

EFFICIENCY COMPARISONS OF GEE VERSUS IEE FOR LONGITUDINAL COMPLEX SURVEY DATA WITH ORDINAL RESPONSES

Abdelnasser Saïdi, Diane Stukel and Susana Rubin-Bleuer ¹

ABSTRACT

We use the proportional odds marginal model and both the “independence estimating equations” (IEE) and the “generalized estimating equations” (GEE) approaches for modeling longitudinal ordinal complex survey data. The IEE approach assumes that observations on a subject over time are independent, whereas the GEE approach assumes an underlying working correlation structure over time. Our objective is to determine “best practices” under this set-up. For this purpose, we compare through simulation the GEE versus the IEE in terms of efficiency under a model with high longitudinal correlation. To measure the efficiency, we consider three approaches to obtaining variance estimates, including the “Liang & Zeger” (Taylor “sandwich”) and the One-Step Jackknife variances.

KEY WORDS: Generalized equation, Jackknife, Ordinal survey data, Proportional odds model, Sandwich.

RÉSUMÉ

Nous utilisons le modèle marginal à cotes proportionnelles et l’approche des « équations estimantes généralisées » (EEG) pour modéliser les données ordinales d’enquêtes longitudinales. L’approche des « équations estimantes indépendantes » (EEI) suppose que les observations répétées d’un sujet dans le temps sont indépendantes alors que l’approche EEG suppose l’existence d’une structure de corrélation longitudinale de travail. Notre objectif est de définir des « bonnes pratiques » sous ce cadre de travail. Nous comparons par une étude de simulation l’approche EEG contre EEI en termes d’efficacité sous un modèles ayant une forte corrélation longitudinale. Pour mesurer l’efficacité, nous considérons trois approches pour obtenir les variances des estimateurs incluant la variance de « Liang & Zeger » (Taylor « sandwich ») et la variance Jackknife en une étape.

MOTS CLÉS : Donnée ordinaire d’enquête; équation généralisée; jackknife; modèle à cotes proportionnelles; sandwich.

1. INTRODUCTION

1.1 Description of the Problem

Ways of modeling longitudinal ordered categorical response variables when the data are obtained from a simple random sample of subjects have been established for some time and have been fully surveyed by Agresti and Natarajan (2001). In this paper we concentrate on marginal modeling and analysis of longitudinal survey data, i.e. longitudinal data obtained from a complex design. We focus on estimating the finite population parameter defined under a working model. We assume that in a longitudinal study of M subjects (or units) there are T occasions of measurement. Suppose that a categorical variable with J ordered categories is the characteristic of interest and let $Y_{ij} = 1$ if subject i has response j at

time t , and $Y_{ij} = 0$ otherwise, with $\sum_{j=1}^J Y_{ij} = 1$, $i = 1, \dots, M$. Each subject has, at each occasion, a $(a_s + a) \times 1$ vector

¹Abdelnasser Saïdi, Statistics Canada, 15 K R.H.Coats Building, Tunney’s Pasture, Ottawa, ON K1A 0T6, saidabd@statcan.ca, Diane Stukel, UNESCO Institute for Statistics, P.O Box 6128 Succursale Centre-Ville, Montreal, QC H3C 3J7, d.stukel@uis.unesco.org, Susana Rubin-Bleuer, Statistics Canada, 11 L R.H.Coats Building, Tunney’s Pasture, Ottawa, ON K1A 0T6, rubisus@statcan.ca

$(\mathbf{x}'_{is}, \mathbf{x}'_{it})'$ of associated covariates containing a_s time-stationary covariates, including intercepts, and a time-varying covariates. Let $\boldsymbol{\beta}_s, \boldsymbol{\beta}_t$ be the corresponding vectors of parameters. Let $\mu_{itj} = pr(Y_{itj} = 1 | x_{is}, x_{it}, \boldsymbol{\beta}_s, \boldsymbol{\beta}_t)$, $j = 1, \dots, J$. The marginal distribution of the $J \times 1$ vector $(Y_{it1}, \dots, Y_{it(J-1)}, Y_{itJ})'$ is multinomial with mean $(\mu_{it1}, \dots, \mu_{it(J-1)}, \mu_{itJ})'$. Now let $\mathbf{Y}_{it} = (Y_{it1}, \dots, Y_{it(J-1)})'$, $\boldsymbol{\mu}_{it} = (\mu_{it1}, \dots, \mu_{it(J-1)})'$ and $\mathbf{A}_{it} = \text{cov}(\mathbf{Y}_{it}) = \text{Diag}(\boldsymbol{\mu}_{it}) - \boldsymbol{\mu}_{it} \cdot \boldsymbol{\mu}'_{it}$ for $1 \leq t \leq T$. Let $\gamma_{itj} = \mu_{it1} + \dots + \mu_{itj}$ denote the cumulative probabilities. The proportional odds model is given by

$$\log\left(\frac{\gamma_{itj}}{1 - \gamma_{itj}}\right) = \theta_j + \boldsymbol{\beta}'_s x_{is} + \boldsymbol{\beta}'_t \mathbf{x}_{it}, \quad j = 1, \dots, J-1, \quad t = 1, \dots, T, \quad i = 1, \dots, M. \quad (1.1)$$

Set $\mathbf{Y}_i = (\mathbf{Y}'_{i1}, \dots, \mathbf{Y}'_{iT})'$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i1}, \dots, \boldsymbol{\mu}'_{iT})'$ and $\mathbf{D}'_i = \partial \boldsymbol{\mu}'_i / \partial \boldsymbol{\beta}$, where $\boldsymbol{\beta}' = (\theta_1, \dots, \theta_{(J-1)}, \boldsymbol{\beta}'_s, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_T)$.

The census estimating equations are given by

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^M \mathbf{U}_i(\boldsymbol{\beta}) = 0, \quad \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{D}'_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) \quad (1.2)$$

Under the working assumption of longitudinal independence, $\mathbf{V}_i = \mathbf{A}_i = \text{Diag}(\mathbf{A}_{i1}, \dots, \mathbf{A}_{iT})$ yielding the census “independence estimating equations” $U_{IEE} = 0$. The census IEE estimator $\boldsymbol{\beta}_{IEE}$ is the solution of the census IEE. Otherwise if we assume a working longitudinal covariance of the form $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, with common correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ across subjects, we have the census “generalized estimating equations” (GEE): $U_{GEE} = 0$. The census GEE estimator $\boldsymbol{\beta}_{GEE}$ is the solution of the census GEE. The GEE method, first proposed by Liang and Zeger(1986), does not fully specify the distribution of \mathbf{Y}_i ; rather, “the model is specified for the mean and the variance function expressing the dependence of the variance on the mean. In the longitudinal context here, one also uses a working guess for the correlation structure and the estimates are solutions of the GEE. Estimates of the model parameter are consistent even if the correlation structure is misspecified” (Agresti & Natarajan, 2001).

Marginal models for longitudinal survey data were first proposed by Rao (1998), where he developed Wald and quasi-score tests for binary longitudinal survey data using Taylor linearization and jackknife methods, which take account of the design and longitudinal features. Next, Sutradhar and Kovacevic (2000) proposed a GEE working model for ordinal longitudinal data, based on a nominal or generalized logit model for the marginal distributions and a longitudinal correlation structure that depends on one time lag only. They used a method of moments to obtain consistent estimates of the longitudinal correlations and estimated the standard error of the estimated regression parameter by the model based sandwich estimator. If the necessary design information was known, they suggested that the middle part of the sandwich variance estimator be substituted by the appropriate design-based estimate. Neither Rao (1998) nor Sutradhar and Kovacevic (2000) gave the asymptotic properties of the survey GEE estimators.

Rubin-Bleuer et al. (2005, 2006) adapted the GEE approach of Rao (1998) to the analysis of ordinal longitudinal survey responses. For the marginal distributions, they used the proportional odds model. They modeled the longitudinal covariance structure under the working independence assumption and under a working unstructured correlation assumption, in the survey data context. They used the longitudinal component of a survey, that is, subjects that responded to all T cycles of the survey and the longitudinal survey weights. They estimated standard error of the estimated regression parameter by the design-based sandwich estimator with the linearized estimating function bootstrap technique of Binder et al.(2004). They proved the asymptotic normality of the survey IEE and GEE estimators under different response models and illustrated the method with data from the Canadian National Population Health Survey (NPHS).

We next describe the NPHS data and discuss the estimates obtained from the model fitting in order to better motivate the questions we studied in this paper. The Canadian National Population Survey produces repeated observations over time on self-perceived health (SPH) along with a wealth of covariate information such as physical, psychological and socio-

economic factors. NPHS has a stratified multi-stage design, where in the first stage strata were formed and independent samples of clusters were drawn from each stratum under a probability proportional to size (*pps*) design. The first cycle of data collection took place in 1994-1995 and subsequent cycles occurred every two years. Shields and Shoostari (2001) studied self-perceived health (SPH) and its covariates using cross-sectional logistic binary model for the first 3 waves of data. Rubin-Bleuer et al. (2005, 2006) examined self-perceived health (SPH) and its relation to age, gender, mobility and other factors, with 4 waves of data, using instead a longitudinal ordinal model, where SPH is now an ordinal variable with 3 response levels, and the covariates are binary. The model was fitted to the data from 8247 individuals aged 25 and older in 1994 (3436 male and 4811 female), who responded to all the four waves. 496 records with item non-response were dropped. Predictors of self-perceived health were studied by fitting model (1.1) under both independence (IEE method) and the unstructured working correlation proposed by Rao (1998) (GEE method).

In the results of Rubin-Bleuer et al (2005, 2006) both the IEE and the GEE fitting methods showed that physical conditions (functional dependency, number of chronic diseases, pain level), socio-economic factors (low education, low income) and psychological factors (low emotional support, low self-esteem, low socio-economic status) were negatively associated with good health perceptions. For the complete list of estimates, see Rubin-Bleuer et al (2006). Table 1 here lists only a portion of the estimates along with the corresponding standard errors. The variable DEPENDENT means the individual needs help in meal preparation, shopping or in personal care. Estimated standard errors of the parameters shown here are very similar under both correlation structures. We remark that the estimated correlations from the GEE model for different time periods were low, for example for periods 1 and 2 the longitudinal correlations between responses (excellent/very good) and good were

$$R_{12} = \begin{pmatrix} 0.25 & -0.04 \\ -0.00 & 0.17 \end{pmatrix}.$$

Table 1 – coefficient estimates and standard error estimates of IEE and GEE

	$\hat{\beta}_{IEE}$	$\hat{\beta}_{GEE}$	$\hat{\sigma}_p(\hat{\beta}_{IEE})$	$\hat{\sigma}_p(\hat{\beta}_{GEE})$
θ_1	1.97	1.86	0.127	0.131
θ_2	4.30	4.09	0.141	0.147
AGE 25-34	0.15	0.18	0.081	0.084
45-54	-0.15	-0.16	0.103	0.112
55-64	-0.30	-0.31	0.075	0.077
65-74	-0.35	-0.36	0.128	0.125
75 et +	-0.30	-0.36	0.181	0.178
DEPENDENT 94	-0.98	-0.84	0.141	0.121
DEPENDENT 96	-1.09	-0.90	0.146	0.140
DEPENDENT 98	-1.31	-1.18	0.115	0.120
DEPENDENT 00	-1.21	-1.09	0.093	0.086

Many issues came up from looking at these results. For example, the two methods yielded “similar” parameter estimates and “similar” design-based standard errors. Does the similarity of the parameter estimates indicate that the assumed marginal model is correct? How will the fitting of an incorrect marginal model affect the estimation of the finite population parameters β_{GEE} and β_{IEE} ? Does the low value of estimated correlations explain the similarity of the standard errors? Will $\hat{\beta}_{GEE}$ have lower design-based variability than $\hat{\beta}_{IEE}$ when the longitudinal correlation is higher? Furthermore, the variance estimators used here account for the various survey processes. How good are these variance estimators? And how different are they from the naïve method used by many analysts for variance estimation?

In this paper we attempt to answer some of these questions: we obtain a longitudinal finite population with two time points from a model with high longitudinal correlation and a marginal proportional odds model. From this population we define the finite population parameters β_{GEE} and β_{IEE} . We then define a stratified two stage sampling design and we compare the performance of the sample estimators $\hat{\beta}_{GEE}$ and $\hat{\beta}_{IEE}$ and three different variance estimators via Monte Carlo simulation. In section 2 we set more notation related to survey estimation, define the sample estimators of β_{GEE} and

β_{IEE} and state some asymptotic results previously studied. We also define different variance estimators for the purpose of comparison. We define the “Naïve Liang & Zeger” variance estimator, used by many analysts, which is a modified Liang & Zeger sandwich estimator, where each of the three multiplicative factors is replaced by their respective survey estimator with normalized survey weights. Next we define the “Proper Liang & Zeger” variance estimator, which properly accounts for all the survey processes by utilizing the correct design-based variance of the middle factor of the sandwich estimator and the One Step Jackknife variance estimator, which is a re-sampling method for estimating the variance of an estimator (Rao and Tausi 2004). Section 3 explains how the longitudinal finite population was simulated and states the specific sample design subsequently used in the Monte Carlo experiment of 1000 samples. In Section 4 we assess the performance of the IEE and GEE methods. We define, for the survey estimators $\hat{\beta}_{IEE}$ and $\hat{\beta}_{GEE}$ several performance measures and then compare the results obtained for the different methods of variance estimation under both the IEE and GEE approaches. We present our discussion and conclusions in Section 5.

2. IEE AND GEE INFERENCE

3.1 The survey data and the asymptotic framework

We assume the census or finite population of size M is composed of L strata, where the h th stratum has N_h primary sampling units (PSUs) and the i th PSU has M_{hi} secondary units hik , $k = 1, \dots, M_{hi}$, $i = 1, \dots, N_h$, $h = 1, \dots, L$. The data are obtained from a stratified sample s of n primary sampling units, distributed as $n = n_1 + \dots + n_L$, where $n_h \geq 2$ PSUs are selected in stratum h and m_{hi} secondary sample units are selected in PSU hi , so $m = \sum_{h=1}^L \sum_{i=1}^{n_h} m_{hi}$ is the overall sample size. In this study we assume complete response, no attrition and no post-stratification adjustments, so the source of survey variation is due exclusively to the sample design. For the sake of simplicity and without loss of generality, we use the notation $i = 1, 2, \dots, M$, rather than hik for the unit labels, without the design classification into stratum, PSU, and secondary sampling unit identification. In this set-up, the census values y_i associated with the finite population labels i are non-stochastic numbers.

3.2 The Survey Independent Estimating Equation (IEE) Estimator of β_{IEE}

Let $w_i = 1 / \text{prob}(i \in s)$ and let $\hat{U}_{IEE}(\beta)$ be the design-consistent estimator of the census function $U_{IEE}(\beta)$ given by:

$$\hat{U}_{IEE}(\beta) = \sum_{i \in s} w_i U_i(\beta) \text{ with } U_i(\beta) = \mathbf{D}'_i \mathbf{A}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i). \quad (2.1)$$

The survey IEE estimator $\hat{\beta}_{IEE}$ is the solution of the survey IEE $\hat{U}_{IEE}(\beta) = 0$. Under some regularity conditions both the survey IEE and the survey IEE estimator $\hat{\beta}_{IEE}$ are asymptotically normal (see, for example, Rubin-Bleuer et al, 2006). Note that for the design consistency and asymptotic normality stated above it is not required that the marginal model be correct, but that there exists a vector β_0 such that the sequence of numbers $U_{IEE}(\beta_0) / M \rightarrow 0$ as $M \rightarrow \infty$.

3.3 The Survey Generalized Estimating Equation (GEE) Estimator of β_{GEE}

Let us now write $U_i^*(\beta) = \mathbf{D}'_i(\beta) \mathbf{A}_i^{-1/2}(\beta) \mathbf{R}^{*-1} \mathbf{A}_i^{-1/2}(\beta) (\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta))$, where

$\mathbf{R}^* = \mathbf{R}(\beta_{IEE})$, β_{IEE} is the solution of the census IEE as in (1.2) and $\mathbf{R}(\beta) = \sum_i \mathbf{R}_i(\beta) / M$ with

$$\mathbf{R}_i(\beta) = \mathbf{A}_i^{-1/2}(\beta) (\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta)) (\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta))' \mathbf{A}_i^{-1/2}(\beta)$$

We define the census GEE as in Rao (1998) by
$$\mathbf{U}_{GEE}(\boldsymbol{\beta}) = \sum_i U_i^*(\boldsymbol{\beta}) = 0, \quad (2.2)$$

and the survey GEE by
$$\hat{\mathbf{U}}_{GEE}(\boldsymbol{\beta}) = \sum_{i \in S} w_i \mathbf{D}'_i(\boldsymbol{\beta}) \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}) \hat{\mathbf{R}}^{*-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}) (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0, \quad (2.3)$$

where
$$\hat{\mathbf{R}}^* = \sum_{i \in S} w_i \mathbf{R}_i(\hat{\boldsymbol{\beta}}_{IEE}) / \hat{M} \quad \text{and} \quad \hat{M} = \sum_{i \in S} w_i.$$

The estimator $\hat{\boldsymbol{\beta}}_{GEE}$ is defined as the solution of the survey GEE (2.3). Rubin-Bleuer et al. (2005, 2006) showed that under certain regularity conditions, as $n \rightarrow \infty$, $\hat{\boldsymbol{\beta}}_{GEE}$ exists, it is consistent and asymptotically normal, with variance matrix which can be consistently estimated by $v(\hat{\boldsymbol{\beta}}_{GEE}) = \hat{\mathbf{I}}_G^{-1}(\hat{\boldsymbol{\beta}}_{GEE}) \cdot v(\sqrt{n} \hat{\mathbf{U}}_{GEE}) \cdot \hat{\mathbf{I}}_G^{-1}(\hat{\boldsymbol{\beta}}_{GEE})$, where $\hat{\mathbf{I}}_G(\boldsymbol{\beta}) = \sum_{i \in S} w_i \mathbf{D}'_i(\boldsymbol{\beta}) \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}) \hat{\mathbf{R}}^{*-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}) \mathbf{D}_i(\boldsymbol{\beta})$ and $v(\sqrt{n} \hat{\mathbf{U}}_{GEE})$, is a consistent estimator of the variance of the survey estimating equation, which is a total in the finite population.

We compare three different variance estimators. The ‘‘Naïve Liang & Zeger’’ (NLZ) variance estimator is:

$$\hat{\text{var}}_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE}) = [\tilde{\mathbf{I}}_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE})]^{-1} [\Gamma_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE})] [\tilde{\mathbf{I}}_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE})]^{-1}$$

where $\tilde{\mathbf{I}}_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE}) = \left[\sum_s \tilde{w}_i \mathbf{D}'_i(\hat{\boldsymbol{\beta}}_{GEE}) \hat{\mathbf{V}}_i^{-1}(\hat{\boldsymbol{\beta}}_{GEE}) \mathbf{D}_i(\hat{\boldsymbol{\beta}}_{GEE}) \right]$, $\tilde{w}_i = m \cdot w_i / \hat{M}$ are the normalized weights and

$$\hat{\mathbf{V}}_i^{-1}(\boldsymbol{\beta}) = \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}) \hat{\mathbf{R}}^{*-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}).$$

The NLZ variance estimator uses for the ‘‘filling’’ of the sandwich the following expression:

$$\Gamma_{\text{NLZ}}(\hat{\boldsymbol{\beta}}_{GEE}) = \left[\sum_s \tilde{w}_i \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right].$$

The ‘‘Proper Liang & Zeger’’ (PLZ) variance estimator, which assumes that the first stage clusters are either drawn with replacement in each stratum or the first stage sampling rates are negligible (see Rao 1998) is defined as:

$$\hat{\text{var}}_{\text{PLZ}}(\hat{\boldsymbol{\beta}}_{GEE}) = [\tilde{\mathbf{I}}_{\text{PLZ}}(\hat{\boldsymbol{\beta}}_{GEE})]^{-1} [\Gamma_{\text{PLZ}}(\hat{\boldsymbol{\beta}}_{GEE})] [\tilde{\mathbf{I}}_{\text{PLZ}}(\hat{\boldsymbol{\beta}}_{GEE})]^{-1}$$

where $\tilde{\mathbf{I}}_{\text{PLZ}}(\hat{\boldsymbol{\beta}}_{GEE}) = \left[\sum_s w_i \mathbf{D}'_i(\hat{\boldsymbol{\beta}}_{GEE}) \hat{\mathbf{V}}_i^{-1}(\hat{\boldsymbol{\beta}}_{GEE}) \mathbf{D}_i(\hat{\boldsymbol{\beta}}_{GEE}) \right]$.

The PLZ variance estimator takes into account the design by both using the survey weights and using the following ‘‘filling’’ (here the expression is easier to write with the *hik* notation, defined in section 3.1):

$$\Gamma_{\text{PLZ}}(\hat{\boldsymbol{\beta}}) = \sum_{h=1}^H \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left[\sum_{k=1}^{n_h} n_h w_{hik} \hat{\mathbf{U}}_{hik}^* - \frac{1}{n_h} \sum_i \sum_k n_h w_{hik} \hat{\mathbf{U}}_{hik}^* \right] \left[\sum_{k=1}^{n_h} n_h w_{hik} \hat{\mathbf{U}}_{hik}^* - \frac{1}{n_h} \sum_i \sum_k n_h w_{hik} \hat{\mathbf{U}}_{hik}^* \right]'$$

Finally the One-step Jackknife variance estimator is defined as in Rao and Tausi (2004):

$$v_{\text{JACK}}(\hat{\boldsymbol{\beta}}_{GEE}) = \left[\sum_{g=1}^L \frac{n_{g-1}}{n_g} \sum_{j=1}^{n_g} (\hat{\boldsymbol{\beta}}_{(gj)} - \hat{\boldsymbol{\beta}}_{GEE}) (\hat{\boldsymbol{\beta}}_{(gj)} - \hat{\boldsymbol{\beta}}_{GEE})' \right] \quad (2.4)$$

and $\hat{\boldsymbol{\beta}}_{(gj)}$ is the one step Newton-Raphson iteration when the (gj) th sample cluster is deleted and obtained from

$$\hat{\boldsymbol{\beta}}_{(gj)} = \hat{\boldsymbol{\beta}}_{GEE} + \hat{\mathbf{I}}_{(gj)}^{-1} (\hat{\boldsymbol{\beta}}_{GEE}) \hat{\mathbf{W}}_{(gj)} (\hat{\boldsymbol{\beta}}_{GEE})$$

Another version named EF (Estimated Function) Jackknife uses the calculation of the matrix \mathbf{I} over the whole sample and then

$$\hat{\mathbf{I}}_{(gj)}^{-1} (\hat{\boldsymbol{\beta}}) = \hat{\mathbf{I}}^{-1} (\hat{\boldsymbol{\beta}}) = \sum_s w_i \mathbf{D}_i' (\hat{\boldsymbol{\beta}}) \hat{\mathbf{V}}_i^{-1} (\hat{\boldsymbol{\beta}}) \mathbf{D}_i (\hat{\boldsymbol{\beta}}) \quad (2.5)$$

The EF Jackknife method avoids repeated inversions of possibly ill-conditioned matrices. We use the customary jackknife (JACK) defined in (2.4) for IEE and the EF jackknife (EFJACK) defined by (2.5) for GEE.

3. SIMULATION SETUP AND SAMPLING DESIGN

In order to have an empirical assessment of the impact of the variance estimation methods, the One-step Jackknife and the ‘‘Proper Liang & Zeger’’ using the GEE versus IEE approach on the analysis of ordinal longitudinal survey data we carried out an extensive simulation study. We simulated a model of the relationship between the self perceived health (SPH) of a person and several factors associated with health status.

The objective was to simulate $M=20000$ bivariate responses (SPH94, SPH96) and associated covariates such that the cumulative probabilities $\gamma_{ij} = pr(SPH_{it} \leq j)$ follow a proportional odds model as in (1.1).

For this purpose, we first fitted model (1.1) to data from the first two cycles ($t = 94/95$ and $t = 96/97$) of the Canadian National Population Health Survey (NPHS) where SPH had three possible responses: excellent/very good, (SPH=1) good (SPH=2) or fair/poor (SPH=3). For our simulated population we chose the covariates: AGE, EDU (completed high school), SELFESTEEM, DEPENDENT94 and DEPENDENT96. All covariates were binary: for example AGE=1 if subject is 55 or older in 1994 and AGE=0 otherwise. SELFESTEEM=1 if subject has low self-esteem in 1994 and SELFESTEEM=0 otherwise. EDU=1 if subject completed high school in 1994 and EDU=0 otherwise. DEPENDENT94=1 if subject needs help in meal preparation, shopping or in personal care in 1994 and DEPENDENT94=0 otherwise. DEPENDENT96 are similarly defined.

We set the super-population parameter $\boldsymbol{\beta}_0$ equal to the estimated $\boldsymbol{\beta}$ from the first two cycles of the NPHS sample using the IEE approach;

$$\boldsymbol{\beta}_0 = (\theta_1, \theta_2, \boldsymbol{\beta}'_s, \boldsymbol{\beta}_{94}, \boldsymbol{\beta}_{96})' = (1.31, 3.33, -0.51, -0.72, -0.86, -1.90, -1.91)'$$

where θ_1, θ_2 are the intercepts corresponding to SPH=1 and SPH=2 respectively, $\boldsymbol{\beta}'_s$ is a 3x1 vector of stationary parameters corresponding to the covariates AGE, EDU and SELFESTEEM, measured in 1994, and $\boldsymbol{\beta}_{94}$ and $\boldsymbol{\beta}_{96}$ are the time-varying covariates corresponding to dependent status.

The marginal model at time $t=94/95$ expressed the logit of the probabilities $\gamma_{i94j} = pr(SPH_{i94} \leq j)$, of an subject i choosing category j or less in 1994:

$$\log \left(\frac{pr[SPH_{i94} \leq j]}{1 - pr[SPH_{i94} \leq j]} \right) = \theta_j - 0.51 * AGE - 0.72 * EDU - 0.86 * SELFESTEEM - 1.90 * DEPENDENT94, \quad j = 1, 2 \quad (4.1)$$

Similarly, the marginal model at time $t=96/97$ expressed the logit of $\gamma_{i96j} = pr(SPH_{i96} \leq j)$:

$$\log \left(\frac{pr[SPH_{i96} \leq j]}{1 - pr[SPH_{i96} \leq j]} \right) = \theta_j - 0.51 * AGE - 0.72 * EDU - 0.86 * SELFESTEEM - 1.91 * DEPENDENT96, \quad j = 1, 2. \quad (4.2)$$

3.1 Simulated Finite Population

We created our finite population and the variables for each member in the following way:

1. We generated $M = 20000$ independent sets of the binary covariates AGE, EDU, SELFESTEEM, DEPENDENT94 and DEPENDENT96 stated in the model given by (4.1) and (4.2) as well as binary variables INCOME and GENDER to be used in the stratification and clustering of the population (INCOME =1 if income in 1994 was greater than the median income, and zero otherwise). It should be mentioned that in a real situation, income and gender could not be used as stratification or cluster variables since their values are most often not

known in advance. For this we followed the methodology used in Binder et al (2004), via the joint probability distributions of such variables estimated from the first two cycles of NPHS. Each set was associated with one subject in the finite population.

2. For each subject $i = 1, \dots, 20000$, we calculated the marginal cumulative probabilities γ_{ij} using equations (4.1) and (4.2). Note that from the cumulative probabilities we can calculate the marginal means $\mu_{ijt} = pr(SP_{H_{it}} = j)$.
3. To establish a correlation between the marginal responses we need to know the joint distribution $F_i(j, k) \equiv pr(SP_{H_{i94}} \leq j, SP_{H_{i96}} \leq k)$ between $SP_{H_{i94}}$ and $SP_{H_{i96}}$. We determined it through fixing the global odds ratio, which is as a 2 X 2 matrix whose elements are defined (see Williamson et al., 1995) as:

$$\Psi_i(j, k) = \frac{pr(SP_{H_{i94}} \leq j, SP_{H_{i96}} \leq k)pr(SP_{H_{i94}} > j, SP_{H_{i96}} > k)}{pr(SP_{H_{i94}} \leq j, SP_{H_{i96}} > k)pr(SP_{H_{i94}} > j, SP_{H_{i96}} \leq k)} = \frac{F_i(j, k)[1 - \gamma_{i94k} - \gamma_{i96k} + F_i(j, k)]}{[\gamma_{i94k} - F_i(j, k)][\gamma_{i96k} - F_i(j, k)]} \quad (4.3)$$

4. For every subject we defined a global odds ratio matrix as $\Psi_i = \begin{pmatrix} 100 & 5 \\ 5 & 100 \end{pmatrix}$, and calculated the joint distribution $F_i(j, k)$ via (4.3) and the known values γ_{ij} and Ψ_i , $i = 1, \dots, 20000$, $j, k = 1, 2, 3$.
5. From the bivariate distribution $F_i(j, k)$ we simulated the longitudinal responses as follows:

- For each $i = 1, \dots, 20000$, we simulated u_i and v_i as independent uniform (0,1) random variables; then we defined the responses as follows:
 - i. If $0 \leq u_i \leq F_i(1,3)$ then $SP_{H_{i94}} = 1$
 - ii. If $F_i(1,3) < u_i \leq F_i(2,3)$ then $SP_{H_{i94}} = 2$
 - iii. If $F_i(2,3) < u_i \leq F_i(3,3) = 1$ then $SP_{H_{i94}} = 3$
 - iv.
- Now given the response $SP_{H_{i94}} = j$, the conditional distribution of $SP_{H_{i96}}$ is defined by :

$$Pr(SP_{H_{i96}} \leq k / SP_{H_{i94}} \leq j) = \frac{F_i(j, k)}{F_i(j, 3)}, \quad j = 1, 2, 3.$$

- i. If $0 \leq v_i \leq F_i(j,1) / F_i(j,3)$ then $SP_{H_{i96}} = 1$
- ii. If $F_i(j,1) / F_i(j,3) < v_i \leq F_i(j,2) / F_i(j,3)$ then $SP_{H_{i96}} = 2$
- iii. If $F_i(j,2) / F_i(j,3) < v_i \leq F_i(j,3) / F_i(j,3) = 1$ then $SP_{H_{i96}} = 3$.

The finite population consists of the 20000 longitudinal responses thus defined and their associated covariates. The global odds ratio was defined so that the correlations between the responses in 1994 and 1996 are high:

$$R(\boldsymbol{\beta}_{IEE}) = \begin{pmatrix} 0.58 & 0.23 \\ 0.28 & 0.43 \end{pmatrix}.$$

We then fitted these data to the model given by (4.1) and (4.2) using both the IEE and GEE approaches to obtain the census parameters $\boldsymbol{\beta}_{IEE}$ and $\boldsymbol{\beta}_{GEE}$ respectively. We obtained

$$\boldsymbol{\beta}_{IEE}^t = (1.578, 3.194, -0.356, -0.708, -0.627, -2.071, -1.275) \quad \text{and} \quad \boldsymbol{\beta}_{GEE}^t = (1.689, 3.410, -0.510, -0.785, -0.760, -2.152, -1.264).$$

3.2 Clustering and Stratification of the Finite Population and Sample Design

We arranged the finite population into three strata in the following way. We created a fictional propensity score of “accessibility to medical cares facility” for each subject using the binary logistic model:

$$P(ACCESS) = \{1 + \exp(-[-2 + 10 * INCOME - 2.5 * GENDER - 1 * SP_{H_{94}}])\}^{-1},$$

We assigned subjects to each of 3 strata according to the value of their accessibility score: $\{p < 0.0125\}$, $\{0.0125 \leq p < 0.95\}$ and $\{p \geq 0.95\}$ so that we have an ‘informative design’. Then, we arranged the finite population in clusters (PSUs). The cluster sizes were between 30 and 50. In each stratum we randomly ordered the subject records and then assigned them to clusters whose sizes were generated as integers uniformly distributed between 30 and 50. We obtained a total 473 PSUs distributed into strata of sizes 154, 136 and 181 respectively. We looked at four stratified two-stage sample designs with probability proportional to cluster size with replacement in the first stage, with respective strata sample sizes: Design 1(10% sampling rate, 16, 14 and 18); Design 2(20% sampling rate, 32, 28 and 36), Design 3(sample sizes 75, 70 and 90) and Design 4(sample sizes 140, 120 and 170). The second stage was simple random sampling with replacement with sampling rate equal to 1/3 within each cluster chosen in the 1st stage. We conducted the simulation under each design setting with respectively 100, 1000 and 4000 Monte Carlo samples. In this paper we present a subset of our results and the rest will be presented in a more extensive article.

4. ASSESSMENT MEASURES AND RESULTS

We select n_s Monte Carlo samples from a design described above. For each sample $k = 1, \dots, n_s$ we estimate β_{IEE} with $\hat{\beta}_{IEE}(k)$ by solving the corresponding survey IEE and estimate β_{GEE} with $\hat{\beta}_{GEE}(k)$ by solving the corresponding survey GEE. We also estimate the variance $Var(\hat{\beta}_{IEE}(k))$ by $\hat{V}_A(k)$ using variance estimation approach A (A= NLZ, PLZ or JACK). Similarly we estimate $Var(\hat{\beta}_{GEE}(k))$. We calculated the following assessment measures for $\hat{\beta}_{IEE}$:

$$\text{Monte Carlo Expected Value: } E_{MC}(\hat{\beta}_{IEE}) = \frac{1}{n_s} \sum_{k=1}^{n_s} \hat{\beta}_{IEE}(k),$$

$$\text{Monte Carlo Variance: } V_{MC}(\hat{\beta}_{IEE}) = \frac{1}{n_s} \sum_{k=1}^{n_s} (\hat{\beta}_{IEE}(k) - E_{MC}(\hat{\beta}_{IEE}))^2,$$

$$\text{Monte Carlo Mean Square Error: } MSE_{MC}(\hat{\beta}_{IEE}) = \frac{1}{n_s} \sum_{k=1}^{n_s} (\hat{\beta}_{IEE}(k) - \beta_{IEE})^2,$$

$$\% \text{Relative Bias of the estimator: } \% \text{RelBias of } (\hat{\beta}_{IEE}) = \frac{E_{MC}(\hat{\beta}_{IEE}) - \beta_{IEE}}{\beta_{IEE}} \times 100$$

$$\% \text{Relative Bias of the variance estimator: } \% \text{RelBias}(\hat{V}_A(\hat{\beta}_{IEE})) = \frac{\sum_{k=1}^{n_s} \hat{V}_A(k) / n_s - V_{MC}(\hat{\beta}_{IEE})}{V_{MC}(\hat{\beta}_{IEE})} \times 100$$

$$\% \text{Coefficient of Total Error: } \% \text{cte}(\hat{V}_A(\hat{\beta}_{IEE})) = \frac{\left(\frac{1}{n_s} \sum_{k=1}^{n_s} [\hat{V}_A(k) - V_{MC}(\hat{\beta}_{IEE})]^2 \right)^{1/2}}{V_{MC}(\hat{\beta}_{IEE})} \times 100.$$

Similarly, we calculated $E_{MC}(\hat{\beta}_{GEE})$, $V_{MC}(\hat{\beta}_{GEE})$, $MSE_{MC}(\hat{\beta}_{GEE})$, $\% \text{RelBias}(\hat{V}_A(\hat{\beta}_{GEE}))$, $\text{RelBias}(\hat{\beta}_{GEE})$ and $\% \text{cte}(\hat{V}_A(\hat{\beta}_{GEE}))$.

Tables 2 and 3 below show results for GEE and IEE under Design 2, respectively. Table 4 compares the MSE of GEE estimator versus IEE estimator. Due to the fact that the % cte of the PLZ and EFJACK variance estimators were high (up to 49% for GEE, see Table 2) we conducted another simulation with a higher sampling rate (Design 3) and 1000 iterations. Results of this simulation are shown in Table 5.

Table 2 RESULTS FOR GEE - Design 2 (20% sampling rate, 3 strata) , $n_s = 1000$

	θ_1	θ_2	AGE	EDU	SELF ESTEEM	DEPENDENT94	DEPENDENT96
% Relative Bias of $\hat{\beta}_{GEE}$	-0.07	1.09	-0.15	0.96	-0.44	-0.06	1.02
% Relative Bias of NLZ	-93.16	-89.41	-36.22	-45.37	-12.76	-35.42	-79.78
% cte of NLZ	93	89	37	46	16	36	80
% Relative Bias of PLZ	-2.22	16.94	-17.43	5.97	-3.87	16.66	2.24
% cte of PLZ	13	28	26	20	21	49	16
% Relative Bias of EF JACK	-2.47	17.92	-17.86	5.68	-4.60	15.17	1.49
% cte of EF JACK	9	31	26	20	19	44	16

Table 3 RESULTS FOR IEE - Design 2 (20% sampling rate, 3 strata) $n_s = 1000$

	θ_1	θ_2	AGE	EDU	SELF ESTEEM	DEPENDENT94	DEPENDENT96
% Relative Bias of $\hat{\beta}_{IEE}$	-0.37	1.17	-0.92	-0.11	-0.72	-0.33	0.03
% Relative Bias of NLZ:	-94.27	-94.11	-24.81	-49.74	-8.88	-38.41	-78.55
% cte of NLZ	94	94	26	50	14	39	79
% Relative Bias of PLZ:	0.68	-1.15	-0.58	0.51	-0.85	4.89	4.61
% cte of PLZ	8	16	28	22	26	30	15
% Relative Bias of JACK	2.89	4.14	1.21	3.64	0.30	0.93	6.76
% cte of JACK	9	20	29	24	27	31	16

Table 4 MSE OF IEE VS MSE OF GEE - Design 2 (20% sampling rate, 3 strata) $n_s = 1000$

	θ_1	θ_2	AGE	EDU	SELF ESTEEM	DEPEN DENT94	DEPEN DENT 96
$\% \left(\text{MSE}_{MC}(\hat{\beta}_{IEE}) / \text{MSE}_{MC}(\hat{\beta}_{GEE}) \right)$	106.54	161.61	76.77	97.43	84.99	80.88	69.91

We notice that Table 4 shows efficiency gains of GEE over IEE only for the intercepts θ_1 and θ_2 . With design 3, as with the previous designs, we obtained low relative bias for the parameter estimators (not reported here), and high relative bias for the PLZ and EFJACK variance estimators (up to -20%, see Table 5). However the % cte of the respective variance estimators is slightly lower (see Table 5). This would be consistent with lower variability of the variance for larger sample sizes while the relative bias remains unchanged. Our simulation shows that for data obtained under the type of designs used here, the “Naïve Liang & Zeger” variance estimator performs badly and that the “Proper Liang & Zeger” and the One-step Jackknife variance estimators are similar and both perform much better than the “Naïve Liang & Zeger” in terms of relative bias and variability. In the GEE case, we can observe in table 5 that the percent relative bias of the “Proper Liang & Zeger” and EF Jackknife variance estimators could range between - 20% and +17%.

Table 5 RESULTS FOR GEE : Design 3_(3 strata) , $n_s = 1000$

	θ_1	θ_2	AGE	EDU	SELF ESTEEM	DEPEND ENT 94	DEPEND ENT 96
% Relative Bias of PLZ:	-7.10	6.27	-19.93	4.92	7.01	16.91	5.04
% cte of PLZ:	10.6	14.3	23.2	14.0	17.2	32.9	15.9
% Relative Bias of EF JACK	-7.16	6.62	-20.14	5.15	7.29	16.98	4.67
% cte of EF JACK:	9.3	15.4	23.4	14.1	16.9	31.1	15.4

6. CONCLUSION

The low relative bias of $\hat{\beta}_{IEE}$ and $\hat{\beta}_{GEE}$ indicates that both the IEE and GEE approaches estimate well the corresponding finite population parameters. To analyse the relative biases observed in the estimator of the variance in the GEE case, let us first make a few remarks about model-based estimation. The finite population was simulated so that the longitudinal correlations of both the responses and covariates were high. The census GEE assumes a ‘working’ constant correlation matrix across subjects. Our simulated population has varying correlation matrices, even if the global odds matrix is constant. Thus the working correlation matrix does not coincide with the model longitudinal correlation. Liang and Zeger (1986) show that when the longitudinal correlation was high and the working model was different from the true longitudinal correlation, the IEE approach was more efficient than the GEE approach. Here we focus on the sampling variability of $\hat{\beta}_{IEE}$ and $\hat{\beta}_{GEE}$. Under the GEE approach, the census estimating equation has a correlation matrix R that is constant across subjects. The survey GEE has a ‘working’ correlation matrix \hat{R} that is design-consistent for R . In spite of this, we observe high relative bias in the PLZ and Jackknife variance estimators and this could be attributed to the difference between the working model for the covariance in the census GEE and the super-population longitudinal covariance matrix. The numbers obtained point to a sensitivity of the variance estimator when the longitudinal correlation is high and the working model for this correlation does not coincide with the super-population correlation. If this thesis is correct, the results are not inconsistent with those of Liang and Zeger (1986) for a high longitudinal correlation and an incorrect model for the longitudinal covariance. The low relative bias obtained for the estimators of the variance of the survey IEE estimator could indicate, as with model-based estimation, that in the presence of high longitudinal correlation it is more efficient to work with the IEE approach. But it is somewhat surprising that under a “purely” design premise the super-population longitudinal model still influences the estimation of the sampling variation of the finite population parameter. We plan to continue with this investigation by comparing through simulation the GEE versus the IEE approaches when the marginal model is correct and the covariates are not longitudinally correlated, and when the marginal model is not correct.

REFERENCES

- Agresti, A. and Natarajan, R. (2001). Modeling Clustered Ordered categorical data: A Survey. *International Statistical Review*, **69**, 3, 345-371. Printed in the Netherlands.
- Binder, D., Kovacevic, M., and Roberts, G. (2004). Design Based Methods for Survey Data: Alternative uses of estimating Functions. *2004 Proceedings of the Section on Survey Research Methods of the American Statistical Association*.
- Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 1, 13-22.
- Rao, J.N.K. (1998), Marginal Models for Repeated Observations: Inference with Survey Data. *1998 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 76-82.
- Rao, J.N.K. and Tausi, M. (2004), Estimation Function Jackknife Variance Estimators Under Stratified Multistage Sampling. *Communications in Statistics theory and methods*, Vol **33**, No 9, 2087-2095, 2004
- Shields, M. and Shooshtari, S. (2001). Determinants of self-perceived health. *Health Reports*, Vol **13**, No.1, 35-64, December 2001 Statistics Canada, catalogue 82-003.
- Rubin-Bleuer, S., Saïdi, A. and Kovacevic, M. (2005). “Analysis of Ordinal Longitudinal Survey Data”, *Proceedings of the 55th Session of the International Statistical Institute*, (invited paper), CDROM, ISI2005.
- Rubin-Bleuer, S., Saïdi, A. and Kovacevic, M. (2006). Analyse des données d’enquêtes longitudinales, Méthodes d’enquêtes et sondages sous la direction de Lavallée, P. et Rivest, L.P., Eds Dunod.
- Rubin-Bleuer, S., Saïdi, A. and Kovacevic, M. (2006). “Analysis of Ordinal Longitudinal Survey Data”, *Statistics Canada Series. Methodology Branch*, BSMD, 2006 (to appear).
- Sutradhar, B.C. and Kovacevic, M. (2000). Analysing ordinal longitudinal survey data: Generalized estimating equations approach. *Biometrika*, **87**, 4, 837-848.
- Williamson, J.M., Kim K. and Lipsitz, S.R (1995). Analyzing Bivariate ordinal Data Using a Global Odds Ratio, *Journal of the American Statistical Association*, Vol **90**, No. 432, *Theory and Methods*, 1432-1437.