# GENERALIZED RESPONSE MODELS FOR SURVEY DATA

Georgia Roberts, Statistics Canada

## 1. INTRODUCTION

Extensive literature and software exists for statistical methods of analyzing cross-classified categorical data. For examining the relationships among the variables determining the cross-classifications, hierarchical loglinear models (see Bishop, Fienberg and Holland (1975)) have been a popular and effective tool. On the other hand, researchers wishing to distinguish between response variables and explanatory variables have used models that describe the effect of a set of explanatory variables on a response variable. Most prevelant has been the case of a two-category or dichotomous response incorporated into a logit model (see, for example, Cox (1970) and Fienberg (1977)). In the situation of a polytomous response variable (i.e. 2 or more response categories), a variety of models have been proposed, some of which take account of an ordering present in the response variable (see Agresti (1984) and McCullough and Nelder (1983)).

Generally, these statistical methods have been developed under the assumption of multinomial sampling. In particular, for response models, an independent multinomial distribution is usually assumed on the response at each combination of levels of explanatory variables. These methods are, therefore, inappropriate for analyzing multidimensional tables arising from survey data because of the clustering and stratification in the survey design, which violate the multinomial assumption. As a result, such methods need to be adjusted to take account of the survey design in order that valid inferences may be made.

Rao and Scott (1984) and Roberts, Rao and Kumar (1986) have shown that the standard Pearson chi-squared statistic $X^2$, if applied to survey data for testing goodness of fit of loglinear or logit models, is asymptotically distributed as a weighted sum of $\chi^2_1$ random variables, rather than as a central chi-squared variable as would be the case under multinomial sampling. This result demonstrates that the survey design can have a substantial impact on the significance level of the test. To take account of the survey design, Rao and Scott (1984) and Roberts, Rao and Kumar (1986) have developed simple corrections to $X^2$ based on certain generalized design effects (deffs). Some of these corrections may be implemented even when the full estimated covariance matrix of cell estimates is not available.

The aim of this article is to develop a method of taking account of the survey design when utilizing a general class of polytomous response models with cross-classified categorical data. In Section 2, the asymptotic distributions of estimated parameters and of the test statistic $X^2$ are obtained. Simple corrections to $X^2$ requiring varying quantities of survey design information, are obtained in Section 3. Section 4 contains some specific examples of models to which the results of Sections 2 and 3 apply - in particular, some dichotomous response models and a polytomous response model with ordered response categories. Finally, Section 5 contains an application of the results of Sections 2 and 3 to a set of data from the Canada Health Survey, 1978-1979.

## 2. ASYMPTOTIC PROPERTIES OF ESTIMATES AND TEST STATISTICS

### 2.1 Assumptions and Notation

Suppose that the finite population of interest of size N is partitioned into R mutually exclusive exhaustive domains, the $i^{th}$ of size $N_i$. (The partitioning generally consists of combinations of categories of the explanatory variables, so that i actually represents a multiple subscript.) In the $i^{th}$ domain, if $N_{ij}$ of the units fall into the $j^{th}$ response category ($j = 1, 2, \ldots J+1$), then the proportion of units with the $j^{th}$ response is $P_{j(i)} = N_{ij}/N_i$ (where

$\sum_{j=1}^{J+1} P_{j(i)} = 1$). Let $\underline{P}_i$ denote the J x 1 vector

$$(2.1)\qquad \underline{P}_i = (P_{1(i)}, P_{2(i)}, \ldots, P_{J(i)})',$$

and $\underline{P}$ denote the JR x 1 vector

$$(2.2)\qquad \underline{P} = (\underline{P}'_1, \underline{P}'_2, \ldots, \underline{P}'_R)'.$$

Suppose that a sample $\tilde{s}$ of n ultimate units is drawn from the population according to a specified sampling plan $p(\tilde{s})$. Let $\hat{N}$, $\hat{N}_i$ and $\hat{N}_{ij}$ be survey estimates of N, $N_i$ and $N_{ij}$ respectively, such that

(a) $\underline{w} = \underline{W} + o_p(n^{-1/2})$ where $\underline{w} = (w_1, w_2, \ldots, w_R)'$, $w_i = \hat{N}_i/\hat{N}$,

(b) $\sqrt{n}(\hat{\underline{P}} - \underline{P})$ converges in distribution to a $N(\underline{0}, \Sigma)$ random vector as $n \to \infty$ where $\hat{P}_{j(i)} = \hat{N}_{ij}/\hat{N}_i$ is the usual survey estimate of $P_{j(i)}$ and $\hat{\underline{P}}$ is obtained from (2.1) by replacing $P_{j(i)}$ with $\hat{P}_{j(i)}$.

A general class of polytomous response models for the proportions $P_{j(i)}$ is given by

$$(2.3)\qquad M_1:\ P_{j(i)} = g_{ij}(\underline{\theta}) \qquad j = 1, 2, \ldots, R; \quad j = 1, 2, \ldots, J+1,$$

where $\underline{\theta}$ is an r-vector of unknown parameters ($r \le JR$) and the $g_{ij}(\underline{\theta})$ are functions of known form satisfying the regularity conditions described in Bishop, Fienberg and Holland (1975, p. 510). Because of the additive restrictions on the $P_{j(i)}$, model $M_1$ may be denoted equivalently by

$$(2.4)\qquad M_1:\ \underline{P} = \underline{g}(\underline{\theta}),$$

where $\underline{g}(\underline{\theta}) = \underline{g} = (\underline{g}'_1, \underline{g}'_2, \ldots, \underline{g}'_R)'$, $\underline{g}_i = (g_{i1}, g_{i2}, \ldots, g_{iJ})'$ and $g_{ij}$ represents $g_{ij}(\underline{\theta})$.

### 2.2 Motivation of Methodology

Under independent multinomial sampling in each domain, the likelihood equations for model $M_1$ are

$$(2.5)\qquad \sum_{i=1}^{R} \frac{n_i}{n} \sum_{j=1}^{J+1} \frac{n_{ij}}{n_i} \frac{1}{g_{ij}(\hat{\underline{\theta}})} \left.\frac{\partial g_{ij}(\underline{\theta})}{\partial \theta_k}\right|_{\underline{\theta}=\hat{\underline{\theta}}} = 0$$

$$k=1,2,\ldots,r$$

where $n_i$ and $n_{ij}$ are, respectively, the size of sample and the number of sampled units with the $j^{th}$ response in the $i^{th}$ domain, where $\sum_{j=1}^{J+1} n_{ij} = n_i$ and $\sum_{i=1}^{R} n_i = n$. As well, the Pearson chi-squared statistic for testing H: $\underline{P}=\underline{g}(\underline{\theta})$ is

$$(2.6)\quad X_c^2 = \sum_{i=1}^{R} \sum_{j=1}^{J+1} (n_{ij} - n_i g_{ij}(\hat{\underline{\theta}}))^2 / n_i g_{ij}(\hat{\underline{\theta}}) ,$$

where $\hat{\underline{\theta}}$ is the solution of (2.5). It is a well-known fact that, given H, $X_c^2$ is asymptotically distributed as a central chi-squared variable.

For a general sampling plan, since appropriate likelihood equations are difficult to obtain, pseudo likelihood equations are produced from (2.5) by replacing $n_i/n$ with $w_i$ and $n_{ij}/n_i$ with $\hat{P}_{j(i)}$, yielding

$$(2.7)\quad \sum_{i=1}^{R} w_i \sum_{j=1}^{J+1} \frac{\hat{P}_{j(i)}}{g_{ij}(\hat{\underline{\theta}})} \frac{\partial g_{ij}(\underline{\theta})}{\partial \theta_k}\Bigg|_{\underline{\theta}=\hat{\underline{\theta}}} = 0 \qquad k=1,2,\ldots,r.$$

The solution of (2.7) for $\hat{\underline{\theta}}$, and the resulting $g_{ij}(\hat{\underline{\theta}})$ are called the pseudo maximum likelihood estimators (pseudo mle's) for $\theta$ and $P_{j(i)}$ respectively. The consistency of $\hat{\underline{P}}$ and $\underline{w}$ ensures the consistency of $\hat{\underline{\theta}}$ and $\hat{\underline{g}} = \underline{g}(\hat{\underline{\theta}})$. The statistic parallel to $X_c^2$ obtained from (2.5) by the same substitutions of $w_i$ and $\hat{P}_{j(i)}$ for $n_i/n$ and $n_{ij}/n_i$, and having $\hat{\underline{\theta}}$ the solution of (2.7), is

$$(2.8)\quad X^2 = n \sum_{i=1}^{R} w_i \sum_{j=1}^{J+1} (\hat{P}_{j(i)} - g_{ij}(\hat{\underline{\theta}}))^2 / g_{ij}(\hat{\underline{\theta}}).$$

As will be shown in section 2.5, $X^2$ is generally not asymptotically distributed as a central chi-squared variable. However, simple modifications to $X^2$ suggested by its asymptotic distribution do produce plausible test statistics.

Both the pseudo likelihood equations (2.7) and the test statistic $X^2$ may be expressed in matrix form in the following way. If $Q_i$ = diag $(\underline{g}_i) - \underline{g}_i\underline{g}_i'$ and Q denotes the JRxJR block-diagonal matrix with the $Q_i$'s on the diagonal, then (2.8) is equivalent to

$$(2.9)\qquad X^2 = n(\hat{\underline{P}} - \hat{\underline{g}})'\hat{Q}^{-1}(D(\underline{w})\otimes I_J)(\hat{\underline{P}} - \hat{\underline{g}}) ,$$

where $D(\underline{w})$ = diag $(\underline{w})$, $I_J$ is the JxJ identity matrix, $\otimes$ denotes the direct product operator and $\hat{Q}$ is obtained from Q by replacing $\underline{g}_i$ with $\hat{\underline{g}}_i$. As well, the pseudo likelihood equations are

$$(2.10)\quad \left(\frac{\partial \underline{g}}{\partial \underline{\theta}}\right)'\Bigg|_{\underline{\theta}=\hat{\underline{\theta}}} (D(\underline{w})\otimes I_J)\hat{Q}^{-1}\hat{\underline{P}} = \left(\frac{\partial \underline{g}}{\partial \underline{\theta}}\right)'\Bigg|_{\underline{\theta}=\hat{\underline{\theta}}} (D(\underline{x})\otimes I_J)\underline{1}_{JR}$$

where $D(\underline{x})$ = diag $(\underline{x})$, $x_i = w_i(1 - \underline{1}_J'\hat{\underline{g}}_i)^{-1}$ , $\underline{1}_{JR}$ and $\underline{1}_J$ are unit vectors of lengths JR and J respectively, and $G = \frac{\partial \underline{g}}{\partial \underline{\theta}}$ is the JRxr matrix of the derivatives of the functions $g_{ij}(\underline{\theta})$ with respect to the elements of $\theta$. The regularity conditions require that G be of full rank.

## 2.3 Nested Models

Suppose that the vector $\underline{\theta} = (\underline{\theta}'_1, \underline{\theta}'_2)'$, where $\underline{\theta}_1$ is sx1 and $\underline{\theta}_2$ is ux1 (s + u = r). If we were to test the hypothesis H(2/1):$\underline{\theta}_2 = \underline{0}$, given $M_1$, we would have the reduced model

$$(2.11)\qquad M_2:\ \underline{P} = \underline{g}(\underline{\theta}_1)$$

where "$\underline{g}(\underline{\theta}_1)$" denotes "$\underline{g}(\underline{\theta}_1, \underline{0})$". Using a similar extrapolation from independent multinomial sampling to a general sampling plan, as done in section 2.2 for model $M_1$, the pseudo mle's $\hat{\hat{\underline{\theta}}}_1$ and $\hat{\hat{\underline{g}}} = \underline{g}(\hat{\hat{\underline{\theta}}}_1)$ of $\underline{\theta}_1$ and $\underline{P}$ respectively under $M_2$ are the solution of the pseudo likelihood equations

$$(2.12)\quad \left(\frac{\partial \underline{g}}{\partial \underline{\theta}_1}\right)'\Bigg|_{\underline{\theta}=(\hat{\underline{\theta}}_1,\underline{0})} (D(\underline{w})\otimes I_J)\hat{\hat{Q}}^{-1}\hat{\underline{P}} = \left(\frac{\partial \underline{g}}{\partial \underline{\theta}_1}\right)'\Bigg|_{\underline{\theta}=(\hat{\underline{\theta}}_1,\underline{0})} (D(\underline{x}_1)\otimes I_J)\underline{1}_J$$

where $\hat{\hat{Q}} = Q(\hat{\hat{\underline{\theta}}}_1, \underline{0})$ and $D(\underline{x}_1)$ = diag$(\underline{x}_1)$ with $x_{1i} = w_i(1 - \underline{1}'_J \underline{g}_i(\hat{\hat{\underline{\theta}}}_1))^{-1}$. As well, the Pearson chi-squared statistic for testing H(2/1) is given by

$$(2.13)\quad X^2(2/1) = n \sum_{i=1}^{R} w_i \sum_{j=1}^{J+1} (g_{ij}(\hat{\underline{\theta}}) - g_{ij}(\hat{\hat{\underline{\theta}}}_1))^2 / g_{ij}(\hat{\hat{\underline{\theta}}}_1)$$

$$= n(\underline{g}(\hat{\underline{\theta}}) - \underline{g}(\hat{\hat{\underline{\theta}}}_1))'\hat{\hat{Q}}^{-1}(D(\underline{w})\otimes I_J)(\underline{g}(\hat{\underline{\theta}}) - \underline{g}(\hat{\hat{\underline{\theta}}}_1)) .$$

Since $G = \frac{\partial \underline{g}}{\partial \underline{\theta}} = \left(\frac{\partial \underline{g}}{\partial \underline{\theta}_1}, \frac{\partial \underline{g}}{\partial \underline{\theta}_2}\right)$ is JRxr of full rank r, $G_1 = \frac{\partial \underline{g}}{\partial \underline{\theta}_1}$ and $G_2 = \frac{\partial \underline{g}}{\partial \underline{\theta}_2}$ must be of full ranks s and u respectively.

It may be shown easily that testing of goodness of fit of model $M_1$ is a special case of testing of a nested model, where nesting is within a saturated model, i.e. a model where s + u = JR.

## 2.4 Asymptotic Distribution of $\hat{\underline{\theta}}$ and $\underline{g}(\hat{\underline{\theta}})$

Lemma 2.4 Under H and the assumed regularity conditions for the $g_{ij}$,

$$(2.14)\qquad \hat{\underline{\theta}} - \underline{\theta} = (G'\nabla G)^{-1}G'\nabla(\hat{\underline{P}} - \underline{g}(\underline{\theta})) + o_p(n^{-1/2})$$

where $\nabla$ is the JRxJR matrix defined by

$$(2.15)\qquad \nabla = (D(\underline{W})\otimes I_J)Q^{-1} .$$

Proof: See Roberts (1985)

Since it was assumed that $\sqrt{n}(\hat{\underline{P}} - \underline{g}(\underline{\theta})) \approx N(\underline{0},\Sigma)$, where "$\approx$" denotes "asymptotically distributed as", the asymptotic covariance matrix of $\hat{\underline{P}}$ is

$$(2.16)\qquad C(\hat{\underline{P}}) = \Sigma/n .$$

This fact, together with (2.14), implies that the asymptotic covariance matrix of $\hat{\underline{\theta}}$ is

$$(2.17)\qquad C(\hat{\underline{\theta}}) = n^{-1}(G'\nabla G)^{-1}(G'\nabla\Sigma\nabla G)(G'\nabla G)^{-1} .$$

In the case of product multinomial sampling, $\Sigma = \nabla^{-1}$ and $C(\hat{\underline{\theta}})$ reduces to $C_M(\hat{\underline{\theta}}) = n^{-1}(G'\nabla G)^{-1}$ .

Because of the assumed regularity conditions and the lemma above,

$$(2.18)\quad \underline{g}(\hat{\underline{\theta}}) - \underline{g}(\underline{\theta}) = G(\hat{\underline{\theta}} - \underline{\theta}) + o_p(n^{-\frac{1}{2}})$$

$$= G(G'\nabla G)^{-1}G'\nabla(\hat{\underline{P}} - \underline{g}(\underline{\theta})) + o_p(n^{-\frac{1}{2}}) ,$$

so that the asymptotic covariance matrix of $g(\hat{\underline{\theta}})$ is

(2.19) $C(g\hat{\underline{\theta}}) = GC(\hat{\underline{\theta}})G'$ .

As well, the vector of residuals $\hat{\underline{P}} - g(\hat{\underline{\theta}})$, often useful in detecting model deviations, has asymptotic covariance matrix

(2.20) $C(\hat{\underline{P}} - g(\hat{\underline{\theta}})) = n^{-1} A \Sigma A'$ ,

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

Consistent estimators of the above covariance matrices are obtained by replacing $\underline{\theta}$ by $\hat{\underline{\theta}}$ and $\Sigma$ by the survey estimate $\hat{\Sigma}$ (when available).

### 2.5 Asymptotic Distribution of the Test Statistics $X^2$ and $X^2(2/1)$

From (2.18),

(2.21) $g(\hat{\underline{\theta}}) - g(\underline{\theta}) = G_1(\hat{\underline{\theta}}_1 - \underline{\theta}_1) + G_2(\hat{\underline{\theta}}_2 - \underline{\theta}_2) + o_p(n^{-\frac{1}{2}})$

under H. Similarly, under H(2/1),

(2.22) $g(\hat{\hat{\underline{\theta}}}_1) - g(\underline{\theta}) = G_1(\hat{\hat{\underline{\theta}}}_1 - \underline{\theta}_1) + o_p(n^{-\frac{1}{2}})$ .

Thus, under H(2/1),

(2.23) $g(\hat{\underline{\theta}}) - g(\hat{\hat{\underline{\theta}}}_1) = G_1(\hat{\underline{\theta}}_1 - \underline{\theta}_1) + G_2\hat{\underline{\theta}}_2 - G_1(\hat{\hat{\underline{\theta}}}_1 - \underline{\theta}_1) + o_p(n^{-\frac{1}{2}})$

However, it may be shown (see Roberts (1985)) that

(2.24) $(\hat{\hat{\underline{\theta}}}_1 - \underline{\theta}_1) = (\hat{\underline{\theta}}_1 - \underline{\theta}_1) + (G_1'\nabla G_1)^{-1}(G_1'\nabla G_2)\hat{\underline{\theta}}_2 + o_p(n^{-\frac{1}{2}})$ ,

which, when substituted into (2.23) yields

(2.25) $g(\hat{\underline{\theta}}) - g(\hat{\hat{\underline{\theta}}}_1) = (I - G_1(G_1'(G_1'\nabla G_1)^{-1}G_1'\nabla)G_2\hat{\underline{\theta}}_2 + o_p(n^{-\frac{1}{2}})$ .

Theorem 2.5

Under H(2/1): $\underline{\theta}_2 = \underline{0}$ (given $M_1$),

(2.26) $X^2(2/1) = n\hat{\underline{\theta}}_2'(H_2'\nabla H_2)\hat{\underline{\theta}}_2 + o_p(n^{-\frac{1}{2}})$ ,

where $\hat{\underline{\theta}} = (\hat{\underline{\theta}}_1, \hat{\underline{\theta}}_2)$ is the pseudo mle of $\underline{\theta}$ under $M_1$, and

(2.27) $H_2 = (I - G_1(G_1'\nabla G_1)^{-1}G_1'\nabla)G_2$ .

Furthermore,

(2.28) $X^2(2/1) \approx \sum_{i=1}^{u} \delta_i Z_i^2$

where the $Z_i^2$ are independent $\chi_1^2$ variables and the $\delta_i$ are the eigenvalues of

(2.29) $nC(\hat{\underline{\theta}}_2)(H_2'\nabla H_2) = (H_2'\nabla H_2)^{-1}(H_2'\nabla\Sigma\nabla H_2)$ .

Proof: See Roberts (1985).

As a special case, under product multinomial sampling, $\Sigma = \nabla^{-1}$ and $(H_2'\nabla H_2)^{-1}(H_2'\nabla\Sigma\nabla H_2) = I$ so that $\delta_i = 1$, $i = 1,2,\ldots,u$ and we get the standard result that $X^2(2/1) \approx \chi_u^2$ under H(2/1).

It should be noted that the asymptotic covariance matrix of $H_2'\nabla\hat{\underline{P}}$ is $n^{-1}(H_2'\nabla\Sigma\nabla H_2)$ under the survey design and $n^{-1}(H_2'\nabla\nabla^{-1}\nabla H_2) = n^{-1}(H_2'\nabla H_2)$ under product multinomial sampling, so that the $\delta_i$ are the "generalized design effects" of the vector $H_2'\nabla\hat{\underline{P}}$, as defined by Rao and Scott (1981).

The asymptotic distribution of $X^2$, the Pearson chi-squared statistic defined in (2.9) for testing the goodness of fit of model $M_1$, may be obtained as a special case of Theorem 2.5. Specifically,

$$X^2 \approx \sum_{i=1}^{JR-r} \delta_i Z_i^2$$

with the $Z_i^2$ being independent $\chi_1^2$ variables and the $\delta_i$ being the eigenvalues of $(H'\nabla H)^{-1}(H'\nabla\Sigma\nabla H)$ where $H = (I - G(G'\nabla G)^{-1}G'\nabla)G_{22}$ and $G_{22}$ is any JRxJR-r matrix of rank JR-r such that $(G\ G_{22})$ is JRxJR of rank JR. Furthermore, the eigenvalues of $(H'\nabla H)^{-1}(H'\nabla\Sigma\nabla H)$ are equivalent to the eigenvalues of the more easily computable matrix $(E'\nabla^{-1}E)^{-1}(E'\Sigma E)$, where E is any (JRxJR-r) matrix of rank JR-r with $E'G = \underline{0}$ (see Appendix).

## 3. ADJUSTMENTS TO $X^2$ AND $X^2(2/1)$

The asymptotic distribution of $X^2(2/1)$ (and of $X^2$, as a special case) is a weighted sum of chi-squared variables for which there are generally no published tables of critical values. It is important, therefore, to have adjustments to $X^2(2/1)$ that can be practically implemented with survey data. Two such corrections that require knowledge of the full estimated covariance matrix of $\hat{\underline{P}}$ (say $\hat{\Sigma}$) are the following.

Compare $X^2(2/1) / \hat{\delta}.$ to the critical values of a $\chi_u^2$ variable, where

(3.1) $u\delta. = \sum_{i=1}^{u} \delta_i = tr\ (H_2'\nabla H_2)^{-1}(H_2'\nabla\Sigma\nabla H_2)$

and $\hat{\delta}.$ is the survey estimate of $\delta.$. The rationale for this approach is the fact that the asymptotic mean of $X^2(2/1) / \hat{\delta}.$ is equal to the mean of a $\chi_u^2$ variable.

A more accurate approach, particularly when the c.v. of the $\delta_i$'s is large, is a Satterthwaite approximation, which consists of comparing

(3.2) $X_S^2 = X^2(2/1)\ /\ [\hat{\delta}.\ (1 + \hat{a}^2)]$

to a $\chi_\alpha^2$ variable where $\alpha = u/(1 + \hat{a}^2)$ and $\hat{a}^2$ is the squared c.v. of the $\delta_i$'s. The first two moments of the asymptotic distribution of $X_S^2$ equal those of the $\chi_\alpha^2$ variable. As demonstrated in Rao and Scott (1984), the individual $\delta_i$'s need not be evaluated to compute $X_S^2$.

In practice, such as in secondary analyses from published reports, $\hat{\delta}.$ cannot be calculated since the full estimated covariance matrix $\hat{\Sigma}$ of $\hat{\underline{P}}$ is unavailable. However, a possible alternative approach is to compare to the critical values of $\chi_u^2$ the value of a statistic of the form $X^2(2/1) / \hat{c}$, where c is an upper bound on $\delta.$. To be workable, calculation of c must require less information than $\hat{\Sigma}$. One method to obtain such a bound is described below.

Since $H_2'\nabla H_2$ is symmetric, there exists a nonsingular matrix B and diagonal matrix D such that $BH_2'\nabla H_2B' = D$ (see C.R. Rao (1965), p. 20). If $Y = BH_2' = (\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_u)'$, then $Y\nabla Y' = D$ and thus $\underline{y}_i'\nabla\underline{y}_j = 0$ for $i \neq j$. Therefore, (3.3) $$u\delta. = \mathrm{tr}\,(B')^{-1}(H_2'\nabla H_2)^{-1}B^{-1}B(H_2'\nabla\Sigma\nabla H_2)B'$$

$$= \mathrm{tr}\,(Y\nabla Y')^{-1}(Y\nabla\Sigma\nabla Y')$$

$$= \sum_{i=1}^{u} [\underline{y}_i'(\nabla\Sigma\nabla)\underline{y}_i] \;/\; \underline{y}_i'\nabla\underline{y}_i \;.$$

Since $Q_i$ is a positive definite covariance matrix for each i, $\nabla$ is also positive definite, so that, extending a matrix result in C.R. Rao (1965), p. 63, it follows that

(3.4) $$u\delta. = \sum_{i=1}^{u} [\underline{y}_i'(\nabla\Sigma\nabla)\underline{y}_i] \;/\; \underline{y}_i'\nabla\underline{y}_i \le \lambda_1 + \lambda_2 + \dots + \lambda_u \;,$$

where $\lambda_1 \geq \lambda_2 \dots \geq \lambda_{JR} > 0$ are eigenvalues of $\nabla\Sigma$. Since the eigenvalues of $\nabla\Sigma$ are independent of any hypothesis, if they were reported along with the published tables, the required number for bounding $\delta.$ for any particular hypothesis could be chosen.

If the estimated eigenvalues of $\nabla\Sigma$ are not available, an upper bound may still be estimated by noting that

(3.5) $$u\delta. \le \sum_{i=1}^{u} \lambda_i \le \sum_{i=1}^{JR} \lambda_i = JR\lambda.$$

and

(3.6) $$JR\lambda. = \mathrm{tr}\,\nabla\Sigma = \sum_{i=1}^{R}\sum_{j=1}^{J} v_{jj(i)}/[P_{j(i)}(1 - P_{j(i)})/w_i]$$

$$= \sum_{i=1}^{R}\sum_{j=1}^{J} D_{j(i)} = JRD.$$

where $v_{jj(i)}$ and $D_{j(i)}$ are the variance and deff respectively of $\hat{P}_{j(i)}$. However, this upper bound on $\delta.$ may not be satisfactory if u is small compared to JR.

When model $M_1$ is saturated, that is s + u = JR, (3.5) and (3.6) lead to

(3.7) $$(JR\text{-}s)\,\delta. \le JRD. \,,$$

so that D. is a satisfactory upper bound on $\delta.$ if s << JR.

## 4. EXAMPLES

Since an extensive variety of response models satisfy the assumptions of section 2.1, the results of sections 2 and 3 have wide applicability. The first example in this section consists of a group of binary response models, including the logit. As well, a model suggested by McCullagh and Nelder (1983) to be particularly applicable to data for which there is a natural ordering of the categories of the response variable, is examined.

### 4.1 Binary Response Models

For a binary response, say "success or failure" or "alive or dead", J=1. Thus, $\underline{P}$ is an Rx1 vector of the scalar quantities $P_i = P_{1(i)}$ from each domain, $\underline{g}(\underline{\theta}) = (g_1(\underline{\theta}), \dots, g_R(\underline{\theta}))'$, $Q = D(\underline{g})D(\underline{1} - \underline{g})$, and $\nabla = D(\underline{W})D(\underline{g})^{-1}D(\underline{1} - \underline{g})^{-1}$. Suppose that the models are restricted even further to those for which a monotone differentiable function f of $P_i$ exists such that $f_i = f(P_i) = \underline{x}_i'\underline{\theta}$. Model $M_1$ then has the form

(4.1) $$M_1\colon f_i = \underline{x}_i'\underline{\theta} = \underline{x}_{1i}'\underline{\theta}_1 + \underline{x}_{2i}'\underline{\theta}_2 \,,$$

and $M_2$ is $f_i = \underline{x}_{1i}'\underline{\theta}_1$. Then G = DX, where D is a diagonal matrix with $i^{th}$ diagonal element equal to $(\partial f_i/\partial P_i)^{-1}$ and X is the Rxr matrix defined by

(4.2) $$X = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_R)' \,.$$

It then follows, by substitution, that $H_2 = D\bar{X}_2$ where $\bar{X}_2 = (I - X_1(X_1'D\nabla DX_1)^{-1}X_1'D\nabla D)X_2$, where $X_1 = (\underline{x}_{11}, \underline{x}_{12}, \dots, \underline{x}_{1R})'$ and $X_2 = (\underline{x}_{21}, \underline{x}_{22}, \dots, \underline{x}_{2R})'$. As well, the asymptotic distribution of $X^2(2/1)$ is $\sum_{i=1}^{u} \delta_i Z_i^2$ where the $Z_i^2$ are independent $\chi_1^2$ random variables and the $\delta_i$ are the eigenvalues of

(4.3) $$(\bar{X}_2'D\nabla D\bar{X}_2)^{-1}(\bar{X}_2'D\nabla\Sigma\nabla D\bar{X}_2) \,.$$

The logit model follows the form of (4.1) with

(4.4) $$f_i = f(P_i) = \ln|P_i/(1 - P_i)| = \underline{x}_i'\underline{\theta} \,.$$

### 4.2 An Ordered Response Model

Let $C_{j(i)}$ denote the $j^{th}$ cumulative probability in the $i^{th}$ domain, defined by

(4.5) $$C_{j(i)} = \sum_{k=1}^{j} P_{k(i)} \,.$$

The model proposed by McCullagh and Nelder (1983) for the situation of a response variable with ordered categories has the form

(4.6) $$\ln\left[C_{j(i)}/(1 - C_{j(i)})\right] = \nu_j - \underline{\beta}'\underline{y}_i \,,\; j=1,2,\dots,J \,,\; i=1,2,\dots,R \,,$$

where $\nu_j$ and $\beta$ are unknown parameters and $\underline{y}_i$ is a known vector of length $\ell$. It follows from (4.5) that

(4.7) $$\underline{C}_i = L\,\underline{P}_i \,,$$

where $C_i = (C_{1(i)}, C_{2(i)}, \dots, C_{J(i)})$ and L is the JxJ nonsingular matrix of the form

(4.8) $$L = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix} .$$

Letting $\underline{\theta} = (\nu_1, \nu_2, \dots, \nu_J, \underline{\beta}')'$, which is of length J + $\ell$, and recalling that G is the matrix of partial derivatives derived from the model (4.7) expressed in the form $P_{j(i)} = g_{ij}(\underline{\theta})$, it follows that

(4.9) $$G = (I_R \otimes L^{-1})C\,|\underline{1}_R \otimes I_J, A|$$

where C is a JRxJR block-diagonal matrix with the JxJ matrix $D(\underline{C}_i)D(\underline{1} - \underline{C}_i)$ forming the $i^{th}$ block on the diagonal and

(4.10) $$A = \begin{bmatrix} -\underline{1}_J \otimes \underline{y}_1' \\ -\underline{1}_J \otimes \underline{y}_2' \\ \vdots \\ -\underline{1}_J \otimes \underline{y}_R' \end{bmatrix} .$$

With this model, nested hypotheses of interest generally take the form $H(2/1)$: $\beta_2 = 0$ where $\beta = (\beta_1', \beta_2')'$, $\beta_1$ is sx1 and $\beta_2$ is ux1 ($s + u = \ell$). Using Theorem 2.5 and (4.9) the asymptotic distribution of the Pearson chi-squared statistic $X^2(2/1)$ is readily obtained.

## 5. APPLICATION

An illustration of the relative performance of the Pearson chi-squared statistic and some of the modifications suggested in Section 3 is now provided, utilizing data from the Canada Health Survey, 1978-1979 and the ordered response model of section 4.2. The Canada Health Survey was designed to provide reliable information on the health status of Canadians. A complex multi-stage design involving stratification and cluster sampling was employed and the estimates of totals and proportions underwent post-stratification adjustment on age-sex to improve their efficiency.

The data set considered consisted of the estimated counts of females aged 20-64 cross-classified by frequency of breast self-examination (with 3 categories: monthly, quarterly, less often or never), education (with 3 categories: secondary or less, some post-secondary, post-secondary) and age (with 3 categories: 29-24, 25-44, 45-64). The frequency of breast self-examination, which obviously has ordered categories, was taken as the response variable while education and age were explanatory variables, so that the number of responses, J+1, equalled 3 and the number of domains, R, was 9.

Models of the type described in section 4.2 were fitted to the cumulated probabilities in each domain. The model considered for goodness-of-fit was

$$\ln\left[C_{j(ik)} / (1 - C_{j(ik)})\right] = \nu_j + A_i + E_k, \quad j=1,2$$

$$\text{with } \sum_{i=1}^{3} A_i = \sum_{k=1}^{3} E_k = 0$$

where $C_{j(ik)}$ is the $j^{th}$ cumulated probability for the $i^{th}$ age level and the $k^{th}$ education level. One nested hypothesis, given this model, $H(2/1)$: $E_k = 0$ (i.e. no education effect), was considered.

Table 1 gives the values of the Pearson chi-squared statistic and some modifications, plus their estimated asymptotic Type I error rates, for testing goodness-of-fit of the model and the nested hypothesis described above. The asymptotic Type I error rates were estimated through Satterthwaite approximations; as well, the Satterthwaite statistics $X_S^2$ were adjusted so that their values could be compared to the same critical values as the other test statistics.

First, considering the goodness-of-fit statistics, since $X^2$ is larger than $\chi^2_{.05}(12) = 21.03$, the model would be rejected if the sample design was ignored. On the other hand, the value of any of the modified statistics would indicate that the model is adequate. Because of the high cv of the $\hat{\delta}_i$, the Satterthwaite statistic would be a better modification than $X^2/\hat{\delta}_.$. The effect of survey design on the (estimated) asymptotic Type I error rate of $X^2$, $\hat{\alpha}(X^2)$ is quite severe: 0.40 compared to the nominal 0.05. The modification $X^2/\hat{\delta}_.$ performs better, but is still elevated, with $\hat{\alpha} = .10$, while $X^2/C$ is overly conservative, with $\hat{\alpha} = .01$.

For testing of the nested hypothesis, again the model would be rejected if $X^2$ was used as the test statistic, while the values of any of the modified statistics, taking the survey design into account, would indicate that the model was adequate. The effect of the survey design on $\hat{\alpha}(X^2)$ was still fairly severe (0.20 compared to the nominal 0.05). The modification $X^2/\hat{\delta}$ performed quite well $\hat{\alpha} = .06$), due to the low cv of the $\hat{\delta}_i$, while $X^2/C$ was overly conservative $\hat{\alpha} = .0004$).

It is thus clear from this example that the survey design should be taken into account when testing the fit of models to survey data. As well, the greater the amount of survey design information used, the better are the properties of the test statistics.

**TABLE 1**

| | Goodness of fit (age & edn) | Nested Hypothesis (Age only) |
|---|---|---|
| $X^2$ | 35.88 | 6.58 |
| $X^2/\hat{\delta}_.$ | 21.21 | 3.43 |
| $X^2/c$ | 14.38 | 1.23 |
| $X_S^2$ | 15.84 | 3.29 |
| $\hat{\delta}_.$ | 1.69 | 1.92 |
| $cv(\hat{\delta}_i)$ | .94 | .40 |
| $c^{(*)}$ | 2.50 | 5.36 |
| $\hat{\alpha}(X^2)^{(**)}$ | .40 | .21 |
| $\hat{\alpha}(X^2/\hat{\delta}_.)$ | .10 | .06 |
| $\hat{\alpha}(X^2/c)$ | .01 | .0004 |
| Critical value | $\chi^2(12)=21.03$ | $\chi^2(2)=5.99$ |

(*) c is the average of the u largest eigenvalues of $\nabla\Sigma$.

(**) $\hat{\alpha}$ is the estimated asymptotic Type I error rate.

## APPENDIX

In the special case of goodness of fit of $M_1$, the asymptotic distribution of $X^2$ has the same form as (2.28) with the $\delta_i$ being the eigenvalues of $A = (H'\nabla H)$, where $H = (I - G(G'\nabla G)^{-1}G'\nabla)G_{22}$ and $G_{22}$ is any JRxJR-r matrix of rank JR-r such that $(G\ G_{22})$ is JRxJR of rank JR. Let $E = \nabla H$. Then $A = (E'\nabla^{-1}E)^{-1}(E'\Sigma E)$, rk E = rk H = JR-r and $E'G = H'\nabla G = 0$. If $\tilde{E}$ is any other JRxJR-r matrix of rank JR-r satisfying $\tilde{E}'G = 0$, there exists a nonsingular matrix B such that $\tilde{E} = EB$. Then,

$$(\tilde{E}'\nabla^{-1}\tilde{E})^{-1}(\tilde{E}'\Sigma\tilde{E}) = B^{-1}(E'\nabla^{-1}E)^{-1}(B')^{-1}B'(E'\nabla E)B$$

which has the same eigenvalues as A. Thus, the $\delta_i$ are the eigenvalues of $(E'V^{-1}E)^{-1}(E'\Sigma E)$ where E is any JRxJR-r matrix of full rank with $E'G = \underline{0}$.

## BIBLIOGRAPHY


Agresti, A. (1984). **Analysis of Ordinal Categorical Data.** John Wiley and Sons, New York.

Birch, M.W. (1964). A new proof of the Pearson-Fisher theorem. **Ann. Math. Statist.,** 35, 817-824.

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

Cook, R.D. & Weisberg, S. (1982). **Residuals and Influence in Regression.** Chapman & Hall, London.

Cox, D.R. (1970). **Analysis of Binary Data.** Chapman and Hall, London.

Fienberg, S.E. (1977). **The Analysis of Cross-Classified Data.** MIT Press, Cambridge, Mass.

McCullagh, P. and Nelder, J.A. (1983). **Generalized Linear Models.** Chapman and Hall, London.

Rao, C.R. (1965). **Linear Statistical Inference and its Applications.** Second Edition. John Wiley and Sons, New York.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. J. Amer. Statist. Assoc., 76, 221-30.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. **Ann. Statist.,** 12, 46-60.

Roberts, G.R. (1985). Contributions to Chi-squared Tests with Survey Data. Unpublished Ph.D. Thesis, Carleton University, Ottawa, Canada.

Roberts, G.R., Rao, J.N.K. and Kumar, S. (1986). Logistic regression analysis of sample survey data. To appear in **Biometrika.**