabla' H_2), \quad (5.11)$$

where  $V_P$  is the covariance matrix of  $\hat{P}$ ,  $\nabla = (D(W) \otimes I)Q^{-1}$ ,  $Q$  is the block diagonal matrix with  $Q_i = \text{diag}(F_i) - F_i F'_i$ ,  $i = 1, \dots, I$ ,  $F_i = F_i(\theta)$ , and  $H_2 = [I - M_1 (M'_1 \nabla M_1)^{-1} M'_1 \nabla] M_2$ , where  $M_1 = (\partial F / \partial \theta_1)'$  and  $M_2 = (\partial F / \partial \theta_2)'$ .

A more accurate, second order correction to  $X^2(2|1)$ , based on the Satterthwaite approximation, is obtained by treating

$$X_S^2(2|1) = X_c^2(2|1)/[1 + \hat{a}(2|1)^2] \text{ as } \chi^2 \text{ with } u/[1 + \hat{a}(2|1)^2] \text{ d.f.} \quad (5.12)$$

Here  $\hat{a}(2|1)^2$  is obtained by replacing  $\theta$  by  $(\hat{\theta}'_1, 0')$  in the following definition of  $a(2|1)^2$ :

$$a(2|1)^2 = \left\{ \sum_{i=1}^u \delta_i(2|1)^2 - u\delta \cdot (2|1)^2 \right\} / u\delta \cdot (2|1)^2, \quad (5.13)$$

where

$$\sum_{i=1}^u \delta_i(2|1)^2 = \text{tr } D(2|1)^2. \quad (5.14)$$

The corrections to goodness-of-fit test  $X^2$  are obtained as special cases of (5.9) and (5.12) by treating the model as nested within a saturated model (*i.e.*, a model where the unknown parameter  $\theta$  is of length  $IJ$ ).

### Example

The previous methods were applied to data from the Canada Health Survey (1978-79). A brief description of the survey is provided in Section 2.

The data set examined consisted of the estimated counts of females aged 20-64 cross-classified by frequency of breast self-examination (with the 3 categories: monthly, quarterly, less often or never), education (with the 3 categories: secondary or less, some post-secondary, post-secondary) and age (with the 3 categories: 20-24, 25-44, 45-64).

The frequency of breast self-examination was considered to be the response variable, while education and age were taken as explanatory variables, so that the number of responses,  $J + 1$ , equalled 3 and the number of domains,  $I$ , was 9. Both response and explanatory variables are ordered.

**Table 3**  
Survey Estimates of Cumulated Probabilities

	Age	Education	$C_{1(ik)}$	$C_{2(ik)}$
$i = 1, k = 1$	20-24	$\leq$ Secondary	.25	.49
$k = 2$		$<$ Post-Secondary	.25	.41
$k = 3$		$\geq$ Post-Secondary	.23	.47
$i = 2, k = 1$	25-44	$\leq$ Secondary	.25	.50
$k = 2$		$<$ Post-Secondary	.27	.44
$k = 3$		$\geq$ Post-Secondary	.26	.44
$i = 3, k = 1$	45-64	$\leq$ Secondary	.28	.51
$k = 2$		$<$ Post-Secondary	.24	.62
$k = 3$		$\geq$ Post-Secondary	.29	.56

**Table 4**  
Statistics for Testing Goodness of Fit and Nested Hypotheses

	Goodness of Fit (Age & Education)	Nested Hypothesis (Age only)
$X^2$	37.7	7.1
$X_c^2$	21.6	3.8
$X_S^2$	18.5*	3.7*
$\hat{\delta}$	1.75	1.9
$\hat{a}^2$	0.83	0.1

\* The Satterthwaite statistic has been adjusted to refer to the same  $\chi^2$  value as  $X_c^2$ .

The following model for the cumulated probabilities of the type described in equation (5.4), was considered:

$$\log\{C_j(ik)/(1 - C_j(ik))\} = \nu_j + \beta a_i + e_k \quad (j = 1, 2; i = 1, 2, 3; k = 1, 2, 3) \quad (5.15)$$

where  $C_j(ik)$  is the  $j^{th}$  cumulated probability for the  $i^{th}$  age group and  $k^{th}$  education group. As well,  $a_i = A_i - \bar{A}$ , where  $A_i$  is the midpoint of the  $i^{th}$  age interval, and  $e_k$  is the effect of the  $k^{th}$  education group ( $\sum e_k = 0$ ), ignoring the order of the education categories. Table 3 contains the survey estimates of the cumulated proportions. Table 4 contains the test statistics  $X^2$ ,  $X_c^2$  and  $X_S^2$  for testing the goodness of fit of (5.15) and also for testing the nested hypothesis of no education effect,  $e_k = 0$  for  $k = 1, 2$ .

First, considering the goodness of fit of (5.15), if the survey design is ignored and the value of  $X^2$  is referred to  $\chi_{0.05}^2(13) = 22.4$ , the upper 5% point of  $\chi^2$  with  $IJ - 5 = 13$  d.f., we would reject the model. On the other hand, the value of  $X_c^2$  or the value of  $X_S^2$  when adjusted to refer to  $\chi_{0.05}^2(13)$ , is not significant at the 5% level, indicating that the model provides a good fit to the data.

For testing of the nested hypothesis, the value of  $X_c^2$ , or the value of  $X_S^2$  when adjusted to refer to  $\chi_{0.05}^2(2) = 5.99$  is not significant at the 5% level, indicating that the nested hypothesis of no education effect is tenable.

## 6. SOFTWARE

Implementation of the methodology of the previous sections requires two stages of computation — calculation of a vector of proportions, along with its estimated covariance matrix, and then calculation of model estimates, test statistics and their adjustments.

Surveys like the Canada Health Survey and the Labour Force Survey, from which examples have been presented, have complex designs and large data bases. Because of these two factors, calculation of covariance matrices was done on a mainframe computer. Custom SAS and Fortran programs were used for this purpose.

Computations required for the fitting and testing of goodness-of-fit models and sub-hypotheses were done either on the mainframe computer using SAS (and the MATRIX procedure in particular), or on a microcomputer using the GAUSS programming package.

These programs are available to other analysts at Statistics Canada.

## REFERENCES

- AMEMIYA, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.
- BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, 279-292.
- BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- BOX, G.E.P., and COX, D.R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209-210.
- FAY, R.E. (1985). A jack-knifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.
- FULLER, W.A. (1986). Estimators of the factor model for survey data. In *Advances in the Statistical Sciences*, Vol. I (Eds. MacNeill, I.B. and Umphrey, G.J.). Dordrecht, Holland: Reidel Publishing Co., 265-284.
- GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series B*, 46, 270-272.
- GUERRERO, V.M., and JOHNSON, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, 69, 309-314.
- HABERMAN, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.
- HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys, Parts I and II. *Journal of Official Statistics*, 3, 117-132 and 133-140.
- HINKLEY, D.V., and RUNGER, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302-309.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KOCH, G.G., FREEMAN, D.H. Jr., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KUMAR, S., and RAO, J.N.K. (1985). Fitting Box-Cox transformation models to labour force survey data. Unpublished Report, Social Surveys Methods Division, Statistics Canada, Ottawa.

- McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109-142.
- NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the American Statistical Association*, 76, 681-689.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- ROBERTS, G. (1985). *Contributions to Chi-Squared Tests with Survey Data*. Unpublished Ph.D. Thesis, Carleton University, Department of Mathematics and Statistics, Ottawa.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SCOTT, A.J. (1986). Logistic regression analysis with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 25-30.
- SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- SCOTT, A.J., RAO, J.N.K., and THOMAS, D.R. (1989). Weighted least squares and quasi maximum likelihood estimation for categorical data under generalized linear models. *Linear Algebra and its Applications*, second special issue on Linear Algebra and Statistics, in press.
- SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Statistics Canada Working Paper No. SSMD 86-002.
- SINGH, A.C., and KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-257.
- SKINNER, C.J., HOLMES, D.J., and SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.
- STATISTICS CANADA (1977). *Methodology of the Canadian Labour Force Survey, 1976*. Catalogue 71-526 occasional. Ottawa: Statistics Canada.
- THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

## COMMENT

**ROBERT E. FAY<sup>1</sup>**

The authors have made an excellent contribution to the literature on the analysis of data from complex samples. By examining in turn four different models for categorical data: i) a log-linear model for a cross-classification; ii) a modification of the approach of Box and Cox to the transformation of binary data; iii) a problem of inference about parameters of a logistic regression model; and iv) a polytomous response model, the authors present solutions to important individual problems and illustrate the ways in which these flexible approaches to inference can be extended to other models for categorical data from complex samples. The applications are connected by an underlying theory, much of it previously appearing in Rao and Scott (1984), but this paper usefully presents in greater detail the implications of the general theory for specific models.

An omission from the paper is understandable but worth noting: for each model illustrated in the paper, replication provides an alternative strategy that, at times, may also be more convenient. In particular, the replication theory is complete for each of the applications, i), ii), and iv), to cross-classified data. In each case, tests of overall fit and comparisons of nested models can be assessed with the jackknifed chi-square test (Fay 1985) and standard errors for the parameters obtained through replication.

Replication also can provide standard errors and covariances for parameters of logistic regression models, as in iii), enabling in some cases a Wald-type test for equality of sets of regression parameters. It also appears likely that the jackknifing approach extends to the likelihood-ratio chi-square test in such situations involving continuous variables, although a firm proof of this conjecture is clearly required before application can be recommended. My point in calling attention to replication as a competing strategy for the problems presented in the paper is not to imply that it represents a methodologically superior approach to the methods of Rao and Scott (1984); instead, the availability of this methodology provides an additional choice to solve these and similar problems of inference. For example, the focus on replication for the estimation of variances from the current demographic surveys at the U.S. Census Bureau provides the potential to carry out analyses such as those presented in the paper.

I also want to point out that the methods presented and the analogues from replication theory have a potential importance beyond the realm of design-based inference from complex sample surveys, which is the focus of the paper. One of these involves the use of multiple imputation or related approaches intended to represent the uncertainty due to missing data. The implied interpretation of variance within the domain of design-based inference can be extended to include uncertainty from missing data without requiring changes to the methodology presented in the paper. The general methodology may also be applicable to some problems of inference from complex designed experiments, in which the design poses problems of clustering or stratification similar to complex sample surveys.

Of the four models discussed, however, I suggest that the Box and Cox transformation not be applied without consideration of alternative strategies, such as transformation of the x-variables instead. My own inclination would be to favor an analysis on a logistic scale, with possibly transformed predictors, unless the adaptation of the Box and Cox transformation obtains some distinct advantage, such as offering an additive model on the transformed scale in an instance where the logistic model does not provide as successful a fit without interaction terms.

I am delighted to have the opportunity to commend the authors on a useful and instructive paper.

---

<sup>1</sup> Robert E. Fay, U.S. Bureau of the Census, Washington, D.C. 20233.

## COMMENT

C.J. SKINNER<sup>1</sup>

This paper provides an excellent discussion of a variety of applications of weighted least squares (WLS) and pseudo maximum likelihood (PML) procedures to categorical data. Its clear presentation and use of real survey examples will, I hope, help to encourage survey analysts to take account of complex designs in their analyses. As the authors indicate, analytical statistical procedures which take account of complex designs have been developed extensively in recent years (see *e.g.* Skinner, Holt and Smith 1989) and are even beginning to be referred to in standard computer software (*e.g.* SAS 1985, pp 61-67).

Commenting first on some specific aspects of the paper, I found Section 5 on polytomous variables to be especially valuable, given the wide occurrence of such data in surveys. A property of ordinal variables is that they may often be expected to possess monotonic relationships and so, for example, lack of monotonicity between the fitted values of  $C_{1(ik)}$  (or  $C_{2(ik)}$ ) and the education variable  $k$  in Table 3 makes the result of the corrected tests, that there is no evidence of an education effect, more plausible than the result of the uncorrected test.

The discussion of testing equality of two logistic regression models in Section 4 also seemed to me to be practically useful, although it would still seem to be possible theoretically to formulate this test as one of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987).

Section 3 provides a useful illustration of how PML may be applied to general parametric models for categorical data. It is, however, gratifying that the more complex transformation model provides no significant improvement in fit over the logistic regression model, since the interpretation of the parameters of the transformation model is more difficult. For example, for the logistic model the coefficient for education may be interpreted as implying that the odds of being employed are increased by 16% for each additional year of education for males of a given age ( $\exp(.1509) = 1.16$ ), whereas this interpretation is not generally available for the transformation model when  $\lambda \neq 0$ .

On a more general note I would be interested in the authors' views on the relative merits of WLS and PML. In the paper, these methods are presented quite separately, although both procedures would seem to be potentially applicable to a very wide class of models for categorical data under complex designs. Indeed both procedures are also applicable to models with continuous variables (Skinner, Holt and Smith 1989, Chapter 3); WLS requires just a statistic consistent for a known function of the parameters together with a consistent estimate of the covariance matrix of the statistic (Fuller 1984, Corollary 2), whereas PML is applicable very widely as described in Binder (1983). As a basis for discussion I list below a number of criteria on which WLS and PML might be compared; M1-M3 are relevant even under multinomial sampling, C1-C3 are specific to complex designs.

- M1 **Flexibility** WLS may be more adaptable than PML for complex problems *e.g.* involving structural zeros.
- M2 **Computation** WLS computation tends to have a more standard form.
- M3 **Small cell counts** WLS is more sensitive to small counts, especially zeros.
- C1 **Adaptability of multinomial methods to complex designs** WLS seems more easily adaptable.

<sup>1</sup> C.J. Skinner, University of Southampton, United Kingdom.

- C2 **Efficiency** Under multinominal sampling WLS is usually asymptotically equivalent to PML (which is then just standard ML). It might be conjectured that WLS will always be at least as efficient as PML under complex designs, although this presupposes a 1-1 correspondence between WLS and PML estimation problems. If WLS is more efficient, is the gain usually negligible (*cf.* Scott and Holt 1982)? Are there general results here?
- C3 **Degrees of freedom** WLS estimators and associated Wald tests may be unstable if the degrees of freedom used to estimate  $V_p$  are low.

#### ADDITIONAL REFERENCES

- FULLER, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* 10, 97-118.
- SAS Institute Inc. (1985). *SAS/IML User's Guide, Version 5 Edition*. Cary NC: SAS Institute Inc.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F., Eds. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

## COMMENT

E.A. MOLINA<sup>1</sup>

I would like to congratulate the authors on bringing together some recent methods developed for analyzing categorical data arising from sample surveys. The paper should be extremely useful for survey analysts who wish to take into account the impact of survey designs on the practical aspects of the analysis of survey data. In particular, it is important to emphasize that the methods discussed cover two different situations arising in practice: so called *primary analyses*, in which the researcher has all the relevant information at hand, and *secondary analyses*, in which the data provided do not include enough information about the population units to enable the calculation of full covariance matrices of the sample estimators.

The methods covered require the existence of a structural model for the data. There are situations, however, in which it is difficult to specify a single structural model that adequately describes categorical data. In large scale surveys there is often need to screen out many cross classifications at minimal cost. In such cases the use of measures of association is a common alternative. These non parametric methods were extended to sample survey data by Molina and Smith (1986, 1988).

For the primary analysis of survey data the paper concentrates on weighted least squares and Wald tests. The results in Scott, Rao and Thomas (1989) are summarized and the relationship with quasi-likelihood is mentioned. I think that an important conclusion from that paper should be included in this section, namely the need to take into account the survey constraints  $K'p(X\beta) = \pi$  when using quasi-likelihood methods. The reader may not be aware of the importance of the careful choice of the  $g$ -inverse in equation (2.9). Quasi-likelihood methods are now widely used and the relationship with weighted least squares methods is a relevant one. In fact, quasi-likelihood functions represent an interesting alternative for the analysis of survey data. However, there are practical problems since the method requires that we specify the covariance matrix as a function of  $p$ , the variance function. Quasi-likelihoods are largely determined by these variance functions (see, *e.g.*, Morris 1982, and Jørgensen 1987). If a matrix of estimates is given instead of a function, the method would be equivalent to the use of a normal distribution.

Most of the paper is devoted to methods involving *pseudo likelihoods*. Since secondary analyses constitute the most common situation in practice, the methods presented are likely to be extensively used by survey analysts. I would like, however, to discuss some alternatives.

The study of the impact of survey design on Guerrero and Johnson's (1982) transformation models is an important addition to the literature. However, Nelder and Pregibon (1987) have proposed a family of functions, the *extended quasi-likelihoods*, that avoid some important disadvantages of transformation models and can be fitted with GLIM. If design effects are available, their methods can be adapted to survey data by incorporating them either in the variance functions or in the form of weights. Alternatively, design variables may be used to adjust the dispersion parameter in the models. In both cases, one advantage is that we can use the goodness of fit statistics and standard errors produced by GLIM under these models to examine the data without the introduction of further corrections.

These comments apply in general to the use of pseudo-likelihoods. The effect of ignoring the survey design may be treated as an increase or decrease in the expected variability that may be modelled as overdispersion or underdispersion by means of quasi-likelihoods or extended quasi-likelihoods. See, *e.g.*, Pocock *et al.* (1981), Breslow (1984), Williams (1982), among

<sup>1</sup> E.A. Molina, Universidad Simon Bolivar, Caracas and University of Southampton, United Kingdom.

others. As an example, I reanalyzed the data in Table 1. The analysis given in the paper is the correct one, since it incorporates the true covariance matrix. Suppose, however, that this matrix is not available and that only the cell design effects are at hand. Using GLIM I fitted model (2.12) with a Poisson error ignoring the sampling scheme. This gives  $X^2 = 5.68$ ,  $G^2 = 5.67$ . The Rao and Scott (1987) approximation for the chi square statistic gives  $X^2(\delta) = 5.68/2.25 = 2.52$ . For the independence model the uncorrected values are  $X^2 = 18.22$ ,  $G^2 = 18.22$ , and the correction gives  $X^2(\delta) = 18.22/1.65 = 11.04$ . What can be done if the deffs are not available? A simple quasi-likelihood approach to overdispersion is to estimate the mean deviance for the larger model,  $D = 5.68/3 = 1.89$ , and to use the inverse of this value as a weight (or as a new scale parameter). This give  $X^2 = 3.01$  for model (2.12) and  $X^2 = 9.65$  for the independence model. The correct approach here is to use the excess in deviance (the difference between the log-likelihood ratio statistics) to test  $\gamma = 0$ , since  $G^2$  will equate the degrees of freedom for the larger model. The value is 6.65, which is just significant at the 1% level. Both analyses are in agreement with the correct analysis given in the paper, but in other situations it may not be so. The quasi-likelihood model presented here is equivalent to assuming that the actual covariance matrix is a multiple of the one obtained under multinomial sampling, a model that may perform badly in several situations. The advantage is that it can be used when the only information available is that given by the variability inherent in the data, and the analysis performed in a standard statistical package like GLIM. If the deffs are available, other models involving them may be proposed, and a paper is in preparation.

There is, however, no completely satisfactory substitute for an analysis involving the actual covariance matrix. The objective of this contribution is to highlight other possibilities when the full covariance matrix is not known. Quasi-likelihoods offer a fertile ground for further exploration, particularly in relation to survey data. The paper under discussion presents several alternatives and is an important contribution to the field.

#### ADDITIONAL REFERENCES

- JØRGENSEN, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B* 127-162.
- MOLINA, E.A., and SMITH, T.M.F. (1986). The effect of sample design on the comparison of associations. *Biometrika* 73, 23-33.
- MOLINA, E.A., and SMITH, T.M.F. (1988). The effect of sampling on operative measures of association. *International Statistical Review* 56, 235-242.
- MORRIS, C.N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics* 10, 65-80.
- NELDER, J.A., and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* 74, 221-232.
- POCOCK, S.J., COOK, D.G., and BERESFORD, S.A.A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Applied Statistics* 31, 286-295.
- WILLIAMS, D.A. (1982). Extra binomial variation in logistic-linear models. *Applied Statistics* 31, 144-148.

## RESPONSE FROM THE AUTHORS

We thank the three discussants, Fay, Molina and Skinner, for their useful comments and for suggesting additional methods useful in the analysis of cross-classified data from complex sample surveys.

### (i) Response to comments of R.E. Fay

We agree with Fay that replication methodology and associated jackknife chi-squared tests provide viable alternatives to the methods presented here, provided the survey design permits the use of a replication method such as the jackknife or the balanced half-sample replication. His CPLX program indeed offers a comprehensive analysis option whenever estimates are available at the individual replicate level. Also, as noted in the Introduction, Fay's jackknife tests and Rao-Scott corrections have performed well under quite general conditions in simulation studies, unlike the Wald tests based on weighted least squares. Rao-Scott corrections are, however, also applicable to survey designs not permitting the use of a replication method.

The software systems for the Canada Health Survey and the Canadian Labour Force Survey were set up to readily provide the estimated covariance matrix of cell estimates but not the replicate level estimates. As a result, the implementation of jackknife tests would have required some changes in the software systems.

We are also thankful to Fay for pointing out that the methods presented here, and the analogues from replication theory, can also handle some problems of inference from complex designed experiments involving clustering and stratification. Indeed, one of us (J.N.K. Rao) recently used Rao-Scott type methods to fit dose-response models and to test hypotheses in teratological studies involving animal litters as experimental units (Rao and Colin 1989). These methods do not assume specific models for the intra-litter correlations, unlike other methods proposed in this area.

We considered Box-Cox transformation models since Guerrero and Johnson (1982) obtained significantly better fits on some Mexican data compared to the logit model. We agree with Fay, however, that the Box-Cox models should not be applied without consideration of alternative strategies, such as transforming the predictors. As noted by Fay, the Box-Cox approach would be useful in these cases where it would lead to additive models on the transformed scale while the logit model would require interaction terms.

### (ii) Response to comments of E.A. Molina

Molina is correct in saying that measures of association can be used to screen out many cross classifications at minimal cost. His joint work with T.M.F. Smith on extending the classical theory for measures of association to sample survey data involving clustering and stratification is an important contribution.

As noted in the Introduction, we assumed throughout the paper that the user has access to a full estimated covariance matrix of cell estimates. However, such detailed information is often not available for secondary analyses, and in fact even cell deffs may not be available, as pointed out by Molina. In the latter case, Rao and Scott (1987) showed that an  $F$  statistic used in GLIM for testing a nested hypothesis, such as  $\gamma = 0$  given the model (2.12), is asymptotically valid whenever the covariance matrix of cell estimates,  $\hat{V}$ , is proportional to the multinomial covariance matrix,  $\hat{P}$ . The  $F$ -test, however, is less powerful than the Rao-Scott tests, unless the denominator degrees of freedom are high. In the latter case, the  $F$  test might work well even if the condition  $\hat{V} \propto \hat{P}$  is not satisfied (see Rao and Scott 1987, p. 392).

For the data in Table 1,  $F = 6.63$  for testing  $\gamma = 0$  given the model (2.12), which is not significant at the 5% level compared to  $F_{1,3}(0.05) = 10.01$ , the upper 1% of the  $F$  distribution with 1 and 3 degrees of freedom (d.f.). On the other hand, the Wald test  $W_1$  and the Rao-Scott test, both requiring detailed information on the estimated covariance matrix, are significant at the 1% level compared to  $\chi_1^2(0.01) = 6.63$ . The  $F$ -test, therefore, appears to be less powerful here since the denominator d.f. is only 3. Molina's proposed test is, in fact, equal to  $F$ , but he was treating  $F$  as a  $\chi^2$  variable with 1 d.f. which may not be valid due to small denominator d.f.

The GLIM method does not provide a statistic for testing the goodness-of-fit of a model. Some information on the design effects is necessary for getting a valid test of goodness-of-fit.

### (iii) Response to comments of C.J. Skinner

Skinner noted that the test of equality of two logistic regression models in Section 4 might be formulated as a test of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987), using dummy  $x$ -variables. The framework of Roberts, Rao and Kumar, however, assumes one fixed sample size  $n$  whereas in Section 4 we have two fixed sample sizes  $n_1$  and  $n_2$  for the two time periods. As a result, their results would need careful modification in order to be applicable to the present case of test of equality of two logistic regression models. Moreover, the dummy variable approach would involve the determination of estimates of  $2s$  parameters iteratively, whereas the approach in Section 4 requires two iterative solutions, each involving only  $s$  parameters. Thus, the dummy variable approach could lead to convergence problems if  $s$  is not small.

We treated WLS with singular covariance matrices separately in Section 2 since the logit-type models in the remaining sections do not involve singular covariance matrices. WLS can also be applied to logit-type models but the resulting estimators and associated Wald tests may be unstable if the degrees of freedom associated with the estimated covariance matrix,  $\hat{V}_P$ , are low (criterion C3 of Skinner). The six criteria proposed by Skinner for comparing WLS and PML are very useful. We prefer PML mainly on the basis of criterion C3. Regarding the relative efficiency of WLS and PML estimators under complex designs, no general results are available, but WLS estimators are not likely to be significantly more efficient (and in fact, may be less efficient) if the degrees of freedom associated with the estimated covariance matrix are low. Clearly, further research on the relative efficiency of WLS and PML estimators would be useful.

### ADDITIONAL REFERENCES

- RAO, J.N.K., and COLIN, D. (1988). Fitting dose-response models and hypothesis testing in teratological studies. Technical Report No. 116, Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa, Ottawa, Ontario.