The Canadian Journal of Statistics
Vol. 38, No. 4, 2010, Pages 540–554
La revue canadienne de statistique

The pseudo-GEE approach to the analysis of longitudinal surveys

Iván A. CARRILLO[1]*, Jiahua CHEN[2] and Changbao WU[3]

[1] *Statistics Canada, Social Survey Methods Division, Tunney's Pasture, R.H. Coats Building, 15th Floor, Ottawa, Ontario, Canada K1A 0T6*
[2] *Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2*
[3] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*

Key words: Complex sampling design; consistency; design-based inference; generalized estimating equations; joint randomization; superpopulation model; variance estimation.

MSC 2000: Primary 62D05; secondary 62G05.

Abstract: Longitudinal surveys have emerged in recent years as an important data collection tool for population studies where the primary interest is to examine population changes over time at the individual level. Longitudinal data are often analyzed through the generalized estimating equations (GEE) approach. The vast majority of existing literature on the GEE method; however, is developed under non-survey settings and are inappropriate for data collected through complex sampling designs. In this paper the authors develop a pseudo-GEE approach for the analysis of survey data. They show that survey weights must and can be appropriately accounted in the GEE method under a joint randomization framework. The consistency of the resulting pseudo-GEE estimators is established under the proposed framework. Linearization variance estimators are developed for the pseudo-GEE estimators when the finite population sampling fractions are small or negligible, a scenario often held for large-scale surveys. Finite sample performances of the proposed estimators are investigated through an extensive simulation study using data from the National Longitudinal Survey of Children and Youth. The results show that the pseudo-GEE estimators and the linearization variance estimators perform well under several sampling designs and for both continuous and binary responses. *The Canadian Journal of Statistics* 38: 540–554; 2010 © 2010 Statistical Society of Canada

Résumé: Les enquêtes longitudinales sont apparues récemment comme un moyen important de collecte d'observations pour les études sur les populations dont nous voulons étudier les changements dans le temps de la population au niveau des individus. Les données longitudinales sont souvent analysées en utilisant les équations d'estimation généralisées (GEE). Cependant, une grande partie des articles publiés sur la méthode GEE sont développés dans un cadre non échantillonnal et ils ne sont pas appropriés pour des données obtenues par des devis échantillonnaux complexes. Dans cet article, les auteurs développent une approche pseudo-GEE pour l'analyse des données d'enquête. Ils démontrent que les poids d'échantillonnage doivent et peuvent être considérés de façon adéquate dans une méthode GEE dans un contexte d'aléation conjointe. La convergence des estimateurs pseudo-GEE ainsi obtenus est démontrée dans le cadre proposé. Les estimateurs de la variance par linéarisation sont développés pour les estimateurs pseudo-GEE lorsque le taux d'échantillonnage de la population fini est petit ou négligeable, scénario qui se produit souvent pour les enquêtes de grande envergure. La performance pour de petits échantillons des estimateurs proposés est étudiée grâce à une étude de simulation exhaustive en utilisant des données provenant de l'enquête longitudinale nationale sur les enfants et les jeunes. Les résultats indiquent que les estimateurs pseudo-GEE et ceux de la variance par linéarisation se comportent très bien sous plusieurs devis échantillonnaux, et ce tant pour les réponses continues que binaires. *La revue canadienne de statistique* 38: 540–554; 2010 © 2010 Société statistique du Canada

* *Author to whom correspondence may be addressed.*
E-mail: ivan.carrillogarcia@statcan.gc.ca

1. INTRODUCTION

There exist two major types of statistical research designs, namely, cross-sectional studies and longitudinal studies. Cross-sectional studies can be described as "one-time" or "one-shot" studies where interest lies in the characteristics of a certain population or model at a particular time point. In longitudinal studies, also called "panel studies," variables of interest are measured on a fixed set of units at several time points during a reference time period. One of the major advantages of longitudinal studies is that they allow for the measurement of time-varying explanatory variables and hence for the exploration of population changes at the individual level. Some of these time-varying explanatory variables can be natural ones, such as age, and some could be specifically designed, such as different treatments or population interventions before and after certain time point. With longitudinal studies it is possible to separate age and cohort effects (Diggle et al., 2002; Hedeker & Gibbons, 2006) or the effect of treatments and interventions from other potential confounders. One of the major challenges of longitudinal studies is the added complexity in data analysis, due to the lack of independence among responses measured from the same unit. Other problems with longitudinal studies include (i) changes of population composition over time (Duncan & Kalton, 1987); (ii) changes in measurement instruments over time (Kish, 1987); and (iii) complexities in missing value problems.

Longitudinal data are often analyzed through the generalized estimating equations (GEE) approach. The vast majority of existing literature on the GEE method; however, is developed under non-survey settings. Liang & Zeger (1986) described the GEE methodology for analyzing longitudinal data; Yuan & Jennrich (1998) and Shao (2003), among others, studied asymptotic properties of the GEE estimators. These results are not directly applicable to complex survey data. The use of survey weights under the estimating equation approach has been examined by several authors, including Godambe & Thompson (1986), Binder & Patak (1994), Godambe (1995), Rao (1998), and Roberts, Ren & Rao (2009), among others. The consistency of the resulting estimators; however, has not been formally established in these earlier investigations.

In this paper we develop a pseudo-GEE approach for the analysis of survey data. We show that survey weights must and can be appropriately accounted in the GEE method under a joint randomization framework. The consistency of the resulting pseudo-GEE estimators is established under the proposed framework. Linearization variance estimators are developed for the pseudo-GEE estimators when the finite population sampling fractions are small or negligible, a scenario often held for large-scale surveys. We illustrate the method and examine finite sample performances of the pseudo-GEE estimators through a simulation study using data from the National Longitudinal Survey of Children and Youth (NLSCY). This survey is designed by Human Resources Development Canada to study child development and well-being. Data from five biennial cycles of the survey conducted from 1994 to 2003 are now available through Statistics Canada's Research Data Centers. One of the main objectives of the survey is to study children's behavioural problems as they grow and identify influential factors. The task is well suited for the pseudo-GEE approach. A key variable in NLSCY data sets is physical aggression score (PAS), derived based on six to eight questions (depending on the age group) included in the survey. Earlier studies (Thomas, 2004; Carrillo et al., 2005; Carrillo-García, 2006) identified several significant factors to the PAS. In this paper, we do not repeat these analyses. Instead, we use their results to construct credible superpopulation models and generate finite populations for our simulation study. The simulation shows that the pseudo-GEE estimators and the linearization variance estimators perform well under several sampling designs and for both continuous and binary responses.

The rest of the paper is organized as follows. In Section 2, we develop a joint randomization framework as the foundation for statistical analysis of complex survey data. In Section 3 we present the pseudo-GEE estimator and establish its consistency under the joint randomization framework. In Section 4 we derive linearization variance estimators for the pseudo-GEE estimators. Results from an extensive simulation study on finite sample performances of the pseudo-GEE estimators are reported in Section 5. Some concluding remarks are given in Section 6.

2. THE JOINT RANDOMIZATION FRAMEWORK FOR COMPLEX SURVEYS

There are three prevailing frameworks for the statistical analysis of complex survey data. In the pure "model-based" approach the parameters of interest are related to a statistical model, often referred to as a superpopulation model. Under this setting the sampling design features are ignored and sampled individuals are treated as independent observations. All inferences are carried out and evaluated with respect to the model. In the conventional "design-based" approach, the parameters of interest are finite population quantities. Values attached to variables of interest are viewed as non-random quantities and inferential procedures are evaluated with respect to the random mechanism induced by the probability selection of sampled units. Design-based inferences focus on the particular finite population from which the sample is taken and their validity does not depend on any model assumptions. In recent years there has been another popular way of inference called "model-assisted" approach. Here again, interest lies exclusively on finite population parameters and all observed quantities are regarded as non-random. Inferential procedures are judged with respect to the probability sampling design. However, these procedures and associated estimators are motivated through an assumed model. Model-assisted approach is essentially design-based but it can have increased efficiency when the finite population is well described by the assumed model.

The two sources of randomization, namely, the probability sampling design for a finite population and the assumed model for a superpopulation, can be jointly considered, resulting in the so-called "joint randomization" inference. Under this framework, the finite population is regarded as a random sample from the superpopulation model and the survey sample is viewed as second phase sampling from the original superpopulation (Binder & Roberts, 2003). The framework is well suited for analytic use of survey data where, for instance, one is interested in possible causal relationships described by the superpopulation model. A preferred inference would be based on the whole data from the entire finite population. The actual survey sample taken from the finite population is usually obtained through a complex sampling design involving stratification, clustering and/or multi-stage unequal probability selection. The superpopulation model can therefore be distorted in the sampled data and becomes invalid. For instance, the GEE model described in the next section assumes that observations from different subjects are independent. This assumption, however, is typically invalid for complex survey samples. Under such scenarios the survey design features cannot be ignored and inferences for the superpopulation parameters need to be carried out by combining both randomization processes.

Inferences under the joint randomization framework may be preferred even if all model assumptions are valid for the survey sample. It provides certain protection against model failure and inferences are valid regardless whether the design features can be ignored or not. Pure model-based methods can be severely affected by things like excluding important variables or interaction terms. Whereas in such situations the inclusion of the sampling design features yields "the best fit of that model for the surveyed population, and hence also a good fit for similar populations where 'similar' relates to the excluded variables" (Kalton, 1983). Another justification for a joint randomization approach is that, even under design-based approach, certain optimality criteria necessarily rely on models (Wu, 2003). There are other scenarios where joint randomization is the only appropriate framework for inference because of the way in which the data are collected (Chen, Thompson & Wu, 2004).

3. THE PSEUDO-GEE METHOD FOR LONGITUDINAL SURVEYS

In this section, we propose a pseudo-GEE approach to longitudinal surveys under the joint randomization framework. We assume that there is a sequence of finite populations, indexed by ν. For a given ν, the finite population is a random sample from a superpopulation model ξ with population size N_ν. Furthermore, a sample of size n_ν is taken from the finite population according to a probability sampling design. As $\nu \to \infty$, both $N_\nu \to \infty$ and $n_\nu \to \infty$. We also assume that the superpopulation model ξ remain the same as $\nu \to \infty$. Hence, the superpopulation parameters, β, ϕ, and α given below, also remain fixed. For notational simplicity, we will drop the index ν.

Let $\{1, 2, \ldots, N\}$ be the set of labels for the N subjects in the finite population. Let $(Y_{ij}; X_{ij1}, \ldots, X_{ijp})'$ be values of the response variable Y and the vector of p covariates $(X_1, \ldots, X_p)'$ for the ith subject at the time of the jth cycle of the survey, $j = 1, \ldots, T_i$. The T_i can be different for different subjects but in many studies $T_i = T$ is common for all subjects. This is typically the case for large-scale surveys. Let $X_{ij} = (1, X_{ij1}, \ldots, X_{ijp})'$ and $X_i = (X'_{i1}, \ldots, X'_{iT_i})'$. We assume that the superpopulation model ξ can be characterized by the following three components:

(1) The conditional mean response $\mu_{ij} = E(Y_{ij} \mid X_{ij})$ is related to the linear predictor $\eta_{ij} = X'_{ij}\boldsymbol{\beta}$ through a monotone link function $g(\cdot)$: $\mu_{ij} = g^{-1}(\eta_{ij}) = g^{-1}(X'_{ij}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$.
(2) The conditional variance of Y_{ij} given X_{ij} is given by $\mathrm{Var}(Y_{ij} \mid X_{ij}) = \phi\upsilon(\mu_{ij})$, where $\upsilon(\cdot)$ is the variance function with known form and $\phi > 0$ is called a dispersion parameter.
(3) The conditional covariance matrix of $Y_i = (Y_{i1}, \ldots, Y_{iT_i})'$ is given by $\mathrm{Cov}(Y_i \mid X_i) = A_i^{1/2}\mathbf{R}_i(\alpha)A_i^{1/2}$, where $A_i = \mathrm{diag}\{\phi\upsilon(\mu_{i1}), \ldots, \phi\upsilon(\mu_{iT_i})\}$ and $\mathbf{R}_i(\alpha)$ is the correlation matrix with a specified structure involving parameter α.

Note that the assumption that the finite population is a random sample from the superpopulation also implies that

(4) The response vectors Y_k and Y_l given X_k and X_l are independent for $k \neq l$.

Among the four components described above, items 1, 2, and 4 are similar to those for the generalized linear models (GLM). However, there are two important and unique features in the GEE model specifications for longitudinal data which are not part of GLM. Firstly, it is possible to include time-dependent covariates in ξ to explore changes over time. Such variables can be as simple as age or variables by specific design features of the study. This allows the examination of the effectiveness of, say, population interventions before and after certain time point while controlling other factors in the study. Secondly, "it is the (third) component, the incorporation of the within-subject association among the repeated responses from the same individual, that represents the main extension of GLM to longitudinal data" (Fitzmaurice, Laird & Ware, 2004). For estimation procedures described below, we use $\Sigma_i = \mathrm{Cov}(Y_i \mid X_i)$ to denote the true variance–covariance matrix but use V_i to represent the so-called working variance–covariance matrix. In other words, $V_i = A_i^{1/2}\mathbf{R}_i(\alpha)A_i^{1/2}$ when $\mathbf{R}_i(\alpha)$ is a chosen working correlation matrix which does not necessarily coincide with the true one.

Following the GEE methodology as described in Liang & Zeger (1986), we can define the so-called "census GEE estimator" of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_N$, as the solution to the following set of estimating equations:

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}'_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{1}$$

Here $\boldsymbol{y}_i$ denotes the observed value of Y_i and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT_i})'$. The census estimator $\boldsymbol{\beta}_N$ has no practical value but serves as an important reference point for theoretical development on the pseudo-GEE estimator defined below.

Let s be the set of n units selected from the finite population by a complex sampling design; let $w_i = 1/P(i \in s)$ be the basic design weights; let $\{(Y_{ij}; X_{ij1}, \ldots, X_{ijp}),\ j = 1, \ldots, T_i,\ i \in s\}$ be the data set from the longitudinal survey. If we treat the left-hand side of Equation (1) as a finite population total, we can estimate it based on the survey sample s using the well-known Horvitz–Thompson estimator (Horvitz & Thompson, 1952). Our proposed sample-based pseudo-GEE estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_n$, is defined as the solution to the following set of estimating equations:

$$\sum_{i \in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{2}$$

The solution $\hat{\boldsymbol{\beta}}_n$ to (2) can be computed through the following iterative steps from a Newton–Raphson type procedure:

$$\boldsymbol{\beta}_{(l+1)} = \boldsymbol{\beta}_{(l)} + \left[\sum_{i \in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_{(l)}} \right]^{-1} \left[\sum_{i \in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i(l)}) \right].$$

The dispersion parameter ϕ can be estimated by

$$\hat{\phi} = \frac{\sum_{i \in s} w_i \sum_{j=1}^{T_i} e_{ij}^2}{(\sum_{i \in s} w_i T_i) - p} = \frac{\sum_{i \in s} w_i \sum_{j=1}^{T_i} (y_{ij} - \hat{\mu}_{ij})^2 / \upsilon(\hat{\mu}_{ij})}{(\sum_{i \in s} w_i T_i) - p}. \tag{3}$$

If the within-subject correlation structure is unspecified but is assumed to be constant across subjects, we can estimate the correlation matrix $\mathbf{R} = (\alpha_{jk})$ by

$$\hat{\alpha}_{jk} = \frac{\sum_{i \in s} w_i e_{ij} e_{ik}}{[(\sum_{i \in s} w_i) - p]\hat{\phi}} = \frac{\sum_{i \in s} w_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik}) / \{\upsilon(\hat{\mu}_{ij})\upsilon(\hat{\mu}_{ik})\}^{1/2}}{[(\sum_{i \in s} w_i) - p]\hat{\phi}}, \tag{4}$$

where the standardized residuals are given by $e_{ij} = (y_{ij} - \hat{\mu}_{ij}) / \{\upsilon(\hat{\mu}_{ij})\}^{1/2}$.

Note that $\hat{N} = \sum_{i \in s} w_i$ and $\sum_{i \in s} w_i T_i$ are used in these formulas as opposed to n and $\sum_{i=1}^{n} T_i$ one would use from expressions for non-survey data. Because of this, the usual GEE software procedures like `gee` in R or `genmod` in SAS are not recommended for calculating the pseudo GEE estimators for survey data. Even if one specifies the `weight` variable as the survey weights w_i, these procedures do not always carry out the appropriate modification of $\hat{\phi}$ and $\hat{\alpha}$.

We now establish the consistency of the pseudo-GEE estimator $\hat{\boldsymbol{\beta}}_n$ under the joint randomization of the model ξ and the sampling design π. All expectations and variances, whether with respect to the model or the design, are conditional on the given covariates. For notational simplicity, we drop these conditions in following discussions. For instance, $E_\xi[h_i^2(Y_i)]$ is a short form for $E_\xi[h_i^2(Y_i) \mid X_i]$. The following theorem is stated in terms of a more general $\psi_i(Y_i, \boldsymbol{\beta})$ than the specific form $(\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ in the definition of $\hat{\boldsymbol{\beta}}_n$.

Theorem 1. *Let $s_n(\boldsymbol{\beta}) = \sum_{i \in s} w_i \psi_i(Y_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \Theta \subset \mathbb{R}^p$ and $\psi_i(Y_i, \boldsymbol{\beta})$ is a function from $\mathbb{R}^{T_i} \times \Theta$ to $\mathbb{R}^p$; let $\boldsymbol{\beta}_0 \in \Theta$ be such that $E_{\xi\pi}[s_n(\boldsymbol{\beta}_0)] = 0$; let $h_i(Y_i) = \sup_{\boldsymbol{\beta} \in \Theta} \|\psi_i(Y_i, \boldsymbol{\beta})\|$, $i = 1, 2, \ldots$, where $\| \cdot \|$ is the usual $\mathcal{L}_1$ norm. Suppose that*

(1) $\sup_i E_\xi[h_i^2(Y_i)] < \infty$ *and* $\sup_i E_\xi\|Y_i\| < \infty$*;*

(2) For any $c > 0$ *and sequence* $\{y_i\}$ *satisfying* $\|y_i\| \leq c$*, the sequence of functions* $\{g_i(\boldsymbol{\beta}) = \psi_i(y_i, \boldsymbol{\beta})\}$ *is equicontinuous on any open subset of* Θ*;*

(3) The function $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[N^{-1}s_n(\boldsymbol{\beta})]$ *has the property that, for any* $\epsilon > 0$*, there exists* $\delta_\epsilon > 0$ *such that* $\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon} |\Delta_N(\boldsymbol{\beta})| > \delta_\epsilon$*;*

(4) There exists a $\hat{\boldsymbol{\beta}}_n \in \Theta$ *which is a solution to* $s_n(\boldsymbol{\beta}) = 0$*, that is,* $\hat{\boldsymbol{\beta}}_n$ *is the pseudo-GEE estimator of* $\boldsymbol{\beta}$ *such that* $s_n(\hat{\boldsymbol{\beta}}_n) = 0$*;*

(5) $\hat{\boldsymbol{\beta}}_n = O_p(1)$*;*

(6) The design weights w_i *satisfy* $N^{-1}\sum_{i\in s} w_i Z_i - N^{-1}\sum_{i=1}^N Z_i = O_p(1/\sqrt{n})$ *for any variable* Z *such that* $N^{-1}\sum_{i=1}^N Z_i^2 = O(1)$*;*

then $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$*, where "p" denotes in probability with respect to both the model* ξ *and the sampling design* π*.*

Condition 5 is weaker than assuming the parameter space is compact, which is what Robins, Rotnitzky & Zhao (1995) assumed for their results. Here the "p" in $O_p(1)$ means in probability with respect to the joint $\xi\pi$ distribution. In condition 6 the "p" in $O_p(1/\sqrt{n})$ means under the distribution induced by the design π. This condition is weaker than assuming $N^{-1}\sum_{i\in s} w_i Z_i$ is asymptotically normally distributed. That is, if $\widehat{\bar{Z}}_{HT} = N^{-1}\sum_{i\in s} w_i Z_i \dot\sim N(\bar{Z}, \sigma^2/n)$, then condition 6 is satisfied. Hájek (1960, 1964) established the asymptotic normality of $\widehat{\bar{Z}}_{HT}$ under simple random sampling and rejective sampling with unequal selection probabilities. Víšek (1979) established the asymptotic normality of $\widehat{\bar{Z}}_{HT}$ for the well-known Rao–Sampford method of unequal probability sampling without replacement.

The following lemma, adapted from Lemma 5.3 of Shao (2003), plays a key role in proving Theorem 1. The proof of the lemma can be found in Carrillo-García (2008) and is omitted here.

Lemma 1. *Suppose that* Θ *is a compact subset of* $\mathbb{R}^p$ *and conditions* 1, 2, *and* 6 *specified in Theorem* 1 *hold. Then, as* $n \to \infty$*,*

$$\sup_{\boldsymbol{\beta}\in\Theta} \left\| \frac{1}{N} s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta}) \right\| \xrightarrow{p} 0,$$

where $s_n(\boldsymbol{\beta}) = \sum_{i\in s} w_i \psi_i(Y_i, \boldsymbol{\beta})$ *and* $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[N^{-1}s_n(\boldsymbol{\beta})] = N^{-1}\sum_{i=1}^N E_\xi[\psi_i(Y_i, \boldsymbol{\beta})]$.

Proof of Theorem 1. We carry out the proof in two cases.

Case 1: Θ is a compact subset of $\mathbb{R}^p$.
The following inequality holds:

$$\begin{aligned} \left|\frac{1}{N} s_n(\boldsymbol{\beta})\right| &= \left|\Delta_N(\boldsymbol{\beta}) + \frac{1}{N} s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \\ &\geq |\Delta_N(\boldsymbol{\beta})| - \left|\frac{1}{N} s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|. \end{aligned}$$

By Lemma 1, for any $\epsilon > 0$, we have

$$\begin{aligned}\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta})\right| &\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}\left\{|\Delta_N(\boldsymbol{\beta})| - \left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|\right\}\\ &\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}|\Delta_N(\boldsymbol{\beta})| - \sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|\\ &\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}|\Delta_N(\boldsymbol{\beta})| - \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|\\ &= \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}|\Delta_N(\boldsymbol{\beta})| + o_p(1). \qquad (5)\end{aligned}$$

It follows from condition 3 stated in the theorem that, for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$P_{\xi\pi}\left(\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\epsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta})\right| > \delta_\epsilon\right) \to 1$$

as $n \to \infty$. Noting that $s_n(\hat{\boldsymbol{\beta}}_n) = 0$ by condition 4, the above limit implies that, for any $\epsilon > 0$, $P_{\xi\pi}(|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| \leq \epsilon) \to 1$ as $n \to \infty$. This completes the proof that $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$.

Case 2: Θ is any subset of $\mathbb{R}^p$.
By condition 5 in the theorem, for any $\epsilon > 0$, there is an $M > 0$ such that $P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n\| \leq M) > 1 - \epsilon$ for all n. The result follows from Case 1 by considering the closure of $\Theta \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| \leq M\}$ as the parameter space. Let Θ^* be the closure of $\Theta \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| \leq M\}$. Then, for any $\delta > 0$,

$$\begin{aligned}P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta) &= P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta,\ \|\hat{\boldsymbol{\beta}}_n\| \leq M) + P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta,\ \|\hat{\boldsymbol{\beta}}_n\| > M)\\ &\leq P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta,\ \|\hat{\boldsymbol{\beta}}_n\| \leq M) + P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n\| > M)\\ &< P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta,\ \|\hat{\boldsymbol{\beta}}_n\| \leq M) + \epsilon\\ &\leq P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta,\ \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \epsilon\\ &= P_{\xi\pi}(\hat{\boldsymbol{\beta}}_n \in \Theta^*)P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta \,|\, \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \epsilon\\ &\leq P_{\xi\pi}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \delta \,|\, \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \epsilon\\ &\leq 2\epsilon,\end{aligned}$$

where the last line is due to the fact that Θ^* is compact and therefore Case 1 applies. It follows that $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$.

■

4. VARIANCE ESTIMATION

The variance of $\hat{\boldsymbol{\beta}}_n$ under the joint $\xi\pi$ randomization is given by

$$\mathrm{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) = \mathrm{Var}_\xi[E_\pi(\hat{\boldsymbol{\beta}}_n)] + E_\xi[\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)]. \qquad (6)$$

The first component in (6), $\mathrm{Var}_\xi[E_\pi(\hat{\boldsymbol{\beta}}_n)]$, is called the "model variance component" and represents the variance in a census fit to the model, using data from the entire finite population. The second component, $E_\xi[\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)]$, is called the "design variance component" or "sampling variance component" and represents the additional variance contributed by sampling from the finite population. It comes from the fact that a sample of n elements is observed rather than the entire finite population of N elements (Särndal, Swensson & Wretman, 1992).

Let $B = E_\pi(\hat{\boldsymbol{\beta}}_n)$; that is, B is the conceptual finite population quantity which is unbiasedly estimated by $\hat{\boldsymbol{\beta}}_n$. If $\mathrm{Var}_\xi(B)$ has the usual order of $1/N$ and suppose that the sampling fraction n/N is small or negligible, which is practically the case for most large-scale surveys, then the leading term in the joint variance is $E_\xi[\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)]$. Note that the B defined above is usually not identical to the census estimator $\boldsymbol{\beta}_N$. Therefore, we can write

$$\mathrm{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) \doteq E_\xi[\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)], \tag{7}$$

and estimate the joint variance of $\hat{\boldsymbol{\beta}}_n$ by $\hat{\mathrm{V}}\mathrm{ar}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) = \hat{\mathrm{V}}\mathrm{ar}_\pi(\hat{\boldsymbol{\beta}}_n)$, where $\hat{\mathrm{V}}\mathrm{ar}_\pi(\hat{\boldsymbol{\beta}}_n)$ is an approximately unbiased estimator of the design-based variance of $\hat{\boldsymbol{\beta}}_n$. The estimator $\hat{\mathrm{V}}\mathrm{ar}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n)$ is also approximately unbiased under the joint randomization since $E_{\xi\pi}[\hat{\mathrm{V}}\mathrm{ar}_\pi(\hat{\boldsymbol{\beta}}_n)] \doteq E_\xi[\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)]$.

We now develop a linearization estimator for $\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n)$. Let

$$U_n(\boldsymbol{\beta}) = \sum_{i\in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \quad \text{and} \quad U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i).$$

It follows that $U_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$ and $U_N(\boldsymbol{\beta}_N) = \mathbf{0}$. Let

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \quad \text{and} \quad \hat{H}(\boldsymbol{\beta}) = \sum_{i\in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}.$$

Applying a first order Taylor series expansion to $U_n(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_N$ and noting that $U_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$ and $U_n(\boldsymbol{\beta}_N) = U_N(\boldsymbol{\beta}_N) + O_p(N/\sqrt{n}) = O_p(N/\sqrt{n})$, we have

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_N = [\hat{H}(\boldsymbol{\beta}_N)]^{-1} U_n(\boldsymbol{\beta}_N) + o_p\left(\frac{1}{\sqrt{n}}\right) = [H(\boldsymbol{\beta}_N)]^{-1} U_n(\boldsymbol{\beta}_N) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

This further leads to

$$\mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n) \doteq [H(\boldsymbol{\beta}_N)]^{-1} \left[\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\Delta_{ij}}{\pi_i\pi_j} \mathbf{z}_i\mathbf{z}_j'\right] [H(\boldsymbol{\beta}_N)]^{-1},$$

where $\mathbf{z}_i = (\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_N$, $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$, π_i and π_{ij} are the first- and second-order inclusion probabilities under the sampling design. An approximately design unbiased variance estimator is given by

$$\hat{\mathrm{V}}\mathrm{ar}_\pi(\hat{\boldsymbol{\beta}}_n) = [\hat{H}(\hat{\boldsymbol{\beta}}_n)]^{-1} \left[\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_i\pi_j\pi_{ij}} \hat{\mathbf{z}}_i\hat{\mathbf{z}}_j'\right] [\hat{H}(\hat{\boldsymbol{\beta}}_n)]^{-1},$$

where $\hat{\mathbf{z}}_i$ is similarly defined as $\mathbf{z}_i$ with $\boldsymbol{\beta}_N$ replaced by $\hat{\boldsymbol{\beta}}_n$. However, this estimator requires the knowledge of the joint inclusion probabilities π_{ij}. Under the assumption that sampling fractions are small or negligible, the with-replacement variance formula can be used to avoid π_{ij}, resulting in the following variance estimator:

$$\hat{\mathrm{V}}\mathrm{ar}_\pi(\hat{\boldsymbol{\beta}}_n) = [\hat{H}(\hat{\boldsymbol{\beta}}_n)]^{-1} \left[\frac{1}{n-1}\left\{n\sum_{i\in s} w_i^2 \hat{\mathbf{z}}_i\hat{\mathbf{z}}_i' - \left(\sum_{i\in s} w_i\hat{\mathbf{z}}_i\right)^{\otimes 2}\right\}\right] [\hat{H}(\hat{\boldsymbol{\beta}}_n)]^{-1},$$

where $A^{\otimes 2} = AA'$.

5. SIMULATION STUDIES

In this section we present results from a comprehensive simulation study. Our simulation models and finite populations were built based on a synthetic data file from the first four cycles of NLSCY which was briefly described in Section 1. We consider several sampling designs and both continuous and binary responses and include several important covariates as identified by previous studies.

5.1. Simulation Settings for Continuous Response

The response variable is the PAS of a child. By design, PAS is an ordinal variable taking values between 0 and 12 or 16 depending on the age group. We treat PAS as a continuous variable in the simulation. Previous studies using data from NLSCY, including Carrillo et al. (2005), Carrillo-García (2006), and Carrillo, Kovacevic & Wu (2006), found that factors which are significant for PAS include the age of the child (AGE), the square of the age (AGE^2), the depression score of the person most knowledgeable about the child (DeprePMK), the punitive/aversive parenting score (Punitive), and the child's gender (GENDER). In the simulation we generated finite populations from the following simpler model:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + \epsilon_{ij}, \tag{8}$$

where Y_{ij} is the PAS of subject i at jth cycle, x_{ij1} is age of subject i at jth cycle, x_{ij2} is depression score of the PMK of subject i at jth cycle, x_{i3} is gender of subject i, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4}) \sim (\mathbf{0}, \sigma^2\mathbf{R})$, and $\mathbf{R}$ is the 4 × 4 correlation matrix. The pseudo-GEE method is then applied to the synthetic NLSCY data set which contains complete observations for 458 children, using model (8) and unspecified correlation structure estimated by the method of moments as in Liang & Zeger (1986). The estimated regression coefficients are $\beta_0 = 5.6225, \beta_1 = -1.0982, \beta_2 = 0.0656, \beta_3 = 0.0609$, and $\beta_4 = -0.2900$, with correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & 0.4123 & 0.3919 & 0.3353 \\ 0.4123 & 1 & 0.4798 & 0.3172 \\ 0.3919 & 0.4798 & 1 & 0.4370 \\ 0.3353 & 0.3172 & 0.4370 & 1 \end{pmatrix}$$

and dispersion parameter $\phi = \sigma^2 = 3.66842$. We set the parameters in our superpopulation model to these values. The finite population used for our repeated simulations was generated as follows. First, the data set of 458 children was duplicated 40 times, resulting in $N = 18,320$ children with complete information on AGE, DeprePMK, and GENDER. Values of the response variable, $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$, were then generated based on $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})$, $\mathbf{R}$ and σ^2 using multivariate normal distributions, where $\mu_{ij} = E_\xi(Y_{ij}|\mathbf{x}_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3}$ and $\mathbf{x}_{ij} = (x_{ij1}, x_{ij1}^2, x_{ij2}, x_{i3})$.

We considered three sampling schemes: (i) simple random sampling (SRS) without replacement; (ii) stratified simple random sampling (STSI); and (iii) cluster sampling with clusters selected by simple random sampling (SIC). For stratified sampling, two strata were formulated based on AGE at first cycle, with first stratum having $N_1 = 9,000$ units and the second stratum having $N_2 = 9,320$ units. The stratum sample sizes were allocated as $n_1 = n/3$ and $n_2 = 2n/3$. For cluster sampling, the population units were artificially grouped into clusters of sizes 5 or 10, and the cluster effect was created by using $\mu_{ijc} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + b_{cj}$ with $b_{cj} \sim N(0, 1)$. Since $\phi = \sigma^2 = 3.66842$, this produces a correlation coefficient of 0.2142 between any two subjects in the same cluster and cycle. The overall sample size for a particular sampling scheme ranges from $n = 120$ to $n = 1,200$, and sampling fractions n/N are in between

0.65% and 6.5%. For cluster sampling, the sample sizes are random. The above numbers are expected sample sizes.

With regard to the estimation procedure, since we deal with a continuous response, the point estimator of Section 3 is obtained as follows. We start with the initial value $\boldsymbol{\beta}^{(0)} = (\sum_{i\in s} w_i X_i X_i')^{-1} \sum_{i\in s} w_i X_i \mathbf{y}_i$. Let $e_{it}^{(0)} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}^{(0)}$ and $\mathbf{R}^{(0)} = (\hat{\alpha}_{tt'}^{(0)})$, where

$$\hat{\alpha}_{tt'}^{(0)} = \widehat{\text{corr}}(y_{it}, y_{it'}) = \frac{\sum_{i\in s} w_i e_{it}^{(0)} e_{it'}^{(0)}}{\phi^{(0)}(\sum_{i\in s} w_i - p)} \quad \text{and} \quad \phi^{(0)} = \frac{\sum_{i\in s}\sum_{t=1}^{4} w_i (e_{it}^{(0)})^2}{\sum_{i\in s}\sum_{t=1}^{4} w_i - p}$$

for $t \neq t'$ and $\hat{\alpha}_{tt}^{(0)} = 1$ for $t = 1, 2, 3$, and 4. The $(l+1)$th iteration on $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}^{(l+1)} = \left(\sum_s w_i X_i [\mathbf{R}^{(l)}]^{-1} X_i'\right)^{-1} \sum_s w_i X_i [\mathbf{R}^{(l)}]^{-1} \mathbf{y}_i .$$

Finally, $e_{it}^{(l+1)}$, $\phi^{(l+1)}$, and $\mathbf{R}^{(l+1)}$ are computed at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(l+1)}$.

5.2. Simulation Settings for Binary Response

In this part of the simulation the PAS was recoded as "low" ($Y_{ij} = 0$) if the original PAS is 1.5 or less and recoded as "high" ($Y_{ij} = 1$) if the original PAS is bigger than 1.5.

For multivariate binary response, the correlation structure is better described by the odds ratio parametrization rather than Pearson correlation. Song (2007) pointed out that "to measure dependence between non-normal variables, there are some better tools than Pearson correlation. For example, odds ratio (OR) is a measure of association for categorical variates." Lipsitz, Laird & Harrington (1991), Liang, Zeger & Qaqish (1992), and Carey, Zeger & Diggle (1993) used odds ratios to measure the association among binary and other categorical data. For binary responses, the odds ratio has some desirable properties and is easier to interpret than the correlation coefficient.

Let $p_{ij} = P(Y_{ij} = 1 \mid \mathbf{x}_{ij})$. We consider the following logistic regression model to generate binary responses for the finite population:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3}. \tag{9}$$

The true values of the model parameters β_0, β_1, β_2, β_3, and β_4 as well as the odds ratios OR_{st} between responses at times s and t were obtained by fitting the logistic regression model (9) to the complete NLSCY data set. They are given by $\beta_0 = 2.7181$, $\beta_1 = -0.8959$, $\beta_2 = 0.0530$, $\beta_3 = 0.0701$, $\beta_4 = -0.2811$, $\text{OR}_{12} = 4.7669$, $\text{OR}_{13} = 3.9257$, $\text{OR}_{14} = 3.0930$, $\text{OR}_{23} = 5.8401$, $\text{OR}_{24} = 4.4069$ and $\text{OR}_{34} = 6.6430$. The dispersion parameter for this case is $\phi = 1$. Values of p_{ij} for given covariates were obtained from model (9) and binary responses Y_{ij} were generated based on p_{ij} and the odds ratios OR_{st}. This was done based on the Gaussian copula method as described in Song (2000).

Once again, the three sampling schemes described in Section 5.1 were used for taking simulation samples, with overall sample sizes ranging from $n = 120$ to $n = 1,200$. For cluster sampling, the clustering effect was created through $p_{ijc} = \{1 + \exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + b_{cj})\}^{-1}$, with $b_{cj} \sim N(0, 0.2)$.

Estimation procedures for binary response proceed as follows. Our initial value for $\hat{\boldsymbol{\beta}}_n$ is $\boldsymbol{\beta}^{(0)} = \left(\beta_0^{(0)}, 0, 0, 0, 0\right)'$, where $\beta_0^{(0)} = \log\left\{\left(\sum_{i\in s}\sum_{t=1}^{4} w_i y_{it}\right)/\left(\sum_{i\in s}\sum_{t=1}^{4} w_i(1 - y_{it})\right)\right\}$. In other words, $\beta_0^{(0)}$ is the estimate of the log odds of high PAS, collapsing all four cycles of responses and ignoring all covariates.

We estimate the six odds ratios as

$$\widehat{\mathrm{OR}}_{st} = \frac{\sum_{i\in s} w_i y_{it} y_{is} \cdot \sum_{i\in s} w_i (1-y_{it})(1-y_{is})}{\sum_{i\in s} w_i y_{it}(1-y_{is}) \cdot \sum_{i\in s} w_i (1-y_{it}) y_{is}}, \tag{10}$$

where $st = 12$, 13, 14, 23, 24, and 34.

At the lth iteration with given $\boldsymbol{\beta}^{(l)}$, we let $\mathbf{R}_i^{(l)} = (\hat{\alpha}_{ist})$ where

$$\hat{\alpha}_{ist} = \widehat{\mathrm{corr}}(Y_{is}, Y_{it}) = \frac{\hat{p}_{ist} - \hat{\mu}_{is}\hat{\mu}_{it}}{\sqrt{\hat{\mu}_{is}(1-\hat{\mu}_{is})\hat{\mu}_{it}(1-\hat{\mu}_{it})}},$$

$\hat{\mu}_{it} = (1+\exp(X_{it}'\boldsymbol{\beta}^{(l)}))^{-1}$ and $\hat{p}_{ist}$, an estimate of $E_\xi(Y_{is}Y_{it}) = P(Y_{is}=1, Y_{it}=1)$, given for example in Liang et al. (1992) or Lipsitz et al. (1991), has the form

$$\hat{p}_{ist} = \begin{cases} \frac{f_{ist} - \{f_{ist}^2 - 4\widehat{\mathrm{OR}}_{st}(\widehat{\mathrm{OR}}_{st}-1)\hat{\mu}_{is}\hat{\mu}_{it}\}^{1/2}}{2(\widehat{\mathrm{OR}}_{st}-1)} & \text{if } \widehat{\mathrm{OR}}_{st} \neq 1 \\ \hat{\mu}_{is}\hat{\mu}_{it} & \text{if } \widehat{\mathrm{OR}}_{st} = 1, \end{cases}$$

with $f_{ist} = 1 - (1-\widehat{\mathrm{OR}}_{st})(\hat{\mu}_{is} + \hat{\mu}_{it})$. The updated $\boldsymbol{\beta}^{(l+1)}$ is computed as

$$\boldsymbol{\beta}^{(l)} + \left(\sum_{i\in s} w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^{(l)}} \left[\hat{A}_i^{1/2}\hat{\mathbf{R}}_i^{(l)}\hat{A}_i^{1/2}\right]^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}^{(l)}}\right)^{-1} \sum_{i\in s} w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^{(l)}} \left[\hat{A}_i^{1/2}\hat{\mathbf{R}}_i^{(l)}\hat{A}_i^{1/2}\right]^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i),$$

where $\partial\hat{\mu}_{it}/\partial\boldsymbol{\beta}^{(l)} = \hat{\mu}_{it}(1-\hat{\mu}_{it})X_{it}$ and $\hat{A}_i = \mathrm{diag}[\hat{\mu}_{i1}(1-\hat{\mu}_{i1}), \ldots, \hat{\mu}_{i4}(1-\hat{\mu}_{i4})]$.

5.3. Results

We report results for small, medium, and large sample sizes ($n = 240$, 720, and 1, 200) here. More results can be found in Carrillo-García (2008). Our simulations were programmed in the R software package, as documented in R Development Core Team (2008), and run on a UNIX machine with 24 CPUs. All results are based on 1,000 simulation runs.

The relative bias of the estimator $\hat{\boldsymbol{\beta}}_n$ is calculated as $\mathrm{RB}(\hat{\boldsymbol{\beta}}_n) = 1,000^{-1}\sum_{k=1}^{1,000}(\hat{\boldsymbol{\beta}}_n^{(k)} - \boldsymbol{\beta})/\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}_n^{(k)}$ is the estimate of $\boldsymbol{\beta}$ from the kth simulated sample. The simulated relative biases of $\hat{\boldsymbol{\beta}}_n$ for continuous response are summarized in Table 1 and the relative biases for binary response are reported in Table 2.

For all three sampling schemes considered and for either continuous or binary responses, the largest relative bias (in absolute value) is about 6%, which occurs with the smallest sample size $n = 120$ (not shown here). For all other cases the largest relative bias is about 3%. For sample sizes of 720 or bigger, the maximum relative bias is bounded by around 2%. A general trend is that, as sample sizes increase, relative biases tend to decrease, although the pattern is not strictly monotone.

To evaluate the performance of our proposed variance estimator, we need to find the true variance–covariance matrix $\mathrm{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n)$ for each simulation model and sampling design. We approximated this variance matrix through 1,000 independently simulated samples based on the following formula:

$$\mathrm{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) \doteq \frac{1}{1,000}\sum_{k=1}^{1,000}(\hat{\boldsymbol{\beta}}_n^{(k)} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n^{(k)} - \boldsymbol{\beta})',$$

TABLE 1: Relative bias of $\hat{\boldsymbol{\beta}}_n$ (in %) for continuous response.

Design	n	β_0	β_1	β_2	β_3	β_4
SRS	240	−0.16	−0.12	−0.08	−0.15	−3.50
	720	0.19	0.15	0.07	−0.71	0.09
	1,200	0.01	−0.03	0.02	−0.56	−0.06
STSI	240	0.32	0.39	0.46	0.01	1.72
	720	−0.05	−0.28	−0.37	0.34	1.42
	1,200	0.06	0.14	0.22	0.52	0.67
SIC	240	−0.19	−0.12	−0.04	0.14	1.96
	720	0.03	0.00	−0.03	−0.17	0.42
	1,200	0.05	−0.15	−0.39	0.71	0.02

TABLE 2: Relative bias of $\hat{\boldsymbol{\beta}}_n$ (in %) for binary response.

Design	n	β_0	β_1	β_2	β_3	β_4
SRS	240	0.47	0.17	−0.17	−1.29	−0.47
	720	−0.15	−0.10	−0.12	−0.62	−0.19
	1,200	0.03	0.05	0.01	−0.07	−0.74
STSI	240	1.38	1.06	1.15	0.14	3.24
	720	0.24	0.24	0.13	−0.56	−1.19
	1,200	−0.21	−0.18	−0.20	−0.18	0.77
SIC	240	0.84	0.90	0.97	−0.04	−1.16
	720	0.25	0.25	0.21	−0.12	−1.05
	1,200	0.11	−0.01	−0.09	−1.22	−1.23

where $\hat{\boldsymbol{\beta}}_n^{(k)}$ is the estimate of $\boldsymbol{\beta}$ computed from the kth simulated sample. To simplify notation, we use $V = (V_{lm})$ to denote the 5 × 5 variance matrix $\mathrm{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n)$ and use $\hat{V} = (\hat{V}_{lm})$ to denote the estimated variance matrix. Relative biases of the variance estimator were calculated through another 1,000 simulated samples using the following formula:

$$\mathrm{RB}(\hat{V}_{lm}) = \frac{1}{1,000}\sum_{k=1}^{1,000}(\hat{V}_{lm}^{(k)} - V_{lm})/(V_{ll}V_{mm})^{1/2}.$$

where $\hat{V}_{lm}^{(k)}$ is calculated from the kth simulated sample. Results on the simulated relative biases of the variance estimator for $n = 240$ are reported in Table 3. Results for other sample sizes can be found in Carrillo-García (2008).

For the model with continuous response, the vast majority of relative biases under simple random sampling and stratified simple random sampling is below 5%, with a few exceptions around 7–10%. Under cluster sampling, the relative biases are generally a bit bigger, with about one quarter of the entries around 8–11%. The observed negative or positive biases do not seem to have a clear pattern in terms of the magnitude of the sample sizes.

TABLE 3: Relative biases (in %) of variance estimators for $n = 240$.

Design	Continuous Response					Binary Response				
	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
SRS	−8					−1				
	5	−2				1	−2			
	−3	1	−1			−1	2	−3		
	3	−3	3	−8		−3	2	−3	−1	
	5	−4	4	6	−10	0	2	−3	−2	−4
STSI	−1					−1				
	1	−2				1	−2			
	0	1	−1			0	1	−1		
	0	4	−4	−5		0	4	−4	−5	
	−4	5	−5	−7	0	−4	5	−5	−7	0
SIC	−8					0				
	8	−8				−1	2			
	−8	9	−9			−1	0	−1		
	−3	2	−2	−10		4	−5	7	−1	
	2	−1	1	7	−4	4	−2	3	−4	−9

For binary responses, relative biases are all smaller than 10%. However, the magnitude of biases does not seem to be closely related to the overall sample sizes or a particular sampling design. It is more related to the actual values of the true variance. For instance, some of the largest relative biases are observed for simple random sampling with $n = 1,200$, where the actual values of the true variances are extremely small.

6. CONCLUDING REMARKS

The GEE methodology has been widely used in analyzing longitudinal survey data in recent years. The use of survey weights in this type of analysis; however, is left open and is usually at the discretion of the data analyst. In this paper we argue that a joint randomization approach is generally appropriate for analyzing complex longitudinal survey data using the GEE method. We have rigorously established the consistency of the proposed pseudo-GEE estimators under the joint randomization framework. Rubin-Bleuer & Schiopu Kratina (2005) presented a similar framework for joint model and design-based inference under a more mathematical treatment using a product probability space. We take a more pragmatic approach in this paper through a conditional argument, conditioning on a particular order of the randomizations involved. We also developed linearization variance estimators for general unequal probability sampling designs assuming that the finite population sampling fractions are small. This later development also echoes the arguments in Binder & Roberts (2003) that design-based inference is usually appropriate even if the goal is to make inference on superpopulation parameters.

Our extensive simulation studies showed that the pseudo-GEE estimator has excellent finite sample performance. The proposed linearization variance estimator performs reasonably well for most cases but the message on scenarios where relative biases are larger than (say) 8% is not clear.

Variance estimation using replication weights has been a popular topic among survey researchers, especially those from Statistics Canada and other large organizations. We are currently examining re-sampling variance estimation techniques for the pseudo-GEE estimator.

The pseudo-GEE method can be extended to handle cases where there are missing values for the response variable. This can be done either by re-weighting the estimating equations using the response probabilities or through imputation. These results will be reported elsewhere.

ACKNOWLEDGMENTS

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and an internship grant to the first author by Mathematics of Information Technology and Complex Systems (MITACS) and National Program for Complex Data Structures (NPCDS).

BIBLIOGRAPHY

D. A. Binder & Z. Patak (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035–1043.

D.A. Binder & G.R. Roberts (2003). Design-Based and Model-Based Methods for Estimating Model Parameters, in "*Analysis of Survey Data*," R.L. Chambers and C.J. Skinner, editors, Wiley, Chichester.

V. Carey, S. L. Zeger & P. Diggle (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80, 517–526.

I. Carrillo, C. Chu, W. Su & X. Xie (2005). A longitudinal study of factors affecting children's behaviour. *Proceedings of the Survey Methods Section*, Saskatoon. Statistical Society of Canada.

I. Carrillo, M. Kovacevic & C. Wu (2006). Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY). *Proceedings of the Survey Methods Section*, London. Statistical Society of Canada.

I. A. Carrillo-García (2006). Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY). *Technical Report*, MITACS/NPCDS Internship Program. Statistics Canada.

I. A. Carrillo-García (2008). *Analysis of Longitudinal Surveys with Missing Responses*. PhD Thesis, University of Waterloo, ON, Canada.

J. Chen, M. E. Thompson & C. Wu (2004). Estimation of fish abundance indices based on scientific research trawl surveys. *Biometrics*, 60, 116–123.

P. Diggle, P. Heagerty, K.-Y. Liang & S. Zeger (2002). "*Analysis of Longitudinal Data*," 2nd ed., Oxford University Press, New York.

G. J. Duncan & G. Kalton (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97–117.

G. M. Fitzmaurice, N. M. Laird & J. H. Ware (2004). "*Applied Longitudinal Analysis*," Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

V. P. Godambe (1995). Estimation of parameters in survey sampling: Optimality. *The Canadian Journal of Statistics*, 23, 227–243.

V. P. Godambe & M. E. Thompson (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127–138.

J. Hájek (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of Hungarian Academy of Science*, 5, 361–375.

J. Hájek (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35, 1491–1523.

D. Hedeker & R. D. Gibbons (2006). "*Longitudinal Data Analysis*," Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

D. G. Horvitz & D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

G. Kalton (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175–188.

L. Kish (1987). "*Statistical Design for Research*," John Wiley & Sons, New York.

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

K.-Y. Liang, S. L. Zeger & B. Qaqish (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society Series B*, 54, 3–40.

S. R. Lipsitz, N. M. Laird & D. P. Harrington (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78, 153–160.

R Development Core Team (2008). "*R: A Language and Environment for Statistical Computing*," R Foundation for Statistical Computing, Vienna.

J. N. K. Rao (1998). Marginal models for repeated observations: inference with survey data. *Proceedings of the Section on Survey Methods Research*, the American Statistical Association, 76–82.

G. Roberts, Q. Ren & J. N. K. Rao (2009). Using marginal mean models for data from longitudinal surveys with a complex design: some advances in methods. in "*Methodology of Longitudinal Surveys*," R. Lynn (Ed.), John Wiley & Sons, New York, pp. 351–366.

J. M. Robins, A. Rotnitzky & L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

S. Rubin-Bleuer & I. Schiopu Kratina (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, 2789–2810.

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

J. Shao (2003). "*Mathematical Statistics*," 2nd ed., Springer-Verlag, New York.

P. X.-K. Song (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27, 305–320.

P. X.-K. Song (2007). "*Correlated Data Analysis: Modeling, Analytics, and Applications*," Springer Series in Statistics, New York.

E. M. Thomas (2004). Aggressive behaviour outcomes for young children: Change in parenting environment predicts change in behaviour. *Children and Youth Research Paper Series*. Catalogue Number 89–599-MIE. Statistics Canada.

J. A . Vısek. (1979) Asymptotic Distribution of Simple Estimate for Rejective, Sampford and Successive Sampling, in "Contributions to Statistics," J. Jureckova (Editor), Reidel, Dordrecht, pp. 263–275.

C. Wu (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937–951.

K.-H. Yuan & R. I. Jennrich (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65, 245–260.

Received 23 November 2009
Accepted 22 July 2010