CATEGORICAL DATA ANALYSIS FOR COMPLEX SURVEYS

A. C. Singh and S. Kumar
Memorial University of Newfoundland and Statistics Canada

ABSTRACT

A class of generalized score tests for the analysis of multidimensional contingency tables from complex survey data is presented. The tests are asymptotically optimal and robust. Furthermore, these tests do not have the problems of instability that are sometimes encountered by the use of Wald type statistics in analysing these data. These problems are eliminated by working with the principal components of the observed count vector. The theory is used to analyze data from the October 1980 Canadian Labour Force Survey.

1. INTRODUCTION

It is well known that the standard X^2 methods for the analysis of categorical data arising from complex surveys could give seriously misleading conclusions. This is so because the type I error rate might be greatly inflated. We consider mainly three methods proposed in the literature for analysing survey data, namely, (i) the Weighted Least Squares (WLS), (ii) the Adjusted X^2, and (iii) the Jackknifed X^2. The WLS method discribed in Grizzle, Starmer, and Koch (1969) for simple random samples, was extended by Koch, Freeman, and Freeman (1975) to the case of complex samples. Fellegi (1978, 1980), Rao and Scott (1979, 1981, 1984), Holt, Scott, and Ewings (1980) and others have considered adjusting the standard X^2 by a scale factor in order to take account of the impact of the survey design. The third method due to Fay (1979, 1985) also adjusts the standard X^2 but by means of jackknifing the X^2 statistic.

The two methods, namely, the Adjusted X^2 and Jackknifed X^2 were developed with the objective of controlling Type I error. By contrast the WLS method provides an asymptotically optimal method based on Wald type statistics. However, as noted by Fay (1979, 1985) and others, the finite sample behaviour of WLS test statistic would generally be unstable with regard to the true asymptotic χ^2 distribution. This is due to inefficient estimation of covariance matrix of sample estimates for various cross-classifications (or domains) under complex designs such as multistage stratified cluster sampling. As a result the estimated covariance matrix is often nearly singular with the consequence of a serious effect in its inversion required in the computation of WLS statistic. This in turn would greatly inflate Type I error rate giving rise to misleading conclusions.

Recent results of Thomas and Rao (1984, 1985) indicate that within the scope of their simulation study for testing goodness-of-fit under cluster sampling, the two tests namely Jackknifed X^2 and Rao-Scott second order corrected X^2 perform better than WLS Wald type test despite its optimality. Thus the problem of instability mentioned above seems to overshadow the optimality of the Wald type statistic. It may be noted that for simple random samples, the instability problem under WLS approach does not usually arise because of the availability of exact formulas for the covariance matrix under consideration. The details for the case of simple random samples are given in Singh (1986).

In this paper an alternative method proposed earlier in a technical report (Singh, 1985) is described which provides an asymptotically optimal test statistic (to be denoted by $Q^{(T)}$) with true χ^2 limiting distribution. The test $Q^{(T)}$ requires a consistent estimate of the covariance matrix but is not susceptible to the problem caused by its near singularity. In section 2, we state the problem with some preliminary considerations. In section 3, the test $Q^{(T)}$ is described and its asymptotic behaviour is given in section 4. An application of $Q^{(T)}$ to a logistic regression analysis of the October 1980 Canadian Labour Force Survey (LFS) data is given in section 5. Finally, some discussion and remarks are presented in section 6.

2. THE PROBLEM AND PRELIMINARIES

Suppose that the population of interest is composed of I disjoint domains (subpopulations or cells), and ν_i denotes the parameter of interest pertaining to the ith domain, $i = 1, 2, \ldots, I$. For example, the ν_i's may be population proportions (or counts). The theory presented is also applicable to the more general case of domain means (or totals). Consider a model for ν_i's as

$$H_0: \nu_i = \nu_i(\theta) \text{ where } h(\nu_i) = x_i'\theta, \qquad (2.1)$$

for $i = 1, \ldots, I$. In (2.1), x's are known forming an I x r matrix of full rank r, θ is a r-vector of unknown paramters, h is a continuously differentiable one-to-one function so that h^{-1} exists. The function h includes, among others, the commonly used functions such as log and logit.

We will use $\hat{\nu}_i$'s, the survey estimates of ν_i's to make inference about H_0. The reason for using $\hat{\nu}_i$'s is not due to sufficiency arguments because likelihood function is difficult to obtain for a general sample design. We choose to use $\hat{\nu}_i$'s because they are generally available for large scale surveys conducted by various organizations. We shall assume that under an appropriate central limit theorem

$$\hat{\nu} \stackrel{a}{\sim} MVN(\nu, \Gamma/n) \qquad (2.2)$$

where $\hat{\nu}$ is the I-vector of $\hat{\nu}_i$'s and the symbol "$\stackrel{a}{\sim}$" stands for "asymptotically distributed as" and n denotes the total sample size.

It is known that the problem of testing fit of the model H_0 can be reduced asymptotically for local alternatives to that of testing a linear hypothesis for Gaussian case with an appropriate covariance matrix. Let $C(\phi)$ denote the class of tests for H_0 based on statistics ϕ. Then it can be shown following Lehmann (1959, pp. 304-313) that for nonsingular Γ, an asymptotically optimal (uniformly most powerful invariant) test in the class $C(\hat{\nu})$ has the rejection region

$$Q(\theta^o) \geq c \; . \tag{2.3}$$

The test statistic $Q(\theta^o)$ is given by

$$Q(\theta^o) = Y(\theta^o)' \, \Delta \, Y(\theta^o) - Z(\theta^o)' \, \Lambda \, Z(\theta^o), \tag{2.4}$$

where $Y(\theta^o) = (\hat{\nu} - \nu(\theta^o))$, $B = (\partial\nu/\partial\theta)$,

$\Delta = n\Gamma^{-1}$, $Z(\theta^o) = B' \, \Delta \, Y(\theta^o)$, $\Lambda = (B' \, \Delta \, B)^{-1}$,

and θ^o is some fixed point in the r-dimensional parameter space specified by H_0. Note that B is I x r matrix. The asymptotic null distribution of the test statistic (2.4) is χ^2_{I-r}. For computing $Q(\theta^o)$, θ^o can be replaced by any root n-consistent estimator $\hat{\theta}$ of θ under H_0, and Γ by a consistent estimator $\hat{\Gamma}$. Some remarks about the possible choices of $\hat{\theta}$ will be made later in this section.

If Γ is singular with rank s(<I), then an asumptotically optimal test based on $\hat{\nu}$ can be obtained by using a g-inverse of Γ. This consideration is useful for motivating the test $Q^{(T)}$ described in the next section. Consider the g-inverse of Γ obtained from its spectral decomposition. In other words, let P_i be the normailized eigen vector corresponding to the ith largest positive eigen value λ_i, i = 1, 2, ..., s. Let Δ_s denote a g-inverse of Γ/n, defined by

$$\Delta_s = n \sum_{i=1}^{s} (P_i P_i'/\lambda_i) \; . \tag{2.5}$$

Now the test statistic Q for the case of singular Γ can be obtained from (2.4) by using Δ_s instead of $n\Gamma^{-1}$ throughout the expression and the asymptotic null distribution of Q would then be χ^2_{s-r}.

It is instructive to note that $Q(\theta^o)$ can also be obained as a score statistic (see Cox and Hinkley, pp. 321-324) by employing the approximate likelihood of ν obtained by (2.2). $Q(\theta^o)$ can be termed as a Generalized Score Statistic (GSS) as in Singh (1986). It follows from the above observation that in the family $C(\hat{\nu})$, the test $Q(\theta^o)$ is also optimal in the sense of Wald (1943). Specifically, it is asymptotically admissible, most stringent and has best average power over certain surfaces, and unique up to asymptotic equivalence.

We make a few remarks about the choice of $\hat{\theta}$. We can use mle of θ, say $\bar{\theta}$, obtained from the likelihood function under the assumption that the data were generated by a simple random sampling plan. Such estimators are commonly known as pseudo mle and have been used quite often for complex designs (Imery, Koch and Stokes, 1982).

Another useful choice of $\hat{\theta}$ is that of the mle corresponding to the likelihood of ν obtained from the asymptotic distribution. We denote this estimator by $\tilde{\theta}$. This approach is equivalent to the minimum X^2-type estimation. It also turns out that, for $\theta^o = \tilde{\theta}$, the test statistic $Q(\theta^o)$ simplifies because the second term in the expression (2.4) becomes zero. Furthermore, if the model H_0 is found to be adequate, then an asymptotically efficient estimator of ν can be obtained by using $\tilde{\theta}$ in (2.1). Such an optimal property is not known for estimator of ν based on $\bar{\theta}$. For details see Kumar and Singh (1986).

Even if Γ were nonsingular, its estimate $\hat{\Gamma}$ may be nearly singular in the sense that some of its eigen values are near zero. Such cases are likely to arise when $\hat{\Gamma}$ is calculated for complex survey designs. In these cases the test statistic Q of (2.3), although asymptotically optimal, would be unstable with regard to the type I error rate. The situation is similar to that for WLS test mentioned in section 1. To circumvent this instability problem with Q, it seems reasonable to modify Q by dropping a few components corresponding to small eigen values as described in the next section. Notice that the statistic $Q^{(s)}$ (for Γ singular case) can be viewed as a special case of modification when components of Q corresponding to zero eigen values are dropped.

3. THE TEST $Q^{(T)}$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_I$ be the eigen values of Γ as before. Some of the λ_i's may be zero depending on s (rank of Γ). In the following we assume that whenever Γ is unknown, it will be replaced by a consistent estimate. We prescribe a very small nonnegative number ε, e.g., .005 or .01 as working values of ε. Now define

$$T = \max\{t\colon t > r \text{ and } \lambda_{[t]}/\lambda_{[1]} \geq \varepsilon\}, \tag{3.1}$$

where r is the number of model parameters under H_0 and

$$\lambda_{[t]} = \sum_{i=t}^{I} \lambda_i, \; t = 1, \ldots, I. \tag{3.2}$$

Note that T depends on ε and is equal to its maximum value s when $\varepsilon = 0$. Let

$$\Delta_T = n \sum_{i=1}^{T} (P_i P_i'/\lambda_i) \; , \tag{3.3}$$

where P_i is the eigen vector corresponding to λ_i. The matrix Δ_T may not be uniquely defined for cases where some of the λ_i's are equal.

3.1. A TEST OF GOODNESS-OF-FIT

We define the test statistic $Q^{(T)}$ for testing the goodness-of-fit of the model H_0 as follows:

$$Q^{(T)}(\theta^o) = Y(\theta^o)'\Delta_T Y(\theta^o) - Z_T(\theta^o)'\Lambda_T Z_T(\theta^o) \tag{3.4}$$

where Z_T and Λ_T are defined as in (2.4) by replacing Δ by Δ_T, and θ^o is a fixed point in the parameter space under H_0. We say that the model is inappropriate for the data when $Q^{(T)}(\theta^o)$ is too large. The asymptotic distribution of the test statistic under H_0 is χ^2_{T-r}. It should be remarked that T, though random when $\hat{\Gamma}$ is empolyed, can be regarded as fixed for our asymptotics (see Proposition 4.1).

As before, θ^o can be replaced by any root n-consistent estimate of θ. For example, the pseudo mle $\bar{\theta}$ may be used. Another choice $\tilde{\theta}$ can be obtained by minimizing the expression

$$(\hat{\nu} - \nu(\theta))' \Delta_T(\hat{\nu} - \nu(\theta)) \qquad (3.5)$$

It may be noted that: (i) $\tilde{\theta}$ depends on Δ_T, i.e., different values of T will give different $\tilde{\theta}$, and (ii) the minimum value of (3.5) is indeed $Q^{(T)}(\tilde{\theta})$ because the second term on the right side in (3.4) becomes zero.

3.2 NESTED MODELS

Let $X = (x_1, x_2, \ldots, x_I)'$ be partitioned as (X_1, X_2) where X_1 is I x p and X_2 is I x q $(p + q = r)$. Then the model (2.1) may be written as

$$h(\nu) = X\theta = X_1\theta_1 + X_2\theta_2 , \qquad (3.6)$$

where $h(\nu) = (h(\nu_1), h(\nu_2), \ldots, h(\nu_I))'$, θ_1 is a p-vector and θ_2 is a q-vector. We are interested in testing the hypothesis H_1: $\theta_2 = 0$ given the model H_0 is accepted. Let $\theta^* = (\hat{\theta}_1, 0)$, where $\hat{\theta}_1$ is a $\sqrt{n}$-consistent estimate of θ_1 under H_1 and 0 is a q-vector of zeros. Then the test statistic for the nested model $H_{1.0}$ (H_1 given H_0) is given by

$$Q^{(T)}_{1.0}(\hat{\theta}_1) = Q^{(T)}_1(\hat{\theta}_1) - Q^{(T)}_0(\theta^*), \qquad (3.7)$$

where $Q^{(T)}_1(\hat{\theta}_1)$ is the goodness-of-fit test statistic for H_1 (non-nested) with p parameters. Thus $Q^{(T)}_1$ is analogous to (3.4). However, $Q^{(T)}_0(\theta^*)$ is simply the computed value of the expression (3.4) when $\theta^o = \theta^*$. It is not a goodness-of-fit statistic for H_0 because θ^* is not a consistent estimator of θ under H_0. The statistic (3.7) is distributed as χ^2_q under H_1 given H_0 and we reject $H_{1.0}$ for large values of $Q^{(T)}_{1.0}$.

An alternate asymptotically equivalent test for the nested problem can be based on the difference between goodness-of-fit test statistics for H_1 and H_0 respectively, i.e. $Q^{(T)}_1(\hat{\theta}_1) - Q^{(T)}_0(\hat{\theta})$, where $\hat{\theta}$ denotes a root n-consistent estimate of θ under H_0. Here two sets of estimates of θ will be required and sometimes, the difference may turn out to be negative for finite samples. The situation is analogous to that encountered with Pearson-Fisher's X^2_{PF} for simple random samples.

3.3 AN INTERPRETATION OF $Q^{(T)}$

A practical interpretation of $Q^{(T)}$ statistic can be given as follows. First define

$$M_T = (P_1, \ldots, P_T),\ W = M_T'\hat{\nu},\ C = M_T'B$$
$$\mu(\theta) = M_T'\nu(\theta), \text{ and } D_T = \text{diag}\ (\lambda_1, \ldots, \lambda_T) \qquad (3.8)$$

It is interesting to note that $Q^{(T)}$ of (3.4) can be expressed as

$$nU(\theta^o)'D_T^{-1}U(\theta^o) \quad -n(V(\theta^o)'(C'D_T^{-1}C)^{-1} V(\theta^o) \qquad (3.9)$$

where $U(\theta^o) = W - \mu(\theta^o)$, $V(\theta^o) = C'D_T^{-1}U(\theta^o)$.

It is seen from the above representation that $Q^{(T)}$ is a generalized score statistic (GSS). It is based on the first T principal components W (of $\hat{\nu}$) whose approximate likelihood is given by

$$W \stackrel{a}{\sim} \text{MVN}\ (\mu, D_T/n), \qquad (3.10)$$

In the class C(W) of tests based on W, $Q^{(T)}$ is asymptotically optimal for our problem. It is due to the fact that the original problem about I-dimensional ν is reduced to testing a hypothesis concerning T-dimensional parameter μ specified by

$$H_0' : \mu = M_T' h^{-1}(X\theta) \qquad (3.11)$$

where $X = (x_1, x_2, \ldots, x_I)'$.

The test $Q^{(T)}$ is expected to control the inflation in type I error rate for finite samples when Q (based on Γ^- without any modification) may be unstable. In fact $Q^{(T)}$ is a conservative test for H_0 because H_0 is a subset of H_0'. This provides additional insurance for controlling type I error rate. In addition to the above property, the test $Q^{(T)}$ for small ε will be nearly optimal for H_0 in the class $C(\hat{\nu})$ when compared to the test Q. This is so because $\hat{\nu}$ and W will be close in the sense that principal components provide the best way for dimensionality reduction with minimum loss of information (see Rao 1973, p. 592). Thus in the absence of instability in Q, the test $Q^{(T)}$ will be robust for fairly small ε. A test for checking instability is given in the next subsection.

3.4 WHEN TO USE $Q^{(T)}$

The test $Q^{(T)}$ is based on the premise that we are willing to sacrifice some information in the data whose contribution to Q is believed to be unstable. Naturally it leads to the following question. How to perform an instability check for Q for a given data? The following test can be used for the above problem.

For a given $\varepsilon > 0$, level α, if

$$Q^{(s)} - Q^{(T)} > \chi^2_{\alpha,s-T} , \qquad (3.12)$$

then we say that there is an evidence of instability in Q and $Q^{(T)}$ should be used.

4. ASYMPTOTIC DISTRIBUTION OF $Q^{(T)}$

First we show that the random variable T defined by (3.1) for a slightly modified ε converges in probability to T_0 where T_0 is T corresponding to the true Γ. This implies that T can be regarded as fixed for our asymptotics.

Proposition 4.1. Assume that

$$\sqrt{n}\,\|\hat{\Gamma} - \Gamma\| = O_p(1)\ , \qquad (4.1)$$

where $\|\cdot\|$ is the Euclidean norm for a matrix and n is the total sample size. Also let ε be modified to ε^* defined by

$$\varepsilon^* = \varepsilon - \beta_n,\ 0 < \beta_n < \varepsilon,\ \beta_n \downarrow 0 \text{ but } \sqrt{n}\beta_n \to \infty \qquad (4.2)$$

Then, $T(\varepsilon^*) \to T_0(\varepsilon)$ (in prob) as $n \to \infty$. (4.3)

We shall outline the proof. It follows from the condition (4.1) that $\sqrt{n}\,(\hat{\lambda}_i - \lambda_i^o)$, $i = 1,\ldots,s$, are bounded in probability where λ_i^o is the ith largest eigen value of the true Γ. This implies that for u = 1, 2, ..., s.

$$\sqrt{n}\ |\ G_n(u) - G(u)| = O_p(1) \qquad (4.4)$$

where $G_n(u) = \lambda_{[s-u+1]}/\lambda_{[1]}$

and G(u) is the corresponding function for Γ. Clearly both G_n and G are increasing with values between 0 and 1. Now, for a given $\varepsilon \geq 0$, let us define a random variable.

$$U_n = G_n^{-1}(\varepsilon^*) = \inf\{u: G_n(u) \geq \varepsilon^*\} \qquad (4.5)$$

By analogy with empirical distribution functions for discrete distributions, it follows from Külperger and Singh (1982) that

$$U_n = G_n^{-1}(\varepsilon^*) \to U_o = G^{-1}(\varepsilon) \text{ in probability} \qquad (4.6)$$

which implies that

$$T(\varepsilon^*) = s - U_n + 1 \to T_o(\varepsilon) = s - U_o + 1 \text{ in prob.}$$

A possible choice of β_n is $(\log n/n)^{\frac{1}{2}}\,\varepsilon$ as suggested in Kulperger and Singh (1982). The modification term β_n will be almost zero for very large n. It may also be noted that the condition (4.1) would generally be satisfied for commonly used estimation procedures of Γ by appealing to an appropriate CLT.

Henceforth we shall regard T as fixed asymptotically. The next proposition states the asymptotic distribution of $Q^{(T)}$ for testing goodness of fit of H_0 .

Proposition 4.2. Under H_0, $Q^{(T)} \mathrel{\dot\sim} \chi^2_{T-r}$ (4.7)

The proof follows from the observation (3.9) that the $Q^{(T)}$ statistic coincides with the Q test statistic (defined by 2.4) applied to the modified hypothesis H_0' based on the T-dimensional transformed data vector W. The next proposition gives the asymptotic test for goodness of fit.

Proposition 4.3. Under H_0

$$Q^{(s)} - Q^{(T)} \mathrel{\dot\sim} \chi^2_{s-T} \qquad (4.8)$$

To prove the above proposition, note that $Q^{(s)} - Q^{(T)}$ equals

$$Y'[(\Delta_s - \Delta_s\,B(B'\Delta_s B)^{-1}B'\Delta_s) - (\Delta_T - \Delta_T B(B'\Delta_T B)^{-1}B'\Delta_T)]Y$$

$$= Y'\,AY \text{ (say)} \qquad (4.9)$$

Now it is seen that $A\Gamma/n$ is idempotent using the relation $\Delta_s\,\Gamma\,\Delta_T = n\Delta_T$. The rank of A can be shown to be s - T from Khatri (1968). See also Rao & Mitra (1971, p. 111). Hence, the proposition. The final proposition states the asymptotic distribution of the test statistic for a nested hypothesis.

Proposition 4.4 Under the nested model $H_{1.0}$,

$$Q^{(T)}_{1.0} \mathrel{\dot\sim} \chi^2_q \qquad (4.10)$$

The proof is similar to that for the proposition 4.3. The quadratic form $Q^{(T)}_{1.0}$ can be expressed as

$$Y(\theta_1^o)'[\Delta_T B_0(B_0'\Delta_T B_0)^{-1}B_0'\Delta_T - \Delta_T B_1(B_1'\Delta_T B_1)^{-1}B_1'\Delta_T]Y(\theta_1^o)$$

$$= Y'\,AY \text{ (say)} \qquad (4.11)$$

where B_0 and B_1 denote the matrices B computed under H_0 and H_1 respectively. Now in new of the relation which follows from the fact that B_1 is a submatrix of B_0, namely

$$B_0(B_0'\Delta_T B_0)^{-1}B_0'\Delta_T B_1(B_1'\Delta_T B_1)^{-1}B_1 = B_1(B_1'\Delta_T B_1)^{-1}B_1, \qquad (4.13)$$

we can show that $A\Gamma/n$ is idempotent and of rank q from Khatri (1968). Hence, the proposition.

5. AN APPLICATION TO THE LABOUR FORCE SURVEY DATA

We shall apply the $Q^{(T)}$ method to the October 1980 labour force survey (LFS) data analysed earlier by Kumar and Rao (1984) and Roberts, Rao and Kumar (1986) using adjusted X^2 and then compare the results. The data consist of males aged 15-64 who were in the labour force and excluded full-time students. Two factors, age and education, were used to explain the variation in employment using a logit model. Ten age groups [10 + 5j, 14 + 5j], j = 1, 2, ..., 10 were formed and the midpoint 12 + 5j (=A_j, say) was assigned to represent age for all persons in the jth group. Similarly each person was assigned one of the six values for education, E_ℓ (ℓ = 1, ..., 6) representing 7, 10, 12, 13, 14 and 16 as median years of schooling. They considered fitting the following logistic regression model to the survey estimates of employment rates ($\nu_{j\ell}$) for the table of 60 cells cross-classified by age and education. The model is

$$\ell n\,\frac{\nu_{j\ell}}{1 - \nu_{j\ell}} = \beta_o + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_\ell + \beta_4 E_\ell^2\ . \qquad (5.1)$$

It should be mentioned that LFS design employed a stratified multistage cluster sampling with two or more stages of sampling. The survey estimates $\hat{\nu}_{j\ell}$ were adjusted for post-stratification using the projected age-sex distribution at the provincial level.

The model (5.1) can be expressed in the notation of section 2 by numbering the sixty cells lexicographically. Thus, (5.1) can be rewritten as

$$h(\nu) = X\theta, \qquad (5.2)$$

where ν is the vector of employment rates, h is the logit function, X is a 60 x 5 matrix with ith row

$(1, A_i, A_i^2, E_i, E_i^2)$, and θ is $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$. We also have

$$B = H^{-1}X, \text{ with } H = (\partial h/\partial \nu) = D_\nu^{-1}D_{1-\nu}^{-1} \quad (5.3)$$

where D_ν and $D_{1-\nu}$ are diagonal matrices with diagonal elements given by the subscripts.

The estimated covariance matrix $\hat{\Gamma}$ as obtained by Kumar and Rao (1984) was singular with rank 58 because two cells had zero observed unemployment rates. For $\varepsilon = .01$, T turns out to be 51. Using the pseudo-mle of θ (under product binomial sampling), we get $Q^{(T)}$ and $Q^{(s)}$ as

$$Q^{(51)}(\bar{\theta}) = 53.5,\ Q^{(58)}(\bar{\theta}) = 86.64 \text{ where} \quad (5.4)$$

$$\bar{\theta} = (-2.76, 0.209, -0.00217, 0.0913, .00276)'.$$

Since the difference, $Q^{(58)} - Q^{(51)}$ (= 33.14) is highly significant when referred to a χ_7^2 distribution, it indicates possible instability in $Q^{(s)}$ and therefore, the test $Q^{(T)}$ is recommended. The test $Q^{(51)}$ accepts the model H_0 at 5% because $\chi^2_{.05,51-5} = 62.83$. The test $Q^{(58)}$ on the other hand rejects H_0 as expected. If we use a higher ε, then we would of course expect to favour H_0. For instance, with $\varepsilon = .05$, T = 37 and

$$Q^{(37)}(\bar{\theta}) = 29.99 < \chi^2_{.05,37-5} = 46.19. \quad (5.5)$$

The conclusion based on $Q^{(T)}$ agrees with that of Rao-Scott's adjusted X^2 used in the earlier analysis. Note that the first order corrected G_c^2 was found to be 53.7 which favours H_0 when referred to a χ_{55}^2 distribution. It may be pointed out that for G_c^2, χ^2 distribution is used only as an approximation to the true limiting null distribution which is known to correspond to a linear combination of χ^2 variables.

Table 1: $Q^{(T)}$ and Adj X^2 Using Pseudo Mle

Nested Hypothesis given H_0	$Q^{(T)}$ $\varepsilon = .05$ T = 37	$\varepsilon = .01$ T = 51	$\varepsilon = 0$ T = 58	Adj X^2 (Rao-Scott) G_c^2
$\beta_1 = 0$	102.3*	178.12*	261.85*	168.4*
$\beta_2 = 0$	65.05*	112.94*	156.51*	102.1*
$\beta_3 = 0$	0.17	0.65	2.94	1.01
P-value	>.65	>.40	>.08	>.30
$\beta_4 = 0$	1.73	0.425	.01	0.46
P-value	>.15	>.50	>.90	>.45
$\beta_3=\beta_4=0$	71.34*	124.92*	227.93*	172.1*
$\beta_2=\beta_4=0$	85.34*	118.17*	156.38*	106.3*

*P-value < .0005

Several nested hypotheses were also tested using $Q^{(T)}$. These are summarized in Table 1 along with the values of G_c^2 obtained from the earlier analysis of the same data. For the nested hypothesis $H_{1.0}: \beta_4 = 0$ given H_0, we use the test statistic $Q_{1.0}^{(T)}$ of (3.7) and we obtain

$$Q_{1.0}^{(51)}(\bar{\theta}^{(1)}) = 0.425 \quad (5.6)$$

where $\bar{\theta}^{(1)} = (-3.10, 0.211, 0.00218, 0.1509)'$. The estimate $\bar{\theta}^{(1)}$ is the pseudo-mle under the model H_1: $\beta_4 = 0$. It is seen that the hypothesis $H_{1.0}$ is accepted by referring $Q_{1.0}^{(51)}$ to a χ_1^2 distribution. As seen from Table 1, the results based on $Q^{(T)}$ for ε even as high as .05 agree with those for $\varepsilon = .01$. Moreover, they are all in agreement with the conclusions based on G_c^2. The test for nested hypothesis suggest that β_4 can be dropped. So we also computed the test $Q^{(T)}$ for H_1. It is found that

$$Q_1^{(51)}(\bar{\theta}^{(1)}) = 54.211,\ Q_1^{(37)}(\bar{\theta}^{(1)}) = 32.15, \quad (5.7)$$

which when referred respectively to χ^2 with 47 and 33 d.f show that H_1 is accepted by the data.

6. DISCUSSION

It is seen that the well-known multivariate technique of principal components can be used to provide an important statistical method for model testing for survey data. The essential idea in the construction of $Q^{(T)}$ is to sacrifice some information in the data that give rise to possibly unreliable components in the metric defined by $nY'\Gamma^-Y$. The modified metric corresponding to $Q^{(T)}$ is defined by $Y'\Delta_T Y$ where Δ_T can be interpreted as a truncated g-inverse of Γ/n. The test $Q^{(T)}$ provides an alternative to WLS, Adjusted X^2 and Jackknifed X^2 approaches. It would be desirable for future investigations to perform a simulation study for level and power comparisons similar to Thomas and Rao (1984, 1985).

It may be noted that for Γ nonsingular, the test Q of (2.4) or $Q^{(T)}$ (with T = I) is asymptotically equivalent to WLS test. Thus WLS test can also be modified in a similar fashion if so desired. However, since the WLS test statistic uses transformed data $h(\hat{\nu})$, it might also encounter another kind of instability namely, due to extreme behaviour of the function h for certain values of $\hat{\nu}$. Clearly, this will not be so with Q or $Q^{(T)}$. Finally we remark that the statistic $Q^{(T)}$ provides a general recipe whether $\hat{\nu}$ represent domain proportions or means or any other statistics. However, the methods based on correcting X^2 do not seem to be applicable to the more general case of domain means.

ACKNOWLEDGEMENT

The first author's research was supported by Statistics Canada and Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Cox, D. R., and Hinkley, D. V. (1974). Theoretical Statistics, Chapman and Hall Ltd., London.

Fay, R. E. (1979). On adjusting the Pearson Chi-square statistic for clustered sampling. Proc. Amer. Statist. Assoc., Social Statistics Section, 402-405.

Fay, R. E. (1985). A jackknifed chi-square test for complex samples. J. Amer. Statist. Assoc., 80, 148-157.

Fellegi, I. P. (1978). Approximate tests of independence and goodness-of-fit based on stratified multistage samples, Survey Methodology 4, 29-56.

Fellegi, I. P. (1980). Approximate tests of independence and goodness-of-fit based on stratified multistage samples, J. Amer. Statist. Assoc. 71, 665-670.

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

Holt, D., Scott, A. J. and Ewings, P. O. (1980). Chi-squared tests with survey data. J. Roy. Statist. Soc. Ser. A. 143, 302-330.

Imrey, P. B., Koch, G. G., and Stokes, M. E. (1982). Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part II: Data analysis. Int. Statist. Rev.50 35-63.

Khatri, C. G. (1968). Some results for the singular multivariate regression models. Sankhya A. 30, 267-280.

Koch, G. G., Freeman, D. H. Jr., and Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. Int. Statist. Rev. 43, 59-78.

Kulperger, R. J., and Singh, A. C. (1982). On random grouping in goodness-of-fit tests for discrete distributions. Journal of Statistical Planning and Inference, 7, 109-115.

Kumar, S., and Rao, J. N. K. (1984). Logistic regression analysis of Labour Force Survey Data. Survey Methodology 10, 62-81.

Kumar, S., and Singh, A. C. (1986). On efficient estimation of unemployment rates from labour force survey data. To appear tentatively in Survey Methodology.

Lehmann, E. L. (1959). Testing Statistical Hypotheses, NY, John Wiley.

Rao, C. R. (1973). Linear Statistical inference and its applications. 2nd ed. NY. John Wiley.

Rao, C. R., and Mitra, S. K. (1971). Generalized inverse of matrices and its applications, NY John Wiley.

Rao, J. N. K., and Scott, A. J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. Proc. Amer. Statist. Assoc., Survey Research Methods Section, 58-66.

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, 221-230.

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, No. 1, 46-60.

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

Singh, A. C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Statistics Canada Methodology Branch, Working Paper No. SSMD 86-002.

Singh, A. C. (1986). On the optimality and a generalization of Rao-Robson's statistic. To appear in Commun. Statist - Theo. Meth.

Singh, A. C. (1986). Categorical data analysis for simple random samples. Proc. Amer. Statist. Assoc., Survey Research Methods Section.

Thomas, D. R., and Rao, J. N. K. (1984). A Monte Carlo study of exact levels of goodness-of-fit statistics under cluster sampling. Proc. Amer. Statist. Assoc., Survey Research Methods Section, 207-211.

Thomas, D. R. and Rao, J. N. K. (1985). On the power of some goodness-of-fit tests under cluster sampling. Proc. Amer. Statist. Assoc. Survey Research Methods Section, 291-296.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426-482.