Survey Methodology, December 1989
Vol. 15, No. 2, pp. 203–223
Statistics Canada

Logistic Regression Under Complex Survey Designs

JORGE G. MOREL[1]

ABSTRACT

Estimation procedures for obtaining consistent estimators of the parameters of a generalized logistic function and of its asymptotic covariance matrix under complex survey designs are presented. A correction in the Taylor estimator of the covariance matrix is made to produce a positive definite covariance matrix. The correction also reduces the small sample bias. The estimation procedure is first presented for cluster sampling and then extended to more complex situations. A Monte Carlo study is conducted to examine the small sample properties of F-tests constructed from alternative covariance matrices. The maximum likelihood estimation method where the survey design is completely ignored is compared with the usual Taylor's series expansion method and with the modified Taylor procedure.

KEY WORDS: Pseudo-likelihood; CPLX procedure; Cluster sampling; Adjusted covariance matrix.

1. INTRODUCTION

In the last few years a lot of attention has been given to the problems that arise when chi-square tests based on the multinomial distribution are applied to data obtained from complex sample designs. It has been shown that the effects of stratification and clustering on the chi-square tests may lead to a distortion of nominal significance levels. Holt, Scott and Ewings (1980) proposed modified Pearson chi-square statistics tests of goodness-of-fit, homogeneity, and independence in two-way contingency tables. Rao and Scott (1981) presented similar tests for complex sample surveys. In all these cases, the correction factor requires only the knowledge of variance estimates (or design effects) for individual cells. Bedrick (1983) derived a correction factor for testing the fit of hierarchical log linear models with closed form parameter estimates. Rao and Scott (1984) presented more extensive methods of using design effects to obtain chi-square tests for complex surveys. They generalized their previous results to multiway tables. Fay (1985) presented the adjustments to the Pearson and likelihood test statistics through a jackknife approach.

The use of the conditional logistic model, Cox (1970), has become increasingly popular in the context of complex survey designs. Under suitable conditions, Binder (1983), proved the asymptotic normality of design-based sampling distribution for a family of parameter estimators that cannot be defined explicitly as a function of other statistics from the sample. His results are applied to binary logistic models. Further applications to the Canada Health Survey are also found in Binder *et al.* (1984).

Chambless and Boyle (1985) derived a general asymptotic distribution theory for stratified random samples with a fixed number of strata and increasing stratum sample sizes. Their theoretical results were illustrated with logistic regression and discrete proportional hazard-smodels. Albert and Lesaffre (1986) discussed the logistic discrimination method for classifying multivariate observations into one of several populations. They restrict their attention to discrimination between qualitatively distinct groups.

[1] Jorge G. Morel is Assistant Professor of the Department of Epidemiology and Biostatistics, University of South Florida, Tampa, Florida 33612.

Extensions to the case where the response consists of a polychotomous variable have been done by Bull and Pederson (1987) and Morel (1987). They show, by using Taylor's series expansion, that the large sample variance of the beta estimates has the form

$$H^{-1} G H^{-1}$$

where H^{-1} is the covariance matrix that wrongly results from assuming independence and multinomial distribution in the response vector, and G is a matrix whose estimation is based in the complex survey design.

More recently, Roberts, Rao and Kumar (1987) showed how to make adjustments that take into account the survey design in computing the standard chi-square and the likelihood ratio test statistics for logistic regression analysis involving a binary response variable. The adjustments are based on certain generalized design effects. Their results can be applied to cases where the whole population has been divided into I domains of study, a large sample is obtained for each domain, and in each domain a proportion π_i, $i = 1, 2, \ldots, I$, is to be estimated. It is assumed

$$\pi_i = \left[1 + \exp\left(x_i \underset{\sim}{\beta}^0\right)\right]^{-1} \exp\left(x_i \underset{\sim}{\beta}^0\right), \; i = 1, 2, \ldots, I,$$

where x_i is a k-vector of known constants derived from the i-th domain and $\underset{\sim}{\beta}^0$ is a k-vector of unknown parameters. This procedure may be most useful when only the summary table of counts and variance adjustment factors are available, instead of the complete data set.

In this paper an estimation procedure is presented for obtaining consistent estimators of the parameter vector of a generalized logistic model and its asymptotic covariance matrix when a complex sampling design is employed. The resulting estimated covariance matrix is always positive definite and asymptotically equivalent to the one obtained from Taylor's series expansion. A correction for reducing the small sample bias in the estimated covariance matrix is also introduced. It is shown, via a Monte Carlo study, that this correction levels off the inflated Type I error that arises from ignoring the complex survey, faster than the Taylor's series expansion. In this sense the correction proposed here produces, for small samples, results that are superior to the usual delta-method.

The new procedure will be termed, henceforth, the CPLX procedure, or simply CPLX. The maximum likelihood estimation method and the Taylor's series expansion method will be termed MLE and TAYLOR, respectively. The CPLX procedure has been incorporated into PC CARP, a personal computer program for variance estimation with large scale surveys, see Schnell *et al.* (1988).

2. LOGISTIC REGRESSION WITH CLUSTER SAMPLING

Consider first single-stage cluster sampling where n clusters or primary sampling units are taken with known probabilities with replacement from a finite population or without replacement from a very large population. Let m_j represent the size of the j-th cluster, $j = 1, 2, \ldots, n$, and let $y_{j\ell}^*$, $\ell = 1, 2, \ldots, m_j$ denote $(d + 1)$ dimensional classification vectors. The vector $y_{j\ell}^*$ consists entirely of zeros except for position r which will contain a one if the ℓ-th unit selected from the j-th cluster falls in the r-th category. Let $x_{j\ell}$ be a k-dimensional row vector of explanatory variables associated with the ℓ-th unit selected from the j-th cluster.

Then, for each $j = 1, 2, \ldots, n$, and each $\ell = 1, 2, \ldots, m_j$, the expectation of the r-th element of $y_{j\ell}^*$ is determined by a logistic relationship as

$$\pi_{j\ell r} = E\{y_{j\ell r}\} = \left[1 + \sum_{s=1}^{d} \exp(x_{j\ell}\beta_s^0)\right]^{-1} \exp(x_{j\ell}\beta_r^0) \quad r = 1, 2, \ldots, d$$

$$= 1 - \sum_{s=1}^{d} \pi_{j\ell s}, \qquad r = d + 1. \tag{2.1}$$

Because the expected value function is nonlinear in the parameter vector $\beta^0 = (\beta_1^{0\prime}, \beta_2^{0\prime}, \ldots, \beta_d^{0\prime})'$, it is necessary to use nonlinear estimation methods. Define the pseudo log-likelihood $L_n(\beta)$ as

$$L_n(\beta) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j(\log \pi_{j\ell}^*)' y_{j\ell}^*, \tag{2.2}$$

where $\pi_{j\ell}^* = (\pi_{j\ell 1}, \ldots, \pi_{j\ell, d+1})'$ and w_j is the sampling weight for the $j\ell$-th sampling unit. This function can be viewed as a weighted log likelihood function, where the weights are the sampling weights and the $y_{j\ell}^*$'s are distributed as multinomial random variables. If the sampling weights are all one, then (2.2) becomes the log-likelihood function under the assumption that the $y_{j\ell}^*$'s are independently multinomially distributed.

Let $\hat{\beta}_{\text{PSEUDO}}$ be the estimator of β^0 that maximizes (2.2). This estimator is a solution to the system of equations

$$\sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j G(\beta, x_{j\ell}) \left[\text{Diag}(\pi_{j\ell}^*)\right]^{-1} (y_{j\ell}^* - \pi_{j\ell}^*) = 0, \tag{2.3}$$

where

$$G(\beta, x_{j\ell}) = \left[(I_{d\times d}, 0_{d\times 1}) \otimes x_{j\ell}'\right] \Delta(\pi_{j\ell}^*),$$

$$\Delta(\pi_{j\ell}^*) = \text{Diag}(\pi_{j\ell}^*) - \pi_{j\ell}^*(\pi_{j\ell}^*)',$$

and $\otimes$ denotes the Kronecker product.

The asymptotic normality of $\hat{\beta}_{\text{PSEUDO}}$ can be proved by defining the parameters of interest implicity as in (2.2) and then by extending the results given in Binder (1983). An alternative approach can be derived by making use of the pseudo-likelihood assumption and Proposition 1 in Dale (1986). Binder and Dale both provide the necessary regularity conditions.

As n increases,

$$\sqrt{n}(\hat{\beta}_{\text{PSEUDO}} - \beta^0) = \sqrt{n}\left[H_n(\beta^0)\right]^{-1} U_n(\beta^0)$$

$$\xrightarrow{L} N_{dk}\left(0, \lim_{n\to\infty} \left[H_n(\beta^0)\right]^{-1} G_n\left[H_n(\beta^0)\right]^{-1}\right) \tag{2.4}$$

where,

$$H_n(\underset{\sim}{\beta}^0) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j \Delta(\underset{\sim}{\pi}_{j\ell}) \otimes \underset{\sim}{x}'_{j\ell} \underset{\sim}{x}_{j\ell},$$

$$U_n(\underset{\sim}{\beta}^0) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j (y_{j\ell} - \underset{\sim}{\pi}_{j\ell}) \otimes \underset{\sim}{x}'_{j\ell},$$

$$G_n = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j^2 \operatorname{Var}(y_{j\ell}) \otimes \underset{\sim}{x}'_{j\ell} \underset{\sim}{x}_{j\ell},$$

$y_{j\ell}$ and $\underset{\sim}{\pi}_{j\ell}$ are the vectors $y^*_{j\ell}$ and $\underset{\sim}{\pi}^*_{j\ell}$, without their last elements, respectively and N_{dk} denotes a dk-multivariate normal distribution.

Nelder and Wedderburn (1972) have shown that under binomial assumption, the pseudo log-likelihood function (2.2) can be solved by an iterative weighted least-squares procedure. Haberman (1974, p.48) shows that under regularity conditions a modified Newton-Raphson converges to the maximum likelihood estimator for the multinomial case. His proof does not depend on the existence of any consistent estimator of $\underset{\sim}{\beta}^0$ which allows the iterative algorithm to be initialized at $\hat{\underset{\sim}{\beta}} = \mathbf{0}$. Jennrich and Moore (1975) proved that when the multinomial assumption holds, the common Gauss-Newton algorithm for finding the maximum likelihood estimator of $\underset{\sim}{\beta}^0$ becomes the Newton-Raphson algorithm. Because of this equivalence of those algorithms and because a modified Newton-Raphson procedure always converge, we have adopted the modified Gauss-Newton algorithm described by Gallant (1987, p.318).

CPLX first finds $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ using an iterative procedure in which the estimate of $\underset{\sim}{\beta}^0$ at the q-th step is

$$\hat{\underset{\sim}{\beta}}_{[q,i(q)]} = \hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]} + (0.5)^{i(q)} \left[H_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,\,i(q-1)]}\right)\right]^{-1} U_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,\,i(q-1)]}\right) \tag{2.5}$$

where $i(q)$ is a nonnegative integer such that

$$L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right) > L_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]}\right). \tag{2.6}$$

The modification of the iteration algorithm provided by $i(q)$ guarantees the convergence of the procedure. The iteration is initiated by setting $\hat{\underset{\sim}{\beta}}_{(0)} = \mathbf{0}$. The algorithm is declared to have converged when the condition

$$\frac{L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right) - L_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]}\right)}{\left|L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right)\right| + 10^{-5}} < \epsilon \tag{2.7}$$

is satisfied, where ϵ can be 10^{-8}.

Observe that a consistent estimator of $H_n(\underset{\sim}{\beta}^0)$ is $H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})$ and a distribution free estimator of G_n is

$$\overset{*}{G}_n = (n-1)^{-1} n \sum_{j=1}^{n} (d_j - \bar{d})(d_j - \bar{d})', \tag{2.8}$$

where

$$d_j = \sum_{\ell=1}^{m_j} w_j(y_{j\ell} - \underset{\sim}{\pi}_{j\ell}) \otimes x'_{j\ell},$$

and $\bar{d} = n^{-1} \sum_{j=1}^{n} d_j$. If within each cluster, the $y^*_{j\ell}$'s are independent and identically distributed according to a multinomial random vector with parameters $(\underset{\sim}{\pi}^*_j, 1)$, then it can be easily shown that the expectation of $\overset{*}{G}_n$ is precisely $H_n(\underset{\sim}{\beta}^0)$. In practice the $\underset{\sim}{\pi}_{j\ell}$'s in (2.8) are replaced with $\hat{\underset{\sim}{\pi}}_{j\ell}$ where $\hat{\pi}_{j\ell r}$ is defined as in (2.1) with $\hat{\beta}_{\text{PSEUDO}}$ substituted by $\underset{\sim}{\beta}^0$, and a small correction is applied to obtain the estimator

$$\hat{G}_n = (n^* - k)^{-1} (n^* - 1)(n-1)^{-1} n \sum_{j=1}^{n} (\hat{d}_j - \hat{\bar{d}})(\hat{d}_j - \hat{\bar{d}})', \tag{2.9}$$

where

$$\hat{d}_j = \sum_{\ell=1}^{m_j} w_j(y_{j\ell} - \hat{\underset{\sim}{\pi}}_{j\ell}) \otimes x'_{j\ell},$$

$$\hat{\bar{d}} = n^{-1} \sum_{j=1}^{n} \hat{d}_j \quad \text{and} \quad n^* = \sum_{j=1}^{n} m_j.$$

The factor

$$(n^* - k)^{-1} (n^* - 1)(n-1)^{-1} n$$

reduces to $(n-k)^{-1} n$ if each cluster contains exactly one element. The factor $(n-k)^{-1} n$ is the degrees of freedom correction applied to the residual mean square for ordinary least squares in which k parameters are estimated. The quantity in (2.9) is well defined for two or more clusters and the factor $(n^* - k)^{-1}(n^* - 1)$ should reduce the small sample bias associated with using the estimated function to calculate deviations. Therefore, a consistent estimator of the asymptotic covariance matrix of $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ under the cluster sampling design is

$$\tilde{A}_n = \left[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\right]^{-1} \hat{G}_n \left[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\right]^{-1} \tag{2.10}$$

which can be used to test any hypothesis of the form H_O: $C\,\underset{\sim}{\beta}^0 = \underset{\sim}{\delta}^*$. Under the null hypothesis, by Moore (1977)

$$(C\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}} - \underset{\sim}{\delta}^*)' \left[C\,\tilde{A}_n\,C'\right]^{-1} (C\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}} - \underset{\sim}{\delta}^*) \tag{2.11}$$

converges in law to a chi-square distribution with $v = \text{rank}\ (\boldsymbol{C}\,\tilde{\boldsymbol{A}}_n\,\boldsymbol{C}')$ degrees of freedom. Here, $[\boldsymbol{C}\,\tilde{\boldsymbol{A}}_n\,\boldsymbol{C}']^{-1}$ is any generalized inverse of $\boldsymbol{C}\,\tilde{\boldsymbol{A}}_n\,\boldsymbol{C}'$.

The sums of squares and products matrix used in the construction of $\hat{\boldsymbol{G}}_n$ is based on n observations, where n is the number of clusters. By analogy to the Hotelling T^2 statistic, it is natural to adjust for degrees of freedom by multiplying (2.11) by the ratio

$$\frac{n - v}{v(n - 1)} \tag{2.12}$$

to obtain an approximate F statistic with v and $n - v$ degrees of freedom. In our case, this adjustment has the disadvantage that v may exceed n in a sample with a small number of clusters but a large number of individual elements.

The covariance matrix constructed as if the elemental observations are a simple random sample is biased, but it can be used to make a small sample adjustment in the estimated covariance matrix. One might view the usual small sample degrees-of-freedom adjustment as the operation of adding to an initial estimator of the covariance matrix the quantity $(n - v)^{-1}\, v\, \hat{\boldsymbol{V}}$, where $\hat{\boldsymbol{V}}$ is also an estimator of the covariance matrix. In the usual case, $\hat{\boldsymbol{V}}$ is also the initial estimator. In our case, we make the adjustment using the covariance matrix based on the elements as the second $\hat{\boldsymbol{V}}$. In our case, the use of the elemental covariance matrix has the advantage that the resulting sum is always positive definite. The adjustment is a function of the number of parameter estimated, dk. The adjustment is

(1) if $n > 3dk - 2$

$$\hat{\boldsymbol{A}}_n = \tilde{\boldsymbol{A}}_n + (n - dk)^{-1}\,(dk - 1)\,\gamma^*\left[\boldsymbol{H}_n\left(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}\right)\right]^{-1}, \tag{2.13}$$

(2) if $n \leq 3dk - 2$

$$\hat{\boldsymbol{A}}_n = \tilde{\boldsymbol{A}}_n + 0.5\,\gamma^*\left[\boldsymbol{H}_n\left(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}\right)\right]^{-1}, \tag{2.14}$$

where $\gamma^* = \max(1, \text{tr}\{[\boldsymbol{H}_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})]^{-1}\,\hat{\boldsymbol{G}}_n\}/dk)$. The upper bound of 0.5 for correction in (2.14) is arbitrary. Then, an approximate F-test with v and $n - v$ degrees of freedom is obtained by substituting $\hat{\boldsymbol{A}}_n$ for $\tilde{\boldsymbol{A}}_n$ in (2.11) and dividing the resulting quadratic form by v. In practice, the approximate degrees of freedom can be taken to be v and infinity.

3. A MONTE CARLO STUDY

In this section a Monte Carlo study is conducted to examine properties of F-Tests (2.11) involving model parameters. Data are generated under two different sampling schemes that correspond to single-stage cluster sampling where the primary units all have the same sampling weight and are taken from an infinite population. In the first sampling scheme all the elements within the cluster have the same explanatory vector $\boldsymbol{x}$ and therefore, the same conditional mean (2.1). This is the case where the logistic regression becomes weighted in the sense of several responses $\boldsymbol{y}$'s with the same covariate vector $\boldsymbol{x}$. Different degrees of intra-class correlation are induced among the $\boldsymbol{y}$'s belonging to the same cluster.

The second sampling scheme, unlike the first, places different vectors of covariates for different subjects within the cluster. The conditional mean (2.1) is also satisfied and different degrees of intra-class correlation are controlled. The effect of the intra-class correlation is studied for both sampling schemes under three different estimation procedures: MLE where the clustering effect is completely ignored, TAYLOR where the large sample covariance matrix (2.10) is used, and CPLX where the adjusted covariance matrix (2.13-2.14) is employed. These last two procedures, for large samples, are asymptotically equivalent. For small samples CPLX performs better than TAYLOR.

3.1 Sampling Scheme I

Suppose that $x_1, x_2, \ldots, x_n$ are k-dimensional independent and identically distributed normal random vectors with vector mean $\underset{\sim}{\mu}$ and covariance matrix Σ. For each j, $j = 1, 2, \ldots, n$, suppose that given x_j, the random vectors $y_{j0}^0, y_{j1}^0, \ldots, y_{j,m_j}^0$ are independent and identically distributed multinomial random vectors, with parameters $(\underset{\sim}{\pi}_j^*, 1)$, where $\underset{\sim}{\pi}_j^*$ satisfies the logistic function (2.1) evaluated at the true parameter vector $\underset{\sim}{\beta}^0$ and at $x = x_j$. Let $U_{j1}, U_{j2}, \ldots, U_{j,m_j}$ be a set of independent and identically distributed uniform (0,1) random variables. For a known and fixed ζ, $0 \leq \zeta \leq 1$, define

$$y_{j\ell}^* \equiv y_{j0}^0 \quad \text{if} \quad U_{j\ell} \leq \zeta \tag{3.1.1}$$

and

$$y_{j\ell}^* \equiv y_{j\ell}^0 \quad \text{if} \quad U_{j\ell} > \zeta, \tag{3.1.2}$$

$\ell = 1, 2, \ldots m_j$.

It can be shown that within the j-th cluster,

$$\mathrm{E}(y_{j\ell}^*) = \underset{\sim}{\pi}_j^*, \tag{3.1.3}$$

$$\mathrm{Cov}(y_{j\ell}^*, y_{jt}^*) = \Delta(\underset{\sim}{\pi}_j^*) \quad \text{if} \quad \ell = t, \tag{3.1.4}$$

and

$$\mathrm{Cov}(y_{j\ell}^*, y_{jt}^*) = \zeta^2 \Delta(\underset{\sim}{\pi}_j^*) \quad \text{if} \quad \ell \neq t. \tag{3.1.5}$$

Therefore, given x_j, the random vector $t_j = \sum_{\ell=1}^{m_j} y_{j\ell}^*$ does not have a multinomial distribution. Instead

$$\mathrm{E}(m_j^{-1} t_j) = \underset{\sim}{\pi}_j^* \tag{3.1.6}$$

and

$$\mathrm{Var}(m_j^{-1} t_j) = \left[1 + \zeta^2 (m_j - 1)\right] m_j^{-1} \Delta(\underset{\sim}{\pi}_j^*), \tag{3.1.7}$$

where ζ^2 represents the intra-cluster correlation. Furthermore, if the m_j's are constant, *i.e.*, $m_j = m$, the factor $\phi = [1 + \zeta^2(m - 1)]$ corresponds to the design effect defined by Kish (1965, p.258). An estimate of the design effect ϕ is

$$\hat{\phi} = (dk)^{-1} \left[\sum_{\ell=1}^{dk} \hat{a}_{(i,i)} / \hat{h}^{(i,i)} \right] \bar{w}^{-1}, \tag{3.1.8}$$

where $\hat{a}_{(i,i)}$ and $\hat{h}^{(i,i)}$ represent the (i,i)-th elements of $\hat{A}_n$ in (2.13)-(2.14) and $\left[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\right]^{-1}$, respectively, and $\bar{w}$ is the average of the sampling weights for the entire sample.

Under this sampling scheme, data $(x_j, y_{j\ell}^*)$, $j = 1, 2, \ldots, n$, $\ell = 1, 2, \ldots, m$, were generated with $k = 4$, $d = 3$, $m = 21$, and parameters

$$\underset{\sim}{\mu} = (1, -2, 1, 5)', \tag{3.1.9}$$

$$\underset{\sim}{\Sigma} = \text{Diag}(0, 25, 25, 25), \tag{3.1.10}$$

$$\underset{\sim}{\beta}_1^0 = (-0.3, -0.1, 0.1, 0.2), \tag{3.1.11}$$

$$\underset{\sim}{\beta}_2^0 = (0.2, -0.2, -0.2, 0.1), \tag{3.1.12}$$

and

$$\underset{\sim}{\beta}_3^0 = (-0.1, 0.3, -0.3, 0.1). \tag{3.1.13}$$

Based on (3.1.9)–(3.1.13), 1000 sets of samples with n clusters of size m, were generated according to (3.1.1)–(3.1.2) for different values of n, ζ^2, and ϕ. The estimated Type I errors obtained from comparing the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ were computed under the three different estimation procedures: MLE, CPLX and TAYLOR. A measure of the distortion of the estimated Type I errors relative to the nominal 0.05 is the relative bias which is defined as

$$(0.05)^{-1} \mid \text{Estimated Type I error} - 0.05 \mid. \tag{3.1.14}$$

Relative biases of the estimated Type I errors are reported in Table 3.1. For data generated with no intra-class correlation, ($\zeta^2 = 0$) the MLE procedure, as it is expected, provides small relative bias of the estimated nominal 5% level. CPLX produces in this case relative biases slightly greater than MLE.This is the penalty of estimating extra parameters in (2.13-2.14).

The MLE procedure shows a strong distortion of the estimated Type I error when a positive intra-class correlation is present. This distortion increases as the intra-class correlation ζ^2 gets bigger. In the case where $\zeta^2 = 0.15$ ($\phi = 4$) the relative bias of the estimated Type I error is about 18 indicating an inflated Type I error of about 95%. For the CPLX procedure, the

Table 3.1

Relative Bias of the Estimated Type I Error for the F-test of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ with nominal 0.05 Level under Sampling Scheme I

			Procedure		
n	ζ^2	ϕ	MLE	CPLX	TAYLOR
20	0.00	1	0.24	0.60	16.42
20	0.05	2	9.66	3.68	17.06
20	0.10	3	15.24	3.98	17.44
20	0.15	4	17.74	4.00	17.70
30	0.00	1	0.08	0.06	12.82
30	0.05	2	9.84	1.20	13.74
30	0.10	3	15.52	1.76	14.22
30	0.15	4	17.74	1.86	14.68
40	0.00	1	0.04	0.32	9.66
40	0.05	2	9.98	0.82	9.62
40	0.10	3	16.20	1.02	11.66
40	0.15	4	17.74	1.80	11.66
50	0.00	1	0.06	0.50	7.40
50	0.05	2	9.76	1.44	8.38
50	0.10	3	16.00	1.96	9.32
50	0.15	4	17.80	2.20	9.70
100	0.00	1	0.06	0.90	2.68
100	0.05	2	10.02	1.66	3.90
100	0.10	3	16.26	2.06	4.70
100	0.15	4	17.78	2.24	5.10
200	0.00	1	0.02	0.74	1.28
200	0.05	2	10.46	1.00	1.64
200	0.10	3	16.30	0.88	1.88
200	0.15	4	18.00	1.52	2.12
400	0.00	1	0.02	0.44	0.70
400	0.05	2	10.14	0.66	0.90
400	0.10	3	16.56	0.64	1.00
400	0.15	4	17.86	0.56	0.84
800	0.00	1	0.08	0.32	0.40
800	0.05	2	10.36	0.22	0.36
800	0.10	3	16.04	0.68	0.80
800	0.15	4	18.12	0.50	0.54

relative bias decreases as the sample size increases from $n = 20$ to the cutting point of correction (2.14) which is 34 in this case. Then it slightly increases as the sample size approaches $n = 100$ and then decreases as the sample size keeps getting bigger. This pattern will be observed throughout the whole simulation. It represents the effect of the correction (2.13-2.14) in small samples.

The Taylor procedure has large relative biases when the sample sizes are small. It varies from 17 to 7 for sample sizes between $n = 20$ and $n = 50$. For large samples both methods CPLX and TAYLOR, provide as expected, similar results. In general, the CPLX shows relative biases smaller than the TAYLOR method.

If the F statistics used for testing H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ are multiplied by the number of parameters being tested, the resulting statistic is distributed as a chi-square random variable with 12 degrees of freedom. The Monte Carlo means and variances for these chi-square statistics are presented in Table 3.2.

As expected, the MLE method produces means and variances around 12 and 24, respectively, when the design effect ϕ is one. CPLX has in this case means around 12 with greater variances that decrease when the sample size gets bigger. However, in the presence of any intraclass correlation, the means and variances under MLE are too large, while CPLX shows consistency with the asymptotic theory and the correction introduced in (2.13-2.14). The TAYLOR method has extremely high variances when the sample size is small. A possible explanation for this is that in some replications of the simulation the covariance matrix (2.10) was ill- conditioned producing very large quadratic forms for (2.11). This problem attenuates when the sample size is bigger. Both methods, CPLX and TAYLOR, become asymptotic equivalent for large samples.

Monte Carlo properties for the estimator (3.1.8) of the design effect are presented in Table 3.3 for both CPLX and TAYLOR methods. The CPLX procedure shows smaller biases and slightly large standard errors. Both methods perform fairly well.

For each category r, $r = 1, 2, 3$ and each covariate s, $s = 1, 2, 3, 4$, "t" statistics for the individual coefficient estimates were also computed as

$$\text{"t"} = \left[\text{Var}\left(\underset{\sim}{\hat{\beta}}_{rs}\right)\right]^{-0.5}\left(\underset{\sim}{\hat{\beta}}_{rs} - \underset{\sim}{\beta}^0_{rs}\right). \tag{3.1.15}$$

The twelve "t"statistics provided by the CPLX estimation procedure were grouped together and the simulated percentiles were computed. Similar computations were performed for the MLE "t" statistics. Consequently, for each run the percentiles are based on 12,000 "t" values. Once these percentiles were calculated, the relative biases were estimated as

$$(\text{Standard Normal Percentile})^{-1} \mid \text{Estimated Percentile} - \text{Standard Normal Percentile} \mid . \tag{3.1.16}$$

The results of the relative bias for the estimated 5th and 95th percentiles for the "t" statistics are presented in Table 3.4 for both MLE and CPLX procedures. Under the MLE it is expected that these relative biases be close to $\phi^{0.5} - 1$. This is true because the "t" statistics under MLE are inflated by the factor $\phi^{0.5}$. This is clearly seen in Table 3.4 under the two columns for the MLE percentiles. The CPLX procedure has satisfactory relative biases for small sample. These biases become negligible, as expected, when the sample sizes get bigger.

Table 3.2

Monte Carlo Properties of the Chi-square Statistic of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ under Sampling Scheme I

			Procedure					
			MLE		CPLX		TAYLOR	
n	ζ^2	ϕ	Mean	Variance	Mean	Variance	Mean	Variance
20	0.00	1	11.5	22.2	12.0	32.7	81.9	$12x10^3$
20	0.05	2	23.9	134.3	16.5	81.2	116.6	$8x10^4$
20	0.10	3	34.2	239.9	16.6	77.8	94.5	$12x10^3$
20	0.15	4	43.8	403.2	17.3	89.3	140.3	$19x10^4$
30	0.00	1	11.8	25.1	11.2	28.5	35.1	702.3
30	0.05	2	23.8	121.4	13.2	41.2	34.1	691.6
30	0.10	3	35.8	268.1	13.8	46.3	41.2	$12x10^2$
30	0.15	4	46.7	450.1	14.1	51.1	44.5	$16x10^2$
40	0.00	1	12.2	24.3	11.9	30.3	25.8	268.3
40	0.05	2	23.2	96.5	12.6	33.6	25.4	201.4
40	0.10	3	35.4	247.7	13.5	43.3	29.1	340.4
40	0.15	4	46.2	428.9	13.8	44.4	30.2	331.4
50	0.00	1	11.9	25.5	12.4	34.6	21.0	140.8
50	0.05	2	23.9	112.5	13.7	43.8	22.7	153.6
50	0.10	3	35.8	231.0	14.3	46.0	24.6	195.8
50	0.15	4	46.7	424.0	14.5	55.4	25.2	234.6
100	0.00	1	12.1	23.6	13.2	35.0	15.8	55.0
100	0.05	2	23.9	102.6	13.8	39.2	16.5	62.1
100	0.10	3	36.5	233.9	14.6	47.0	17.6	75.8
100	0.15	4	47.5	350.4	14.6	43.0	17.9	70.6
200	0.00	1	11.7	24.1	12.6	32.4	13.6	38.2
200	0.05	2	23.9	93.9	13.1	33.1	14.1	39.1
200	0.10	3	35.7	194.1	13.3	31.5	14.3	37.4
200	0.15	4	48.0	399.6	13.5	35.7	14.6	42.7
400	0.00	1	11.9	24.9	12.3	29.3	12.7	31.3
400	0.05	2	24.1	96.6	12.7	29.2	13.1	31.3
400	0.10	3	36.9	208.5	13.1	29.2	13.6	31.4
400	0.15	4	47.3	390.7	12.7	31.6	13.1	34.0
800	0.00	1	11.9	24.0	12.1	26.4	12.3	27.2
800	0.05	2	24.0	99.3	12.3	27.3	12.5	28.2
800	0.10	3	36.4	239.3	12.6	30.1	12.8	31.1
800	0.15	4	48.7	396.3	12.6	26.7	12.7	27.5

Table 3.3
Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme I

			Procedure			
			CPLX		TAYLOR	
n	ζ^2	ϕ	Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.00	1	0.28	0.23	0.23	0.22
20	0.05	2	0.01	0.63	0.35	0.48
20	0.10	3	0.07	0.93	0.40	0.70
20	0.15	4	0.15	1.15	0.46	0.85
30	0.00	1	0.33	0.22	0.17	0.20
30	0.05	2	0.14	0.62	0.25	0.47
30	0.10	3	0.08	0.88	0.30	0.66
30	0.15	4	0.04	1.18	0.33	0.90
40	0.00	1	0.26	0.18	0.14	0.18
40	0.05	2	0.14	0.53	0.19	0.42
40	0.10	3	0.10	0.83	0.22	0.67
40	0.15	4	0.07	1.13	0.25	0.91
50	0.00	1	0.18	0.18	0.11	0.17
50	0.05	2	0.09	0.48	0.16	0.41
50	0.10	3	0.07	0.75	0.18	0.64
50	0.15	4	0.04	0.97	0.21	0.83
100	0.00	1	0.07	0.13	0.06	0.13
100	0.05	2	0.04	0.34	0.08	0.32
100	0.10	3	0.01	0.54	0.10	0.51
100	0.15	4	0.01	0.69	0.11	0.65
200	0.00	1	0.03	0.10	0.03	0.09
200	0.05	2	0.02	0.25	0.04	0.24
200	0.10	3	0.01	0.38	0.05	0.36
200	0.15	4	0.01	0.49	0.05	0.48
400	0.00	1	0.01	0.07	0.01	0.07
400	0.05	2	0.01	0.19	0.02	0.19
400	0.10	3	0.00	0.27	0.02	0.27
400	0.15	4	0.00	0.37	0.02	0.37
800	0.00	1	0.01	0.05	0.01	0.05
800	0.05	2	0.00	0.13	0.01	0.13
800	0.10	3	0.00	0.19	0.01	0.18
800	0.15	4	0.00	0.24	0.01	0.24

Table 3.4

Relative Bias of the Estimated 5th and 95th Percentiles for the "t" Statistics for the Coefficient Estimates under Sampling Scheme I

			Procedure			
			MLE Percentile		CPLX Percentile	
n	ζ^2	$\phi^{0.5} - 1$	5th	95th	5th	95th
20	0.00	0.00	0.02	0.00	0.10	0.09
20	0.05	0.41	0.40	0.38	0.04	0.02
20	0.10	0.73	0.68	0.65	0.07	0.04
20	0.15	1.00	0.84	0.79	0.07	0.04
30	0.00	0.00	0.00	0.02	0.10	0.09
30	0.05	0.41	0.43	0.38	0.01	0.02
30	0.10	0.73	0.73	0.70	0.02	0.01
30	0.15	1.00	0.97	0.91	0.01	0.01
40	0.00	0.00	0.01	0.01	0.07	0.08
40	0.05	0.41	0.38	0.41	0.03	0.02
40	0.10	0.73	0.70	0.72	0.03	0.01
40	0.15	1.00	0.96	0.93	0.01	0.03
50	0.00	0.00	0.01	0.01	0.05	0.07
50	0.05	0.41	0.43	0.40	0.00	0.01
50	0.10	0.73	0.71	0.70	0.01	0.00
50	0.15	1.00	0.97	0.96	0.02	0.01
100	0.00	0.00	0.00	0.02	0.01	0.00
100	0.05	0.41	0.42	0.42	0.02	0.01
100	0.10	0.73	0.71	0.74	0.01	0.03
100	0.15	1.00	1.03	0.99	0.04	0.04
200	0.00	0.00	0.01	0.01	0.00	0.00
200	0.05	0.41	0.42	0.43	0.01	0.01
200	0.10	0.73	0.71	0.72	0.01	0.01
200	0.15	1.00	1.00	1.00	0.02	0.02
400	0.00	0.00	0.01	0.01	0.01	0.01
400	0.05	0.41	0.39	0.40	0.01	0.00
400	0.10	0.73	0.76	0.77	0.03	0.04
400	0.15	1.00	1.02	0.89	0.02	0.00
800	0.00	0.00	0.00	0.01	0.00	0.01
800	0.05	0.41	0.43	0.44	0.01	0.02
800	0.10	0.73	0.76	0.70	0.02	0.01
800	0.15	1.00	1.07	1.04	0.04	0.02

3.2 Sampling Scheme II

Let $x_1, x_2, \ldots, x_n$ be a set of k-dimensional independent and identically distributed normal random vectors with vector mean $\underset{\sim}{\mu}$ and covariance matrix $\underset{\sim}{\Sigma}_B$. These vectors x represent cluster means for the explanatory variables in the logistic function (2.1). Suppose that for the j-th cluster, $j = 1, 2, \ldots, n$, $x_{j0}^0, x_{j1}^0, \ldots, x_{j,m_j}^0$ are independent and identically distributed normal random vectors with vector mean x_j and covariance matrix $\underset{\sim}{\Sigma}_W$. Given $x_{j\ell}^0$, $\ell = 0, 1, \ldots, m_j$, the $(d + 1)$-dimensional random vector $y_{j\ell}^0$ has a multinomial distribution with parameters $(\underset{\sim}{\pi}_{j\ell}^0, 1)$, where the elements of $\underset{\sim}{\pi}_{j\ell}^0$ satisfy the logistic function (2.1) evaluated at the true parameter vector $\underset{\sim}{\beta}^0$ and at $x = x_{j\ell}^0$. Furthermore, suppose that given the $x_{j\ell}^0$'s, the $y_{j\ell}^0$'s are independent.

Let $U_{j1}, U_{j2}, \ldots, U_{j,m_j}$ be m_j independent and identically distributed uniform (0,1) random variables that are also jointly independent from the $x_{j\ell}^0$'s and from the $y_{j\ell}^0$'s . Let ζ be a fixed and known number, $0 \leq \zeta \leq 1$. Then define $(x_{j\ell}, y_{j\ell}^*)$, $\ell = 1, 2, \ldots, m_j$ in the following way:

$$(x_{j\ell}, y_{j\ell}^*) \equiv (x_{j0}^0, y_{j0}^0) \text{ if } U_{j\ell} \leq \zeta \tag{3.2.1}$$

and

$$(x_{j\ell}, y_{j\ell}^*) \equiv (x_{j\ell}^0, y_{j\ell}^0) \text{ if } U_{j\ell} > \zeta. \tag{3.2.2}$$

Observe that within each cluster, the $x_{j\ell}$'s all have the same vector of conditional means x_j and that the covariance matrix between $x_{j\ell}$ and x_{jt} is $\underset{\sim}{\Sigma}_W$ if $\ell = t$ and $\zeta^2 \underset{\sim}{\Sigma}_W$ otherwise. Also, note that the conditional mean of each $y_{j\ell}^*$ is the logistic function (2.1) evaluated at $\underset{\sim}{\beta}^0$ and $x = x_{j\ell}$, and that the vectors $(x_{j\ell}, y_{j\ell}^*)$, $\ell = 1, 2, \ldots, m_j$, exhibit an intra-class correlation of ζ^2 and an approximate design effect of $\phi = [1 + \zeta^2 (m - 1)]$ when all the m_j's are constant.

Data $(x_{j\ell}, y_{j\ell}^*)$, $j = 1, 2, \ldots, n$, $\ell = 1, 2, \ldots, m_j$, were generated under this cluster sampling scheme with $k=4$, $d=3$, and parameters

$$\underset{\sim}{\mu} = (1, -6, 4, 8)', \tag{3.2.3}$$

$$\underset{\sim}{\Sigma}_B = \text{Diag}(0, 25, 25, 49), \tag{3.2.4}$$

$$\underset{\sim}{\Sigma}_W = \text{Diag}(0, 25, 36, 36), \tag{3.2.5}$$

$$\underset{\sim}{\beta}_1^0 = (0.30, -0.05, -0.06, 0.08), \tag{3.2.6}$$

$$\underset{\sim}{\beta}_2^0 = (0.06, -0.08, -0.10, 0.07), \tag{3.2.7}$$

and

$$\underset{\sim}{\beta}_3^0 = (0.70, -0.08, -0.10, 0.11), \tag{3.2.8}$$

Based on (3.2.3)-(3.2.8), 1000 sets of samples with n clusters of size $m_j = m = 6$, were generated according to (3.2.1)-(3.2.2) for different values of n, ζ^2 and ϕ. The relative biases defined in (3.1.14) of the estimated Type I errors from comparing the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ are presented in Table 3.5 under three different estimation techniques: MLE, CPLX and TAYLOR.

In the presence of intra-class correlation, there is a strong distortion of the Type I error for MLE even in the case where ζ^2 is relatively small ($\zeta^2 = 0.2$) for cluster size $m = 6$. This distortion is reflected in the relative bias which ranges from approximately 7 to 18. These values indicate inflated Type I errors between 40% and 95%. The CPLX procedure provides satisfactory relative biases even for the case of small samples. The TAYLOR procedure has too high values for small samples. It becomes equivalent to CPLX for large samples. One more time CPLX seems to be superior to TAYLOR when the sample size is small.

Table 3.5

Relative Bias of the Estimated Type I Error for the F-test of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ with Nominal 0.05 Level under Sampling Scheme II

			Procedure		
n	ζ^2	ϕ	MLE	CPLX	TAYLOR
20	0.0	1	0.54	0.46	13.52
20	0.2	2	7.30	0.46	12.96
20	0.4	3	13.70	0.68	13.96
20	0.6	4	17.08	0.60	14.72
30	0.0	1	0.28	0.78	7.78
30	0.2	2	8.72	0.72	8.16
30	0.4	3	14.84	0.72	9.32
30	0.6	4	17.50	0.82	9.23
40	0.0	1	0.36	0.56	5.16
40	0.2	2	9.28	0.56	5.76
40	0.4	3	15.38	0.64	5.84
40	0.6	4	17.76	0.70	5.80
50	0.0	1	0.44	0.56	3.44
50	0.2	2	9.34	0.08	4.86
50	0.4	3	15.48	0.38	4.36
50	0.6	4	17.56	0.46	4.16
100	0.0	1	0.16	0.04	1.26
100	0.2	2	9.46	0.26	1.46
100	0.4	3	15.94	0.44	2.00
100	0.6	4	18.16	0.14	1.46
200	0.0	1	0.10	0.26	0.76
200	0.2	2	10.20	0.34	0.82
200	0.4	3	16.22	0.02	0.48
200	0.6	4	18.06	0.06	0.52

Table 3.6

Monte Carlo Properties of the Chi-square Statistic of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ under Sampling Scheme II

			Procedure					
			MLE		CPLX		TAYLOR	
n	ζ^2	f	Mean	Variance	Mean	Variance	Mean	Variance
20	0.0	1	11.3	18.9	10.2	19.7	40.5	$15x10^2$
20	0.2	2	20.3	62.8	10.5	21.4	39.2	$11x10^2$
20	0.4	3	28.3	106.4	10.5	18.4	111.3	$42x10^5$
20	0.6	4	35.2	152.6	10.3	18.2	$11x10^3$	$50x10^9$
30	0.0	1	11.6	21.6	9.4	16.3	22.0	147.3
30	0.2	2	21.8	75.2	9.9	17.5	22.7	161.2
30	0.4	3	30.4	117.6	9.8	16.5	24.3	224.6
30	0.6	4	39.3	191.0	9.5	14.5	$24x10^2$	$60x10^8$
40	0.0	1	11.6	21.3	9.9	19.4	18.1	86.7
40	0.2	2	22.4	76.5	10.4	18.3	18.9	80.8
40	0.4	3	31.8	153.2	10.2	17.8	19.2	90.4
40	0.6	4	41.4	223.1	10.1	16.9	19.3	104.4
50	0.0	1	11.5	19.9	10.6	20.0	16.1	56.9
50	0.2	2	22.7	80.6	11.4	23.9	17.5	70.9
50	0.4	3	32.3	160.1	11.1	22.9	17.4	73.7
50	0.6	4	41.7	262.3	10.7	19.7	17.0	63.8
100	0.0	1	11.8	21.5	11.8	25.2	13.9	36.2
100	0.2	2	22.9	87.3	11.9	27.0	14.0	38.5
100	0.4	3	34.7	191.8	12.3	27.9	14.4	40.7
100	0.6	4	45.1	297.7	12.0	25.0	14.1	37.2
200	0.0	1	12.0	23.8	12.1	26.3	13.0	30.3
200	0.2	2	24.0	88.6	12.4	25.9	13.3	30.0
200	0.4	3	34.5	175.2	12.0	23.3	12.8	27.0
200	0.6	4	46.8	320.0	12.2	24.0	13.0	27.9

Monte Carlo properties of the chi-square statistics of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ (chi-square = 12 × F) are presented in Table 3.6 for the three estimation procedures under study. CPLX shows means and variances slightly below 12 and 24, respectively, when the sample sizes are small. This underestimation vanishes when the sample size increases. The TAYLOR procedure has too large means and variances when the sample size is small. For instance, for $\zeta^2 = 0.6$, the variance is in the order of billions when n is 30 or less. For large samples, both CPLX and TAYLOR, seem to provide similar results. The MLE method has acceptable results only when $\zeta^2 = 0.00$. Otherwise the estimated mean and variances are too large.

Table 3.7

Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme II

			Procedure			
			CPLX		TAYLOR	
n	ζ^2	ϕ	Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.0	1	0.48	0.22	0.04	0.20
20	0.2	2	0.16	0.53	0.26	0.42
20	0.4	3	0.05	0.87	0.34	0.72
20	0.6	4	0.01	1.24	0.39	1.03
30	0.0	1	0.49	0.18	0.02	0.16
30	0.2	2	0.25	0.48	0.19	0.40
30	0.4	3	0.19	0.84	0.24	0.69
30	0.6	4	0.16	1.12	0.27	0.94
40	0.0	1	0.38	0.16	0.02	0.14
40	0.2	2	0.22	0.45	0.14	0.38
40	0.4	3	0.16	0.70	0.20	0.60
40	0.6	4	0.16	0.98	0.19	0.86
50	0.0	1	0.27	0.14	0.02	0.13
50	0.2	2	0.15	0.42	0.12	0.37
50	0.4	3	0.12	0.67	0.15	0.60
50	0.6	4	0.11	0.89	0.16	0.81
100	0.0	1	0.12	0.10	0.01	0.10
100	0.2	2	0.06	0.32	0.07	0.31
100	0.4	3	0.05	0.50	0.07	0.48
100	0.6	4	0.06	0.59	0.07	0.57
200	0.0	1	0.05	0.07	0.01	0.07
200	0.2	2	0.03	0.24	0.03	0.23
200	0.4	3	0.02	0.34	0.04	0.33
200	0.6	4	0.02	0.40	0.03	0.40

Monte Carlo properties for the estimator of the design effect proposed in (3.1.8) are presented in Table 3.7 under the CPLX and TAYLOR procedures. The TAYLOR procedure seems to perform slightly better than CPLX for small samples. Both procedures, in general, provide reasonable values. They seem to be equivalent for large samples.

Table 3.8

Relative Bias of the Estimated 5th and 95th Percentiles for the "t" Statistics for the Coefficient Estimates under Sampling Scheme II

			Procedure			
			MLE Percentile		CPLX Percentile	
n	ζ^2	$\phi^{0.5} - 1$	5th	95th	5th	95th
20	0.0	0.00	0.01	0.00	0.15	0.18
20	0.2	0.41	0.37	0.32	0.06	0.09
20	0.4	0.73	0.63	0.57	0.02	0.05
20	0.6	1.00	0.79	0.74	0.05	0.05
30	0.0	0.00	0.02	0.00	0.15	0.16
30	0.2	0.41	0.39	0.38	0.10	0.10
30	0.4	0.73	0.68	0.63	0.07	0.08
30	0.6	1.00	0.91	0.86	0.05	0.07
40	0.0	0.00	0.01	0.00	0.12	0.15
40	0.2	0.41	0.39	0.40	0.10	0.06
40	0.4	0.73	0.65	0.60	0.07	0.09
40	0.6	1.00	0.99	0.89	0.04	0.05
50	0.0	0.00	0.01	0.01	0.10	0.10
50	0.2	0.41	0.39	0.40	0.05	0.04
50	0.4	0.73	0.73	0.72	0.02	0.01
50	0.6	1.00	1.00	0.95	0.00	0.01
100	0.0	0.00	0.01	0.01	0.04	0.05
100	0.2	0.41	0.40	0.37	0.02	0.02
100	0.4	0.73	0.72	0.73	0.00	0.00
100	0.6	1.00	1.00	1.02	0.01	0.02
200	0.0	0.00	0.02	0.01	0.00	0.01
200	0.2	0.41	0.40	0.45	0.01	0.02
200	0.4	0.73	0.71	0.68	0.01	0.01
200	0.6	1.00	1.03	0.95	0.02	0.02

The relative biases (3.1.16) of the 5th and 95th percentiles of the "t" statistics (3.1.15) are presented in Table 3.8 under the MLE and CPLX procedures. MLE has a relative bias, as expected, close to zero in the absence of intra-class correlation. This bias increases when the ζ^2 gets bigger. On the other hand, CPLX has small relative bias in general and for large sample this bias becomes negligible.

4. EXTENSION TO STRATIFIED SAMPLING AND MORE COMPLEX DESIGNS

A generalization of CPLX procedure to stratified sampling can be done as follows. Suppose that the population has been divided into $i = 1, 2, \ldots, L$ strata. Let m_{ij} represent the size of the j-th cluster in the i-th stratum, n_i the number of clusters selected in the i-th stratum, and $y^*_{ij\ell}$ the multinomial response of the ℓ-th element in the j-th cluster in the i-th stratum, $\ell = 1, 2, \ldots, m_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, L$. It is assumed that $\underset{\sim}{\pi}^*_{ij\ell}$, the expected value of $y^*_{ij\ell}$, satisfies the logistic relationship (2.1) for a given explanatory vector $x_{ij\ell}$.

A consistent estimator of $\underset{\sim}{\beta}^0$, say $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$, can be found by maximizing the function

$$L_n\left(\underset{\sim}{\beta}\right) = \sum_{i=1}^{L} \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij}\left(\log \underset{\sim}{\pi}^*_{ij\ell}\right)' y^*_{ij\ell}. \tag{4.1}$$

Algorithm (2.5) is performed with three indexes i, j, ℓ. The adjustment given by (2.13) and (2.14) is applied with

$$n = \sum_{i=1}^{L} n_i, \tag{4.2}$$

$$H_n\left(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}\right) = \sum_{i=1}^{L} \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij}\, \Delta\left(\hat{\underset{\sim}{\pi}}^*_{ij\ell}\right) \otimes x'_{ij\ell} x_{ij\ell}, \tag{4.3}$$

$$\hat{G} = \left[\left(n^* - k\right)^{-1}\left(n^* - 1\right)\right] \sum_{i=1}^{L} \left(n_i - 1\right)^{-1} n_i\left(1 - f_i\right) \sum_{j=1}^{n_i} \left(\hat{d}_{ij} - \hat{\bar{d}}_i\right)\left(\hat{d}_{ij} - \hat{\bar{d}}_i\right)', \tag{4.4}$$

$$\hat{d}_{ij} = \sum_{\ell=1}^{m_{ij}} w_{ij}\left(y_{ij\ell} - \hat{\underset{\sim}{\pi}}_{ij\ell}\right) \otimes x'_{ij\ell}, \tag{4.5}$$

$$\hat{\bar{d}}_i = n_i^{-1} \sum_{j=1}^{n_i} \hat{d}_{ij}, \tag{4.6}$$

f_i = sampling rate of i-th stratum, and (4.7)

$$n^* = \sum_{i=1}^{L} \sum_{j=1}^{n_i} m_{ij}. \tag{4.8}$$

The estimation procedure can be extended in a stepwise manner to multi-stage sampling designs by maximizing (4.1) up to elemental units. The summation of (4.3) should be extended in order to include all the final sampling units. The key part is (4.4). The construction of $\hat{G}$ must be based on the complex survey. This could be a difficult task for multi-stage sampling. Results for stratified two-stage sampling are presented in Fuller, *et al.* (1986, p. 82).

5. SUMMARY

In this paper, we have outlined a methodology for obtaining asymptotic normal estimators of the parameters of a generalized logistic function involving a multinomial response variable under complex survey designs. A consistent estimator of the asymptotic covariance matrix under the complex sampling design is (2.10), which results from the usual Taylor's series expansion. This covariance matrix produces for large samples correct Type I errors for the F-tests involving model parameters. More important, it is shown that correction (2.13-2.14) provides a covariance matrix that reduces the small sample bias. This adjusted covariance matrix has some important characteristics:

1. It levels off the inflated Type I error, originated from ignoring the complex survey, faster than the usual delta-method.
2. It is positive definite when $H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})$ is positive definite regardless if (2.9) is singular or not.
3. It is asymptotic equivalent to (2.10).

The results of a Monte Carlo study were reported in Section 3. Data satisfying the logistic conditional mean (2.1) were generated under two different single-stage cluster sampling schemes. It was studied, among other things, the effect of the intra-class correlation and the design effect on the relative biases of the estimated Type I errors for the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$. The simulation showed, as expected, a strong relative bias when the naive maximum likelihood method is employed. For small samples, the Monte Carlo results favor the use of the adjusted covariance matrix over the one that arises from the usual delta-method.

ACKNOWLEDGEMENTS

This work was begun while the author was a student at Iowa State University. I thank Professor Wayne A. Fuller for introducing me to the topic and for suggesting a number of the small sample modifications that were incorporated into the estimation procedure. The author wants also to thank the referees for useful comments.

REFERENCES

ALBERT, A., and LESAFFRE, E. (1986). Multiple group logistic discrimination. *Computers and Mathematics with Applications*, 12A, 209-224.

BEDRICK, E.J. (1983). Adjusted chi-square tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.

BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A., GRATTON, M.A., HIDIROGLOU, M.A., KUMAR, S., and RAO, J.N.K. (1984). Analysis of categorical data from surveys with complex designs: some Canadian experiences. *Survey Methodology*, 10, 141-156.

BULL, S.B., and PEDERSON, L.L. (1987). Variance for polychotomous logistic regression using complex survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, Theory and Methods*, 14, 1377-1392.

COX, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.

DALE, J.R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society*, Ser. B, 48, 48-59.

FAY, R.E. (1985). A jackknife chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). *PC CARP*. Statistical Laboratory, Iowa State University, Ames, Iowa.

GALLANT, A.R. (1987). *Nonlinear Statistical Methods*. New York: John Wiley & Sons.

HABERMAN, S.J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.

HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.

JENNRICH, R.I., and MOORE, R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Section on Statistical Computing, American Statistical Association*.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

MOORE, D.S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72, 131-137.

MOREL, J. (1987). Multivariate nonlinear models for vectors of proportions: A generalized least squares approach. Unpublished Ph.D. dissertation. Iowa State University, Ames, Iowa.

NELDER, J.A., and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Ser. A, 135, 370-384.

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. *Journal of the American Statistical Association*, 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. *The Annals of Statistics*, 12, 46-60.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.