Statistics Canada — Statistique Canada

Methodology Branch

Household Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

MARGINAL MODELS FOR LONGITUDINAL DATA ANALYSIS USING COMPLEX SURVEY DATA

HSMD -2001 - 004E

Ioana Şchiopu-Kratina

Expenditure Survey and Analysis Methods Section

Household Survey Methods Division
Statistics Canada

April 2001

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

MARGINAL MODELS FOR LONGITUDINAL DATA ANALYSIS USING COMPLEX SURVEY DATA

Ioana Şchiopu-Kratina [1]

ABSTRACT

This article covers three aspects of statistical inference with data from longitudinal surveys with complex sample designs. Firstly, the consistency of the parameter obtained as a root of a generalized estimating equation (GEE) is proved. Next, we prove a central limit theorem and thirdly we prove the consistency of the Jackknife estimator of the asymptotic variance.

[1] Ioana Şchiopu-Kratina, Expenditure Survey Analysis Methods Section, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

MODÈLES MARGINAUX POUR DES DONNÉES PROVENANT D'UNE ENQUÊTE LONGITUDINALE AU PLAN COMPLEXE

Ioana Şchiopu-Kratina[2]

RÉSUMÉ

Cet article porte sur trois aspects de l'inférence statistique avec des données provenant d'une enquête au plan complexe. On démontre d'abord la convergence ponctuelle de l'estimateur défini comme racine d'une équation de type GEE. Un théorème limite centrale et la convergence de l'estimateur Jackknife associé à la variance asymptotique sont également démontrés.

[2] Ioana Şchiopu-Kratina, Méthodes d'enquêtes sur les dépenses et d'analyse, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6.

1. Introduction. In longitudinal surveys subjects are observed on at least two different occasions, which makes such surveys suitable for studying change over time at the individual, or unit level. In addition to the production of crossectional estimates, data from longitudinal surveys may be used, for instance, to estimate gross flows (important in the study of labour market dynamics), or in event history modelling, which may be used to uncover determinants of survival for individuals afflicted with a serious health condition. More generally, longitudinal data may be used for modelling a response variable as a function of covariates and time, with applicability in many areas. Rao (1998) and the references therein give a more complete description of possible uses of data from longitudinal surveys and the statistical techniques that are available to explore them.

Large scale longitudinal surveys are often carried out by large organizations like Statistics Canada. Their primary goal in conducting a survey is to obtain design - based estimates of totals, means or proportions for a target population, which is finite. The selection of the sample generally follows a complex plan with goals like reducing the cost of the survey. The conditions for model based inference are often not met by the data collected according to the survey design, even if the finite population is large. Design - based inference, introduced by Binder (1983), offers a solution, as it allows for the use of modelling techniques in the context of survey randomization. We follow this approach, which is also that of Rao (1998), and consider marginal models for longitudinal data as in Liang and Zeger (1986) in the context of design - based inference.

In longitudinal data, observations on the same subject are dependent, and this dependence is different from the clustering effect due to the sampling selection. Liang and Zeger (1986) introduced Generalized Estimating Equations (GEE), which require only specification of the marginal model mean and variance for each individual. Correlation across time for the same individual is assumed to exist, but it is not specifically modelled. In the special situation when the observations across time are assumed independent for each individual (the working independence assumption), GEE becomes the Independence Estimating Equation (IEE).

Three areas of design - based inference are presented in this article: consistency of the estimator of the regression coefficient implicitly defined by estimating equation (EE), the corresponding Central Limit Theorem (CLT) and the consistency of the jackknife estimator of the asymptotic variance.

The topic of consistency of the main estimator seems to have been neglected in the literature - it is not even mentioned in Liang and Zeger (1986). Yet consistency is an essential ingredient in the proof of asymptotic normality. Due to the nature of design inference, we can only define weak consistency, i.e. in terms of convergence in probability. The sets on which the estimators are roots of estimating equations (REE) have asymptotic probability 1 (see the statement of Theorem 1). When discussing the asymptotic variance of these estimators we will make the stronger assumption that the estimators of the main parameters are REE's on every selected sample. We do not make any assumptions regarding the uniqueness of the REE. We first give an analytical proof of consistency (Theorem 1), then show how it applies in the GEE situation (Corollary 3). The proof of Theorem 1 is loosely patterned after the proof of Theorem, p. 145, Serfling (1980).

We prove asymptotic normality (CLT) in Theorem 2. The proof exploits the form of the GEE , which "separates" into factors related to the covariance structure over time and linear statistics, for which CLT's are available.

The consistency of the jackknife estimator of the asymptotic variance was proven only for the IEE case. It could be generalized to cover the GEE situation using the proof and the results of Theorem 2.

The technical problems that we had to overcome were due to the estimation of the variance structure across time and to obtaining asymptotic results in finite populations with survey randomization. The first problem was solved by Liang and Zeger (1986) in a model - based context. However, they do not supply proof for their asymptotic results. In order to do design inference, we tried to give simple analytical proof which do not depend on model assumptions in a superpopulation, as in Binder (1983). However, some of our conditions appear more natural if a superpopulation were assumed to exist and some model assumptions were present (e.g. (i) and (ii) of Assumption 1 - see Example 3 and model assumption (1)). The results of Binder (1983) had to be extended from the IEE case to GEE.

This article is organized as follows: Section 2 presents marginal models as in Liang and Zeger (1986). Example 1 illustrates the classical use of EE's in calculating an estimator of a regression coefficient for the linear model. This estimator becomes the census parameter in the context of design - based inference, which is outlined in Section 3. The design that we consider is stratified, multistage and with replacement at the first stage. Example 2 shows the calculation of the design based estimator from the 'weighted' EE in Example 1. Section 4 is devoted to design - consistency. Example 3 illustrates conditions for consistency on the EE in Example 2. Example 4 is a genuine example of GEE on which we illustrate the conditions for consistency. Section 5 deals with asymptotic normality and the consistency of the jackknife estimator of the asymptotic variance. The proof of Theorem 3 is completed in the Appendix.

2. Model set-up. We describe briefly the set-up in Liang and Zeger (1986). Consider M individuals observed on d_i occasions (i = 1, ... M). The univariate responses y_{it} and the p - covariates x_{it} are recorded, t = 1, ... d_i , i = 1,... M. We assume that $d_i = d$, i = 1, ... M. Typically, d is small for marginal models. Otherwise, time series techniques may be more appropriate. Here only $E_m[y_{it}]$ and $Var_m(y_{it})$ are specified, where *m* stands for model- based, for all t, i ≥ 1. Liang and Zeger (1986) consider probability densities of the following type: $\rho(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\varphi]$ with $\theta_{it} = h(\eta_{it})$, $\eta_{it} = x_{it}^T\beta$, where a, b and h are known (differentiable) functions, $\{\theta_{it}\}$, φ are parameters , x_{it}^T is an $1 \times p$ matrix of covariates and β is an $p \times 1$ vector of main parameters, for all t, i≥ 1. Here T stands for transposition of matrices. Note that for random variables with such densities we have:

$$(1) \qquad E_m[y_{it}] = \mu_{it} = a'(\theta_{it}), \text{ for all } t, i \geq 1:$$

Let $\mu(\eta) = a'(\eta)$, for any η in a space of parameters Θ. The function g is a link function if $g \circ \mu(\theta_{it})$

$= x_{it}^T \beta$, for all t, i≥ 1. If g $= \mu^{-1}$, then g is called the canonical or natural link, the function h above can be taken to be the identity and the parametric form of the model is the natural one. With binary response, the logit link function $g(\mu) = \log\{\mu/(1-\mu)\}$is the natural link associated with the logistic regression model. EE's are formed that mimic log likelihood equations associated with exponential distributions (e.g. normal, binomial, logistic, Poisson). These are quasi - likelihood equations if the original distributions belong to the normal family upon further restrictions, e.g. knowledge on the dispersion parameter φ (Shao 1999, p. 242). The idea is to produce estimators for β which are REE by making few assumptions on the distribution of the observed data, and then study the properties of these estimators.

When GEE's are used, it is assumed that correlation of observations y_{it} across time for the same individual is the same for all individuals, and is represented by a matrix $R(\alpha)$, with α a "nuisance parameter". More precisely, let $U_i(\beta, \alpha, \varphi) = D_i^T V_i^{-1} S_i$, $V_i = 1/\varphi[A_i^{1/2} R(\alpha) A_i^{1/2}]$, $D_i = A_i \Delta_i X_i$, $S_i = Y_i - a'(\theta_i)$, $A_i = \text{diag } a''(\theta_{it})$ in R^d and $\Delta_i = \text{diag } [d\theta_{it} / d\eta_{it}]$, which could be taken to be the identity matrix I_d, for all i ≥ 1. Notice that the covariates are contained in D_i and that A_i as well as S_i (through a') contain the main parameter β, i ≥ 1. The GEE, or equation (7) of Liang and Zeger (1986), is:

$$(2) \qquad \sum_{i=1}^{M} U_i(\beta, \hat{\alpha}(\beta), \varphi(s, \beta)) = 0$$

Equation (2) above is called a pseudo-likelihood equation in Shao (1999), p. 315. Note that it consists of p scalar equations. In equation (2) $\hat{\alpha}$ and $\varphi(s, \beta)$ are estimates of nuisance parameters that are obtained from the sample and generally contain β. When the solution to (2) exists and is unique , i.e. when β is defined implicitly by (2), it is denoted by $\hat{\beta}_G$in Liang and Zeger (1986). Note that this approach is different from the one presented in Section 5 of Rao (1998). It is important to note that (2) contains only β as unknown parameter and that, due to the estimation of the nuisance parameters, the left hand side of (2) is, in general, a nonlinear function of the sample observations.

When the observations across time are assumed independent for each individual (the working independence assumption), equation (2) becomes IEE. In this case $R(\alpha) = I_d$ and there is no need to estimate nuisance parameters in (2). This is the situation discussed, in a design randomization context, by Binder (1983). In the context of IEE and survey randomization (see Section 3), $\hat{\beta}_G$ becomes the "census" parameter defined in Binder (1983). The example below illustrates the calculation of $\hat{\beta}_G$ from an IEE. Notice that the presence of the time dimension is accounted for by the increase in the number of data points (from M to 2×M in this case).

<u>**Example 1**</u> Assume that the individual observations are independent, identically distributed (i.i.d.) and that they follow a normal distribution. Take $\varphi = 1$ and $d = 2$ occasions. We have $R(\alpha) = I_2$ (case IEE). Assume that x_{it}, β are scalars, $i,t \geq 1$ and that h is the identity.

$$\rho\,(y_{it}) = \exp - \frac{(y_{it} - \theta_{it})^2}{2} = \exp\,\{y_{it}\,\theta_{it} - a\,(\theta_{it}) + b\,(y_{it})\,\} \Rightarrow a\,(\theta_{it}) = \frac{\theta_{it}^2}{2},\; b(y_{it}) = -\frac{y_{it}^2}{2}$$

$$E\,[y_{it}] = \theta_{it} = \frac{da}{d\theta_{it}};\; \frac{d^2a}{d\theta^2_{it}} = 1\,,\; \theta_{it} = x_{it}\beta,\; i,t \geq 1.$$

Note that each x_{it} has as many components as β ($p = 1$ components here) and, for $i, t \geq 1$:

$$\frac{d \log \rho(y_{it})}{d\beta} = y_{it}\, x_{it} - x_{it}^2 \beta$$

Now $a'\,(\theta_{it}) = \theta_{it} = x_{it}\beta,\; i, t \geq 1$ and (2) is:

$$(3) \qquad \sum_{i=1}^{M} \sum_{t=1}^{2} x_{it}\, y_{it} - \sum_{i=1}^{M} \sum_{t=1}^{2} x_{it}^2\, \beta = 0 \Rightarrow \hat{\beta}_G = \frac{\sum_{i=1}^{M} \sum_{t=1}^{2} x_{it}\, y_{it}}{\sum_{i=1}^{M} \sum_{t=1}^{2} x_{it}^2} \quad \blacksquare$$

3. The design and the design-based inference. In the article, inference is done in the design - based randomization as proposed by Binder (1983). As mentioned in his paper, conclusions can be drawn only in designs in which conditions have been given for the Central Limit Theorem (CLT) to hold. The design that we consider here is stratified, multistage in which the p.s.u.'s (clusters) are selected with replacement from a population of M individuals (or 'ultimate' selection units). Conditions for the CLT to hold in such designs have been given by Krewski and Rao (1981) and by Yung (1996). Here the cluster totals (or 'normalized' cluster totals) are i.i.d.'s in the design randomization within each stratum and independent random variables (r.v.'s) across strata. Thus, the r.v.'s involved in the limiting theorems are the clusters rather than the individuals. The populations change with the increase in the number of units involved in the inference. The sampling distributions of these variables change with the changing populations and so does the finite population parameter. It is therefore appropriate to consider CLT's for arrays. To simplify notation, we index the populations by the total number of associated r.v.'s involved in the limiting process, i.e. the total number of clusters N from which n p.s.u.'s are selected. Thus, the census parameter defined by (3) for the IEE case will be denoted $\beta_N = \hat{\beta}_G$, rather than β_M, which would be more appropriate. The parameter to estimate in the design randomization context changes as $n \to \infty$ (which implies that $N, M \to \infty$). In this article β_0 is a limit point, e.g.: $\hat{\beta}_N \xrightarrow{p_N} \beta_0$, where $\xrightarrow{p_N}$ means convergence in the design probability, which is consistent with Binder (1983). In some instances, one might wish to link β_0 to the superpopulation parameter, e.g. if one wishes to give an interpretation to the finite population parameter. We do not attempt to do this here.

For simplicity, we consider that the selected sample s consists of respondents only. The generalization to the situation where nonresponse occurs completely at random is straightforward (see J.N.K. Rao, 1998). Consider a population that consists of M individuals and which is partitioned into L strata. Each stratum consists of M_h individuals from which N_h clusters are formed, h =1,...L. From each stratum h, n_h clusters are selected with replacement and a further selection of m_{hi} individuals takes place within each cluster i, i = 1,... n_h, h = 1,... L. We denote by n the total number of clusters selected. To each individual k we attach a basic weight appropriate to the sample selection mechanism. As in Yung (1996), we 'normalize' it by dividing the basic weight by M, the total number of individuals in the finite population. We denote the resulting weight by w_{hik} and, when no confusion may arise, by w_k, k = 1, ... M, i = 1, ... n_h, h = 1, ... L

Definition 1. In the case of the GEE (2), the census parameter β_N is defined as the solution (when it exists and is unambiguously defined) of equation (4) below:

$$\sum_{k=1}^{M} U_k(\beta,\ \alpha_N(\beta),\ \varphi_N(\beta)) = 0 \qquad (4) \qquad ■$$

We will define next a sample - based estimator $\hat{\beta}_N$, which will serve to make design based inference on the census parameter β_N. In conjunction with the GEE (2), we define, for $\beta \in \Theta$:

$$\hat{\psi}_N(\beta) = \psi_N(s,\beta) = \sum_{k\varepsilon s} w_k U_k(\beta, \hat{\alpha}_N(\beta), \varphi_N(s,\beta)) \qquad (5)$$

In (5) $\hat{\alpha}_N(\beta)$ and $\varphi_N(s,\beta)$ are sample based estimators of the census parameters α_N, respectively φ_N. Notice that in case of with - replacement sampling, s is an ordered sample, i.e. the same p.s.u.'s may appear several times in the sample *s* (Särndal et al 1992, p.72)

Definition 2. The REE estimator $\hat{\beta}_N$ of the census parameter β_N is defined as a solution to $\psi_N(s,\beta) = 0$, with $\psi_N(s,\beta)$ as in (5) above. ■

Example 2. Consider the simpler situation of an IEE presented in Example 1. The census parameter in Example 1 is $\beta_N = \hat{\beta}_G$ in (3). A design based estimator $\hat{\beta}_N$ is a solution to $\hat{\psi}_N(\beta) = \psi_N(s,\beta) = 0$, where:

$$\psi_N(s,\beta) = \sum_{k\epsilon s} w_k \sum_{t=1}^{2} x_{kt}(y_{kt} - x_{kt}\beta)$$

This estimator can be found explicitly as the EE above has the unique solution:

$$\hat{\beta}_N = \frac{\sum_{k\epsilon s}\sum_{t=1}^{2} w_k x_{kt} y_{kt}}{\sum_{k\epsilon s}\sum_{t=1}^{2} w_k x_{kt}^2} \qquad (6)$$

Note that in (6) the normalized weights can be replaced by the original design weights. ■

4. Consistency of $\hat{\beta}_N$. We first give conditions for the existence of an REE estimator $\hat{\beta}_N$ as well as on its convergence to a constant, which is a major step in proving its design consistency.

<u>Assumption 1</u> (also included in Binder (1983)):

(i) $\psi_N(s,\beta) \xrightarrow{p_N} \psi(\beta)$, for any $\beta \in \Theta$, where $\psi(\beta)$ is a non random function defined on the space of parameters Θ which may be unbounded. Recall p_N is the design probability.

(ii) $\psi(\beta_0) = 0$, and all partial derivatives of $\psi(\beta)$ exist and are continuous around β_0.

(iii) $D_\beta[\psi(\beta)]|_{\beta_0} = -J_0$ is invertible (it suffices to have $\det|D_\beta[\psi(\beta)]|_{\beta_0} \neq 0$), where $D_\beta[\psi(\beta)]$ is the pxp matrix of partial derivatives of $\psi(\beta)$. ■

<u>Remark 1</u> Assume that β_0 is the true superpopulation parameter used in S_i, i = 1,...N. Then $E_m[Y_i - a'(\theta_i)] = 0$, $i \geq 1$, by the first model assumption in equation (1).■

<u>Assumption 2</u> For $K_0 = K(\beta_0)$ a compact containing $\beta_0, K_0 \subseteq \Theta$ and any $\eta > 0$, there exist a constant h_0 and an integer n such that, for the partial derivatives of $\psi_N(s,\beta) = (\psi_j^N(s,\beta))_{j=1,\ldots p}$, $\beta = (\beta_k)_{k=1,\ldots,p}$

(iv)
$$\sup_{n \geq n_0} p_N\left\{s: \sup_{\beta \in K_0} \left|\frac{\partial \psi_j^N(s,\beta)}{\partial \beta_k}\right| \geq h_0\right\} \leq \eta$$

for all j, k = 1, ... p .■

We note that (iv) is equation (4.69) of Shao (1999).

Example 3 : We consider again Example 2 above.

$$\psi_N(s\,,\beta) = \sum_{k\in s} w_k \sum_{t=1,2} x_{kt}(y_{kt} - x_{kt}\beta)$$

$$\cong \frac{1}{M}\sum_{k=1}^{M}\sum_{t=1,2} x_{kt}\,(y_{kt} - x_{kt}\beta)$$

(if design consistency holds).

If the Strong Law of Large Numbers holds in the superpopulation, we have

$$\frac{1}{M}\sum_{k=1}^{M}\sum_{t=1,2} x_{kt}\left[E_m\left[y_{kt}\right] - x_{kt}\beta\right] \to \psi\,(\beta)$$

$$\text{if } \lim \frac{1}{M}\sum_{k=1}^{M}\sum_{t=1,2} x_{kt}^2 = \bar{x}_2 < \infty$$

by the model assumption (1), where :

$$\psi\,(\beta) = \bar{x}_2(\,\beta_0 - \beta)$$

Therefore (i) of Assumption 1 holds. Now clearly $\psi\,(\beta_0) = 0$ and so (ii) also holds. For (iii) to hold, we notice that the derivative of $\psi\,(\beta)$ is $-\,\bar{x}_2$, which is different from zero if at least one of the covariates is.

To verify Assumption 2, we first take the derivative of $\hat{\psi}_N(\beta)$ with respect to β, N ≥ 1. We note that the survey weights do not depend on β and neither does $D_\beta\,\hat{\psi}_N(\beta)$ in this example. Furthermore, if we have design consistency of totals, we conclude that:

(7) $$D_\beta\,[\,\hat{\psi}_N(\beta)] \simeq D_\beta[\,\psi_N(\beta)] = \frac{-1}{M}\sum_{k,t} x_{kt}^{\;2}\,.$$

Note that the right hand side of the equation above is bounded if the covariates are equibounded (there exists a common upper bound for all (k, t)) or if the right hand side converges, i.e. $\bar{x}_2 < \infty$ ■

The proof of the following result was given in the scalar case only (p = 1).

Theorem 1. (Existence and convergence of $\hat{\beta}_N, N \geq 1$). Assume that $\hat{\psi}_N, N \geq 1$ are continuous and that the convergence in (i) of Assumption 1 is uniform in β. Assume further that (ii) and (iii) of Assumption 1 hold. There exist then estimators $\hat{\beta}_N$ such that, for any $\eta, \delta > 0$, there exists $n_0(\eta, \delta) = n_0$ and :

$$\sup_{n \geq n_0} p_N \{ s : | \hat{\beta}_N - \beta_0 | \leq \delta, \ \psi_N(s, \hat{\beta}_N) = 0 \} \geq 1 - \eta$$

The same conclusion holds if Assumptions 1&2 hold.

Proof The proof relies on functional properties of ψ, which is invertible in a neighbourhood of β_0(by (ii), (iii) and the inverse function theorem), and the uniform convergence of the sequence $\hat{\psi}_N(\beta), N \geq 1$. The uniform convergence (in β) of the $\hat{\psi}_N$'s is essential in proving consistency of the estimators. More precisely, it leads to $\psi(\hat{\beta}_N) \to 0 = \psi(\beta_0)$ and, because ψ^{-1} exists and is continuous around β_0, $\psi^{-1}\psi(\hat{\beta}_N) = \hat{\beta}_N \to \psi^{-1}\psi(\beta_0) = \beta_0$. The intermediate value theorem is also used, which makes the proof unsuitable for p > 1.

From the uniform convergence assumption we have that, on a compact set K_0, $\beta_0 \in K_0$, for any pair $\varepsilon, \eta > 0$ there exists $n_0(\varepsilon, \eta) = n_0$ such that :

$$\sup_{N \geq n_0} p_N \{ s : \sup_{\beta \in K_0} | \psi_N(s, \beta) - \psi(\beta) | > \varepsilon \} \leq \eta \tag{8}$$

We now show the existence and tightness of MLE's $\hat{\beta}_N$, $N \geq 1$. More precisely, for any $\eta > 0$, there exist a compact set in the space of parameters, (which contains β_0), say $K_0 = K_\eta(\beta_0)$, $\hat{\beta}_N$, as well as $N_0 = N_0(\eta)$, such that:

(9) $$\sup_{N \geq N_0} p_N \{ s : \hat{\beta}_N \in K_0, \psi_N(s, \hat{\beta}_N, \hat{\alpha}_N, \hat{\phi}_N) = 0 \} \geq 1 - \eta$$

We prove (9) for $\beta \in \mathbb{R}$ and scalar ψ only. This is not just for the sake of simplicity, but also because we use an order relation on the space of parameters to define monotone functions and the estimators $\hat{\beta}_N(s)$, $N \geq 1$ as in (10) below. Let $\hat{\alpha}_N = \alpha_N(s, \hat{\beta}_N, \hat{\varphi}_N)$, $\hat{\varphi}_N = \varphi_N(s, \hat{\beta}_N)$. Now $\psi(\beta_0) = 0$ and $\psi(\beta)$ is smooth at β_0. It is also 1:1, by (ii) and (iii) and the inverse function theorem (see Rudin 1964, Theorem 9.17, p. 193) . There exists therefore β_+ and $\beta_- \in \Theta$, $\beta_+ < \beta_0 < \beta_-$, $\psi(\beta_+) > 0$, $\psi(\beta_-) < 0$ say, and $\psi(\beta)$ is a homeomorphism on $U \supset [\beta_+, \beta_-]$, where U is an open set. By (8), for $\eta > 0$ and for all large $N \geq n_0$ $p_N(\Omega_N) = p_N\{s: \psi_N(s, \beta_+) > 0, \psi_N(s, \beta_-) < 0\} \geq 1 - \eta$. By the continuity of the $\hat{\psi}_N$, on Ω_N, there exists $\hat{\beta}_N'$ such that $\hat{\psi}_N(\hat{\beta}_N') = 0$. Note that we used the intermediate value theorem which restricts our proof to the case p = 1. For every *N*, define the REE:

(10) $$\hat{\beta}_N = \begin{cases} \inf\{\beta \in [\beta_+, \beta_-],\ \hat{\psi}_N(\beta) = 0\} & \textit{if there is such } \beta \\ \beta_N & \text{otherwise} \end{cases}$$

If $K_0 \supseteq [\beta_+, \beta_-]$, we have (9). We can take β_+ , β_- closer to β_0 so $K_0 \supseteq [\beta_+, \beta_-]$, as ψ is strictly decreasing. To complete the proof, note first that:

$$| \hat{\psi}_N(\hat{\beta}_N) - \psi(\hat{\beta}_N) | = | \psi(\hat{\beta}_N) | \leq \varepsilon$$

on a large set and for a large *N*, by the definition of $\hat{\beta}_N \varepsilon K_0$ and (8). Note that ψ is 1:1 and locally a homeomorphism and so its inverse ψ^{-1} exist and is continuous (locally) . Take any $\delta > 0$, $\varepsilon(\delta) = \varepsilon$ small so that $B_\varepsilon(0) \subset V$, $B_\delta(\beta_0) \subset U$ and $y \,\varepsilon\, B_\varepsilon(0) \Rightarrow \psi^{-1}(y) \in B_\delta(\beta_0)$.Thus we have $\psi(\hat{\beta}_N) \,\varepsilon\, B_\varepsilon(0) \Rightarrow \psi^{-1}\psi(\hat{\beta}_N) \,\varepsilon\, \psi^{-1}(B_\varepsilon(0)) \subseteq B_\delta(\beta_0) \Rightarrow \hat{\beta}_N \,\varepsilon\, B_\delta(\beta_0)$, which proves weak consistency.

To prove the last assertion, we have that Assumptions 1&2 hold. Then (8) is the probabilistic equivalent of the classical theorem which states that pointwise convergence to a continuous function on a compact set (Assumption 1) and equicontinuity of a sequence (follows from Assumption 2) imply uniform convergence in β. Thus (8) could follow, for example, from a modification of the proof of b) of Theorem 7.23 on p. 144 of Rudin (1964). ■

Corollary 1 Assume that:

(11) $\hat{\psi}_N(\beta_0) \xrightarrow{P_N} 0$

(12) $D_\beta \ \hat{\psi}_N|_\beta \xrightarrow{P_N}$ D(β), uniformly in $\beta \in \Theta \cap K$, where D(β) is a nonrandom function continuous on K, a compact set containing β_0.

(13) $J_0 = -$ D(β_0) is positive definite and invertible.

We conclude then that there exists a function $\psi(\beta)$, $\beta \in$K, where K is a compact set in Θ, such that (8) (i.e. uniform convergence in probability), holds for $\hat{\psi}_N(\beta)$ and $\psi(\beta)$, as well as for $D_\beta \ \hat{\psi}_N|_\beta$ and $D_\beta \ \psi|_\beta = D(\beta)$. Furthermore, the conclusion of Theorem 1 also holds.

Proof: The fact that { $D_\beta \ \hat{\psi}_N|_\beta$ }, N $\geq$ 1 converges to D(β) uniformly in β on a compact set containing β_0 (i.e. condition (8) for { $D_\beta \ \hat{\psi}_N$ }), follows from (12) and (14) below, as in the proof of Theorem 1.

From Theorem 7.17 on p. 140 of Rudin (1964), we obtain a function $\psi(\beta)$, condition (8) for { $\hat{\psi}_N(\beta)$}, N$\geq$1 and $\psi(\beta)$, $\beta \in K$, as well as the fact that the derivatives { $D_\beta \ \hat{\psi}_N|_\beta$}, N $\geq$ 1 converge in probability to the derivative $D_\beta \psi|_\beta = D(\beta)$. Then (13) ensures that ψ is invertible in a neighbourhood of β_0 and the proof of Theorem 1 follows as before. Note that J_0 in Example 3 (see (7)) is positive . ■

Remark 2: Note that uniform convergence in (12) is implied by pointwise convergence and:

(14) Condition (iv) of Assumption 2 holds for $D_\beta(\hat{\psi}_N)$ rather than $\hat{\psi}_N$.

The conditions of Corollary 1 are stronger than those of Theorem 1. However, some of the conditions could be required in the proof of the Central Limit Theorem anyhow. Note that conditions (12) and (14) imply condition 7 of Binder (1983), p. 291.

In Example 3 the verification of the assumptions was done in two stages, the first based on assumptions of design consistency and the second on Assumptions 1&2 holding for $\psi_N(s,\beta),$, $N \geq 1$. Under this new set of conditions, we also obtain $\beta_N \to \beta_0$ and consequently design consistency:

Corollary 2. If Assumptions 1&2 hold and $\psi_N(s,\beta.)$, $D_\beta\psi_N(s,\beta)$ $N \geq 1$ are design consistent, then $\beta_N \to \beta_0$, $\hat{\beta}_N \xrightarrow{p_N} \beta_0$ and so $\hat{\beta}_N - \beta_N \xrightarrow{p_N} 0$ as $n \to \infty$. Furthermore, the convergence in (i) of Theorem 1 is uniform in $\hat{\beta}_N - \beta_N \xrightarrow{p_N} 0$. ■

Theorem 1 is also valid in the GEE situation. Conditions for Assumptions 1&2 to hold are more complex if V_i, $i \geq 1$ are not known and must be estimated from the sample. In this case, the EE are no longer sums of independent r. v.'s in the design, due to the presence of the estimated correlation structure across time (see the pseudo-likelihood equation (2)).

In order to do statistical inference for GEE, we must find a sample based estimator of V_k = $V_k(\alpha, \beta)$ (see Rao (1998)), and replace it in $U_k(\alpha, \beta, \varphi) = D_k^T V_k^{-1} S_k$, k = 1,...M. This corresponds to the case when $R(\alpha)$ is completely unspecified in Example 5 of Liang and Zeger (1986). In this instance there is no need to estimate the overdispersion parameter φ. To estimate $V_k(\beta)$ $= A_k^{1/2} C_N(\alpha,\beta) A_k^{1/2}$, $k \geq 1$ for fixed values of the parameters, we estimate the common correlation structure across time, denoted here $C_N(\alpha,\beta)$, by $\sum_{k\in s} w_k A_k^{-1/2}(\beta) S_k(\beta) S_k^T(\beta) A_k^{-1/2}(\beta)$. The entries of this matrix are: $\hat{c}_N^{ij}(\beta) = \sum_{k\in s} w_k [a''(\eta_{ki}(\beta)) a''(\eta_{kj}(\beta))]^{-1/2} s_{ki}(\beta) s_{kj}(\beta)$, where $s_{ki}(\beta) = y_{ki} - \mu_{ki}(\beta)$

$k = 1,\ldots M$, $i, j = 1,\ldots d$, $\beta \in \Theta$. Let $\hat{g}_N^{ij}(\beta)$, $i, j = 1, \ldots d$, $\beta \in \Theta$, be the entries of $\hat{C}_N^{-1}(\beta)$, which is assumed to exist. Then $\hat{V}_k^{-1}(\beta)$ has entries $\hat{g}_N^{ij}(\beta)[a''(\eta_{ki}(\beta)) a''(\eta_{kj}(\beta))]^{-1/2}$, $i, j = 1,\ldots d$. We substitute in GEE (5): $U_k(\beta, \hat{V}_k(\beta)) = D_k^T \hat{V}_k^{-1}(\beta) S_k$, $S_k = Y_k - a'(\theta_k)$, for any $k \geq 1$ and obtain:

$$\psi_N(s,\beta) = \sum_{i,j=1,\ldots d} \hat{g}_N^{ij}(\beta)\ \psi_N^{ij}(s,\beta),$$

with

$$\psi_N^{ij}(s,\beta) = \sum_{k\in s} w_k \left[\frac{a''(\eta_{ki}(\beta))}{a''(\eta_{kj}(\beta))}\right]^{1/2} x_{ki} s_{kj}(\beta) \qquad i, j = 1, \ldots d.$$

Therefore, the GEE in (5) can be written as a finite sum of terms with each of these terms equal to a product of two estimators. Furthermore, each $\psi_N^{ij}(s,\beta)$, $i, j = 1,\ldots d$, is a sum of random variables for which the conditions for consistency in Theorem 1 can easily be applied.

<u>**Example 4.**</u> The marginal model is that of Example 1. Recall that $a''(\theta_{kt}) = 1$, $k, t \geq 1$ in this case. An estimator of $V_k(\beta) = V(\beta)$, $k = 1,\ldots M$ is the 2×2 matrix $\hat{C}_N(\beta)$ with entries $\hat{c}_N^{ij}(\beta) = \sum_{k\in s} w_k s_{ki}(\beta) s_{kj}(\beta)$, $i, j = 1,2$, where $s_{ki}(\beta) = y_{ki} - \mu_{ki}(\beta)$, $i = 1, 2$ and $k = 1, \ldots M$. The condition for $\hat{G}_N(\beta) = \hat{C}_N^{-1}(\beta)$ to exist becomes $\Delta(\beta) = \hat{c}_N^{11}(\beta)\hat{c}_N^{22}(\beta) - (\hat{c}_N^{12}(\beta))^2 \neq 0$.

The entries of the matrix $\hat{G}_N(\beta) = \hat{C}_N^{-1}(\beta)$, when it exists, are: $\hat{g}_N^{11}(\beta) = \hat{c}_N^{22}(\beta) \times \Delta(\beta)^{-1}$,

$\hat{g}_N^{12}(\beta) = \hat{g}_N^{21}(\beta) = -\hat{c}_N^{12}(\beta) \times \Delta(\beta)^{-1}$, $\hat{g}_N^{22}(\beta) = \hat{c}_N^{11}(\beta) \times \Delta(\beta)^{-1}$, $\beta \in \Theta$. As above, we have

$$\hat{\psi}_N(\beta) = \sum_{i,j=1,2} \hat{g}_N^{ij}(\beta)\ \hat{\psi}_N^{ij}(\beta), \text{ where } \hat{\psi}_N^{ij}(\beta) = \sum_{k\in s} w_k x_{ki} s_{kj}(\beta) \ \ i, j = 1,2. \blacksquare$$

We assume first that a symmetric, invertible matrix $C(\beta)$ exists and is continuous at β_0 and that:

$$\text{(v)} \qquad \hat{C}_N(\beta) \xrightarrow{p_N} C(\beta),$$

This implies: $\hat{g}_N^{ij}(\beta) \xrightarrow{p_N} g^{ij}(\beta)$, $\forall\ \beta \in \Theta,\ i,j = 1,\ldots d$. Therefore the asymptotic behaviour of $\psi_N(s,\beta)$ is the same as the asymptotic behaviour of $\psi_N^1(s,\beta) = \sum_{i,j=1,\ldots d} g^{ij}(\beta)\ \psi_N^{ij}(s,\beta)$, which can now be written as a linear combination of sample indicators. We can verify the conditions of Theorem 1 for $\psi_N^1(s,\beta)$. However, we often need uniform convergence on a compact space containing β_0 and so we would have to impose conditions stronger than (v). Consider now the information matrices $\hat{J}_N(\beta) = -\dfrac{d\hat{\psi}_N(\beta)}{d\beta}$, N ≥ 1.

__Corollary 3.__ Assume that (12) holds for $\hat{\psi}_N^{ij}(\beta)$ and $\hat{g}_N^{ij}(\beta)$, i, j = 1... d. Then, by the proof of Corollary 1, $\psi_N^{ij}(s,\beta) \xrightarrow{p_N} \psi^{ij}(\beta)$ and $\hat{g}_N^{ij}(\beta) \xrightarrow{p_N} g^{ij}(\beta)$ for some $\psi^{ij}(\beta)$, $g^{ij}(\beta)$ i, j = 1,... d. Let $\psi(\beta) = \sum_{i,j=1,\ldots d} g^{ij}(\beta)\ \psi^{ij}(\beta)$. We assume that $\psi(\beta_0) = 0$ and that $-J_0 = \dfrac{d\psi(\beta_0)}{d\beta} \neq 0$. Then the conclusions of Theorem 1 hold for $\hat{\psi}_N(\beta)$ and $\psi(\beta)$. Furthermore, $\hat{J}_N(\beta)$ are equicontinuous at β_0 and $\hat{J}_N(\beta_0) \xrightarrow{p_N} J_0$. Under the additional conditions $\psi^{ij}(\beta_0) = 0$, i,j = 1,...d, we have that $J_0 = -\sum_{ij} g^{ij}(\beta_0)\dfrac{d\psi^{ij}(\beta_0)}{d\beta} \neq 0$.

__Proof:__ By the proof of Corollary 1, we also have :

$$\psi_N^{ij}(s,\beta) \xrightarrow{p_N} \psi^{ij}(\beta) \text{ and } \hat{g}_N^{ij}(\beta) \xrightarrow{p_N} g^{ij}(\beta) \text{ uniformly in } \beta \text{ so,}$$

$\psi_N(s,\beta) = \sum_{i,j=1,\ldots d} \hat{g}_N^{ij}(\beta)\ \psi_N^{ij}(s,\beta)$ converges uniformly to $\psi(\beta) = \sum_{i,j=1\ldots d} g^{ij}(\beta)\ \psi^{ij}(\beta)$ and $\psi(\beta_0)$

$= 0$. Now : $-\hat{J}_N(\beta) = \dfrac{d\,\hat{\psi}_N(\beta)}{d\beta} = \sum_{i,j=1,\dots,d} \dfrac{d\hat{g}_N^{ij}(\beta)}{d\beta}\,\hat{\psi}_N^{ij}(\beta) + \sum_{i,j=1,\dots,d} \hat{g}_N^{ij}(\beta)\dfrac{d\,\hat{\psi}_N^{ij}(\beta)}{d\beta}$. At β_0, we have convergence to $-J_0 \neq 0$. By the proof of Corollary 1, the conclusions of Theorem 1 hold . The statement about uniform convergence follows from the expression above and the properties of the functions on the right hand side. The last assertion also follows from the expression above. ■

5. The asymptotic variance and the jackknife estimator. We discuss first the asymptotic distribution of $n^{1/2}[\hat{\beta}_N - \beta_N]$, as $n \to \infty$

<u>Theorem 2</u> Assume that conditions of Corollary 3 and the conclusions of Corollary 2 hold . We assume that there exist finite population parameters $g_N^{ij}(\beta_N)$ such that , for i, j = 1,...d:

(vi)
$$\hat{g}_N^{ij}(\beta_N) - g_N^{ij}(\beta_N) \xrightarrow{p_N} 0$$

and that
$$n^{1/2} \sum_{i,j=1,\dots d} \hat{\psi}_N^{ij}(\beta_N) \text{ is } O_{p_N}(1),$$

Then $n^{1/2}[\hat{\beta}_N - \beta_N]$ and $-J_0^{-1} n^{1/2} \sum_{i,j=1,\dots d} g_N^{ij}(\beta_N)\,\hat{\psi}_N^{ij}(\beta_N)$ have the same asymptotic distribution.

Furthermore, if F is the limiting distribution and σ^2 is the limit of the variance $J_0^{-1} n\, V_N[\sum_{i,j=1,..d} g_N^{ij}(\beta_N)\,\hat{\psi}_N^{ij}(\beta_N)]\, J_0^{-1}$ as $n \to \infty$, then σ^2 is the variance associated with F.

<u>Proof</u>: By Corollaries 1 and 2, $\bar{\beta}_N \to \beta_0$, for any $\bar{\beta}_N$ contained in the closed interval defined by $\beta_N, \hat{\beta}_N$ and $\hat{J}_N(\bar{\beta}_N) \to J_0 \neq 0$ in p_N . From Theorem 1, we have that $\psi_N(s, \hat{\beta}_N) = 0$ on a set of large probability and for a large n, and so $\psi_N(s, \hat{\beta}_N) - \psi_N(s, \beta_N) = -\psi_N(s, \beta_N)$ on that set. By the mean value theorem and on the same large set, we have : $-\psi_N(s, \beta_N) = -\hat{J}_N(\bar{\beta}_N)[\hat{\beta}_N - \beta_N]$, or

$n^{1/2}[\hat{\beta}_N - \beta_N] = \hat{J}_N^{-1}(\bar{\beta}_N)\, n^{1/2}\psi_N(s,\beta_N)$. Now $n^{1/2}\psi_N(s,\beta_N) = n^{1/2} \sum_{i,j=1,\ldots d} g_N^{ij}(s,\beta_N)\ \psi_N^{ij}(s,\beta_N)$.

We add and subtract $g_N^{ij}(\beta_N)$, and use (vi) for $\hat{\psi}_N^{ij}(\beta_N)$ and $\hat{g}_N^{ij}(\beta_N)$ to obtain that $n^{1/2}[\hat{\beta}_N - \beta_N] \approx J_0^{-1} n^{1/2} \sum_{i,j=1,\ldots d} g_N^{ij}(\beta_N)\ \psi_N^{ij}(s,\beta_N)$ in p_N. The statement about the variance

follows from (viii - B) p. 121 of C.R. Rao (1973). ■

Condition (v) appears in Shao (1999), p. 315. Notice that it is stronger than condition (vi) above. It is, however, not sufficient to ensure the asymptotic normality of $n^{-1/2}[\hat{\beta}_N - \beta_N]$. Note that $-J_0$ is 'the bread of the sandwich' in the asymptotic variance of $n^{-1/2}[\hat{\beta}_N - \beta_N]$.

We denote by $u_{i\ell}(s,\beta)$ the ℓ th component of U_i above, $\ell = 1,..p$, $i = 1,\ldots M$. For the rest of the section, we assume equicontinuity of the components of U_i, $i = 1, \ldots M$ (see Shao (1999), p. 318 and Theorem 3 below) i.e. we require:

(E) The family of functions $\{u_{i,\ell}(\beta)\}_{i,\ell}$ is equicontinuous in β at β_0.

Note that this condition is related to the previous conditions (e.g. in Corollary 3) once we take the normalized weights into account. Such assumptions appear elsewhere in the technical literature in connection with asymptotic results for GEE or IEE (e.g. Theorem 5.14 of Shao (1999) and Rubin - Bleuer (1998) when p =1). In many interesting instances, condition (E) is implied by continuity of functions of the covariates and boundedness of the covariates (e.g. equation (3) in Example 1).

We consider now the one - step jackknife estimator with the design described in section 3. From the sample of n clusters, let us delete cluster i, which we assume belongs to stratum h. All individual weights in each of the remaining n_h - 1 clusters in stratum h are multiplied by the factor $n_h/(n_h - 1)$ to compensate for the deletion of one cluster. The weights in other strata are left unchanged. Of course, all individual weights in cluster i are set to 0. Let $w_k^*{}_k$ be the new weights, $k \in s$. The estimator that corresponds to $\hat{\psi}_N(\beta)$ in (5) will be denoted, for simplicity, by $\hat{\psi}_{-i}(\beta)$. More precisely, we have, for each cluster $i \in s$:

(15) $$\hat{\psi}_{-i}(\beta) = \sum_{k \varepsilon s \backslash i} w_k^* U_k(\beta, \hat{\alpha}(\beta), \hat{\varphi}(\beta))$$

We introduce, for cluster $i \in s$, the p×p "information" matrices $\hat{J}_{-i}(\beta)$ along with the p - component vectors $\hat{\beta}_{-i}$, with $\hat{\beta}_N$ as in Definition 2:

(16) $$\hat{J}_{-i}(\beta) = -\frac{\partial \hat{\psi}_{-i}(\beta)}{\partial \beta}, \quad \hat{\beta}_{-i} = \hat{\beta}_N + \hat{\psi}_{-i}(\hat{\beta}_N) \times \hat{J}_{-i}^{-1}(\hat{\beta}_N), \; i \in s,$$

Note that in (16) one must state conditions for the definitions to be valid (i.e. the inverse matrices, should exist at least asymptotically). We can now define the jackknife estimator $v_J(\hat{\beta})$. For simplicity of notation, we write $\hat{\beta}$ for $\hat{\beta}_N$ when no confusion may arise.

(17) $$v_J(\hat{\beta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{\beta}_{-i} - \hat{\beta})(\hat{\beta}_{-i} - \hat{\beta})^T$$

For the remainder of the section we assume that $\hat{\psi}_N(\hat{\beta}) = 0$, N = 1,...The main result presented in this section is the consistency of the jackknife estimator of $V_N(\hat{\beta}_N)$, the variance of $\hat{\beta}_N$, i.e.:

(18) $$n[v_J(\hat{\beta}_N) - V_N(\hat{\beta}_N)] \xrightarrow{p_N} 0 \text{ as } n \to \infty$$

We consider the following condition (see Krewski and Rao, 1981):

(19) $$\max_i |\hat{J}_{-i}(\hat{\beta}_N) - J_0| \xrightarrow{p_N} 0 \text{ as } n \to \infty$$

We state condition C4 of Yung (1996), which appears in Shao and Tu (1995):

(C4) $$n \max_{hik} m_{hi} w_{hik} = O_n(1)$$

This condition can be interpreted in terms of the weights associated with the sample design. It requires that the design weights be comparable in size. We will state the main result of this article, for the IEE and the case p = 1 only. Let $u_{hi}(\beta) = \sum_{k \in s} w_{hik} u_{hik}(\beta)$ be an unbiased estimator of the mean of cluster i in stratum h (recall that the weights are normalized), where $i = 1, \ldots N_h$, h = 1, ...L. Note that $u_{hi}(\beta)$ are independent random variables, $i = 1, \ldots N_h$, h = 1, ...L, and that

$\hat{\psi}_N(\beta) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} u_{hi}(\beta)$. Condition (C1) below is a Lyapounov type condition. Both (C1) and (C2)

are typical conditions required in the proof of CLT.

Theorem 3 Assume that the conditions of Corollary 1 hold and that $\beta_N \rightarrow \beta_0$, so $\hat{\beta}_N \xrightarrow{p_N} \beta_0$ and $\hat{\beta}_N - \beta_N \xrightarrow{p_N} 0$ as $n \rightarrow \infty$. We also assume (19), (C4) and:

(C1) $$n^{1+\delta} \sum_{hi} E|u_{hi}(\beta_0) - Eu_{hi}(\beta_0)|^{2+\delta} = O_n(1), \text{ for some } \delta > 0.$$

(C2) $$n V_N(\hat{\psi}_N(\beta_0)) \longrightarrow \sigma^2 > 0.$$

(E) $$\max_{hik} |u_{hik}(\beta) - u_{hik}(\beta_0)| \le \varepsilon \quad \text{if} \quad |\beta - \beta_0| \le \delta(\varepsilon), \text{ for some } \delta(\varepsilon) > 0.$$

Then the jackknife estimator is consistent as in (18).

Proof: Write $v_J(\hat{\psi})(\beta) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i \in s_h} \hat{\psi}^2_{-i}(\beta)$, where s_h represents the sample of clusters in stratum h. We have :

(20) $$n v_J(\hat{\beta}) = n \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i \in s_h} [\hat{J}_{-i}^{-2} - J_0^{-2}] \hat{\psi}^2_{-i}(\hat{\beta}) + J_0^{-2} n\, v_J(\hat{\psi})(\hat{\beta})$$

We will show :

(21) $$n[v_J(\hat{\psi})(\hat{\beta}) - v_L(\hat{\psi})(\beta_0)] \xrightarrow{p_N} 0$$

where $v_L(\hat{\psi})(\beta) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i \in s_h} [\bar{u}_h(\beta) - u_{hi}(\beta)]^2$, $\bar{u}_h(\beta) = \frac{\sum_{i \in s_h} u_{hi}(\beta)}{n_h}$

is the usual estimator of the variance of $\hat{\psi}_N(\beta) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} u_{hi}(\beta)$. As in Yung (1996), (C1) and (C2) imply that $nv_L(\hat{\psi})(\beta_0) \xrightarrow{p_N} \sigma^2$. Let us assume (21) for now. Then $nv_J(\hat{\psi})(\hat{\beta}) \xrightarrow{p_N} \sigma^2$, the first term of (20) converges to 0 by (19) and the second converges to $J_0^{-2}\sigma^2$ thus:

(22) $$nv_J(\hat{\beta}) \xrightarrow{p_N} J_0^{-2}\sigma^2$$

On the other hand, it can be shown (using (C2) and condition (E), as in the Appendix) that :

$$n[V_N(\hat{\psi}_N(\beta_N)) - V_N(\hat{\psi}_N(\beta_0))] \longrightarrow 0.$$

Therefore, by (C2), the equicontinuity of the information matrices and the proof of Theorem 2,

$$nV_N(\hat{\beta}) = n J_N^{-2}(\bar{\beta}_N)\ V_N(\hat{\psi}_N(\beta_N)) \longrightarrow J_0^{-2}\sigma^2,$$

which is the limit in (22). To complete the proof of (18), we need to prove (21). After some algebraic manipulations, we have that, for i $\geq$ 1, $\hat{\psi}_{-i}(\hat{\beta}) = n_h/(n_h-1)[\bar{u}_h(\hat{\beta}) - u_{hi}(\hat{\beta})]$ and so $nv_J(\hat{\psi})(\hat{\beta}) = nv_L(\hat{\psi})(\hat{\beta})$. We used the fact that $\hat{\psi}_N(\hat{\beta}) = 0$. For the proof of (21), we recall

$v_L(\hat{\psi})(\beta) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i \in s_h} [\bar{u}_h(\beta) - u_{hi}(\beta)]^2$, and define $\varepsilon_{hi} = \sum_{k \in s_{hi}} w_{hik}[u_{hik}(\hat{\beta}) - u_{hik}(\beta_0)]$,

$\bar{\varepsilon}_h = \frac{\sum_{i \in s_h} \varepsilon_{hi}}{n_h}$, where s_{hi} is the second stage sample selected in cluster i of stratum h. Then

$\bar{u}_h(\hat{\beta}) - u_{hi}(\hat{\beta}) = \bar{\varepsilon}_h - \varepsilon_{hi} + \bar{u}_h(\beta_0) - u_{hi}(\beta_0)$ and

$[\bar{u}_h(\hat{\beta}) - u_{hi}(\hat{\beta})]^2 = [\bar{\varepsilon}_h - \varepsilon_{hi}]^2 + 2[\bar{\varepsilon}_h - \varepsilon_{hi}][\bar{u}_h(\beta_0) - u_{hi}(\beta_0)] + [\bar{u}_h(\beta_0) - u_{hi}(\beta_0)]^2$. The contribution of the last term is $n v_L(\hat{\psi}(\beta_0))$, which is what is needed. We must show that the contribution of the first terms is 0 asymptotically. We will first deal with the cross terms.

They contribute $2\, n\sum_{h=1}^{L} \frac{n_h}{n_h-1}\sum_{i\in s_h} [\bar{\varepsilon}_h - \varepsilon_{hi}][\bar{u}_h(\beta_0) - u_{hi}(\beta_0)]$, and by Schwarz inequality, this is bounded by $2[n\sum_{h=1}^{L} \frac{n_h}{n_h-1}\sum_{i\in s_h} [\bar{\varepsilon}_h - \varepsilon_{hi}]^2]^{1/2}\ [n\ v_L(\hat{\psi})(\beta_0)]^{1/2}$. It suffices to show that the contribution of the first term is asymptotically 0, i.e. :

$$(23) \qquad n\sum_{h=1}^{L} \frac{n_h}{n_h-1}\sum_{i\in s_h} [\bar{\varepsilon}_h - \varepsilon_{hi}]^2 \xrightarrow{p_N} 0$$

Since $[\bar{\varepsilon}_h - \varepsilon_{hi}]^2 \le 2(\bar{\varepsilon}_h^2 + \varepsilon_{hi}^2) \le 2[\sum_{i\in s_h} \varepsilon_{hi}^2/n_h + \varepsilon_{hi}^2]$,

we have $[\bar{\varepsilon}_h - \varepsilon_{hi}]^2 \le 4\max_{hik}\varepsilon_{hi}^2 \le 4\,\varepsilon^2\max_{hik}(w_{hik}m_{hi})^2$ by (E) when $\hat{\beta}$ is close to β_0. Since $n_h \ge 1 \Leftrightarrow n_h/(n_h-1) \le 2$, we have that the left hand side of (23) is bounded by $\le 8\,\varepsilon^2\max_{hik}(n\,w_{hik}m_{hi})^2$ which can be made as small as possible by (C4). Recall that $\sum_{h=1}^{L}\sum_{i\in s_h} 1 = \sum_{h=1} n_h = n$, the total number of clusters selected. ■

6. Conclusions. Design inference is a useful, interesting and challenging subject. Inference is generally more difficult in finite populations than in infinite populations. In the finite population situation, we have to deal with 2 levels for each of the main and 'nuisance' parameters. Many of the techniques that are used in classical inference can be adapted to the context of survey randomization. However, 'regularity conditions' that involve the interchange of derivatives and expectations taken with respect to the superpopulation model must be replaced by functional conditions. We tried to reduce the model assumptions to a minimum. As in Rao (1998), we retained the first moment model assumption in (1). Even though convergence of census parameters (including population averages in Example 3) can be treated as limits of functions, it is more natural to view them as realizations of sums of r.v.'s, as indicated in Example 3. Therefore, it appears more natural to view design - inference within the more general set-up presented in Rubin-Bleuer (1998), which allows for joint model and design-based inference.

Acknowledgement. The author would like to thank Professor J.N.K. Rao for suggesting the research topic and for his continuous assistance and encouragement. She would also want to thank Emmanuel Benhin for reading the manuscript and suggesting some changes.

APPENDIX

To complete the proof of Theorem 3, we must show that $n\,[V_N(\hat{\psi}_N(\beta_N)) - V_N(\hat{\psi}_N(\beta_0))] \longrightarrow 0$,

We write $\hat{\psi}_N(\beta_N) = \Delta_N + \hat{\psi}_N(\beta_0) - E\hat{\psi}_N(\beta_0) = \Delta_N + Z_N$, where $\Delta_N = \hat{\psi}_N(\beta_N) - \hat{\psi}_N(\beta_0) + E\hat{\psi}_N(\beta_0)$. Notice that , by the definition of the census parameter in (4), Δ_N is centered and so is $\hat{\psi}_N(\beta_N)$. Thus $V_N(\hat{\psi}_N(\beta_N)) = E(\Delta_N^2) + EZ_N^2 + 2E[\Delta_N Z_N]$, where $V_N(\hat{\psi}_N(\beta_0)) = EZ_N^2$. Now $nEZ_N^2 \longrightarrow \sigma^2$ by (C2) and , applying Schwarz inequality to the cross term we reduce the problem to showing that $nE\Delta_N^2 = nV_N(\Delta_N) \longrightarrow 0$. We calculate the design variance of Δ_N, which equals $V_N[\sum_{h=1}^{L}\sum_{i\in s_h}[u_{hi}(\beta_N) - u_{hi}(\beta_0)]]$. Because of the independent selection in each stratum, this variance is $\sum_{h=1}^{L} V_N(\sum_{i\in s_h}[u_{hi}(\beta_N) - u_{hi}(\beta_0)])$. We can write: $\sum_{i\in s_h} u_{hi}(\beta_N) - u_{hi}(\beta_0) = (1/n_h)\sum_{j=1}^{n_h}\sum_{i=1}^{N_h}\hat{\delta}_{hi} I_i(r_{hj})$, where $\hat{\delta}_{hi} = n_h \sum_{k\in s_{hi}} w_{hik}[u_{hik}(\beta_N) - u_{hik}(\beta_0)]$, I_i is the sample indicator of cluster i, r_{hj} is the sample selected in stratum h at the j th independent draw. Recall that s_{hi} is the second stage sample selected in cluster i of stratum h. By the independence due to the "with replacement" selection, we have $V_N[(1/n_h)\sum_{j=1}^{n_h}\sum_{i=1}^{N_h}\hat{\delta}_{hi} I_i(r_{hj})] = n_h^{-2}\sum_{j=1}^{n_h} V_N(\sum_{i=1}^{N_h}\hat{\delta}_{hi} I_i(r_{hj})) \le n_h^{-2}\sum_{j=1}^{n_h} E[(\sum_{i=1}^{N_h}\hat{\delta}_{hi} I_i(r_{hj}))]^2$. Since in each stratum only one cluster is selected at each draw, the expectation above equals $\sum_{i=1}^{N_h} E[\hat{\delta}_{hi}^2\; I_i(r_{hj}))]$.

Now $\hat{\delta}_{hi}^2 \le n_h^2\delta^2[\max_{hik} w_{hik} m_{hi}]^2$ if β_N is close to β_0, by condition (E). Thus,

$nV_N(\Delta_N) \le n\delta^2[\max_{hik} w_{hik} m_{hi}]^2 \sum_h n_h \sum_{i=1}^{N_h} E[I_i(r_{hl})] = n\delta^2[\max_{hik} w_{hik} m_{hi}]^2 \sum_h n_h$, which can be made arbitrarily small by (C4), as δ can be made arbitrarily small. We used the fact that the probability of selecting at each draw one cluster from each stratum is 1. ■

REFERENCES

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

Binder , D.A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys, *JASA*, **89**, 1035- 1043.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Stat.* **13**, 342 - 368.

Fuller W.A. (1975). Regression analysis for sample surveys. *Sankhya* **(C)**. **37**, 117-132.

Hájek, J. (1960). Limiting distributions in simple random sampling from finite populations, *Publ. Math. Inst. Hungarian Acad. Sci.*, **5**, 361 - 374.

Krewski, D. and Rao, J.N.K.(1981). Inference from stratified samples: Properties of linearization, jackknife and balanced repeated replication methods, *Ann. Stat.* **9**, 1010-1019.

Liang, K-Y and Zeger, S. L.(1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1990). Using the jackknife to estimate the variance of regression estimators from repeated measurement studies, *Communications in Statistics* 1990, 821-844.

McCullagh, P. and Nelder, J.A.(1989), *Generalized Linear Models*, second edition, Chapman& Hall, London.

Rao, C. R. (1974). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rao, J.N.K. (1998). Marginal models for repeated observations: inference with survey data. *Proceedings of the ASA*, 1998.

Rudin,W(1964). *Principles of Mathematical Analysis*, second edition, Mc Graw-Hill, New York.

Rubin-Bleuer, S.(1998). Sabbatical Research Report, Statistics Canada.

Särndal, C-E, Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling,* Springer-Verlag, New York.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shao, J., Tu, D.(1995) . *The Jackknife and Bootstrap*, Springer Verlag, New York.

Shao, J.(1999). *Mathematical Statistics*, Springer Verlag, New York.

Wu, C.F.J.(1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Stat.* **9**, 501-513.

Yung, W.(1996). Contributions to poststratification in stratified multi-stage samples. Ph. D. thesis, Carleton University, Ottawa, Canada.

c. 3