The Canadian Journal of Statistics
Vol. 42, No. 3, 2014, Pages 404–422
La revue canadienne de statistique

Longitudinal data analysis using the conditional empirical likelihood method

Peisong HAN[1]*, Peter X.-K. SONG[2] and Lu WANG[2]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.*

Key words and phrases: Generalized estimating equations (GEE); marginal model; unbalanced longitudinal data; variance–covariance matrix; within-subject correlation.

MSC 2010: Primary 62F12; secondary 62J12

Abstract: This paper studies a new approach to longitudinal data analysis using the conditional empirical likelihood (CEL) method within the framework of marginal models. The possible unbalanced follow-up visits are dealt with via stratification according to distinctive follow-up patterns. The CEL method does not require any explicit modelling of the variance–covariance of the longitudinal outcomes. Instead, it implicitly incorporates a consistently estimated variance–covariance matrix in a nonparametric fashion. The proposed CEL estimator is connected to the generalized estimating equations (GEE) estimator, and achieves the same efficiency as the GEE estimator employing the true variance–covariance. The asymptotic distribution of the CEL estimator is derived, and simulation studies are conducted to assess the finite sample performance. Data collected from a longitudinal nutrition study are analysed as an application. *The Canadian Journal of Statistics* 42: 404–422; 2014 © 2014 Statistical Society of Canada

Résumé: Les auteurs proposent une nouvelle approche pour l'analyse de données longitudinales à l'aide de la méthode de la vraisemblance empirique conditionnelle (VEC) dans le cadre de modèles marginaux. Ils prennent en compte la possibilité d'un suivi irrégulier en stratifiant selon les séquences de suivis observées. La VEC ne nécessite pas la modélisation explicite de la variance-covariance des résultats longitudinaux, mais en intègre plutôt implicitement un estimateur non paramétrique convergent. La VEC est associée aux équations d'estimation généralisées (EEG), et les estimateurs découlant de la VEC atteignent la même efficacité que ceux des EEG basées sur la vraie structure de variance-covariance. Les auteurs présentent la distribution asymptotique de l'estimateur de la VEC, ainsi qu'une étude de simulation afin d'évaluer la performance de la méthode sur des échantillons finis. Ils effectuent finalement l'analyse des données d'une étude longitudinale portant sur la nutrition. *La revue canadienne de statistique* 42: 404–422; 2014 © 2014 Société statistique du Canada

1. INTRODUCTION

Longitudinal data are repeated measurements collected from the study subjects over time. To efficiently evaluate the relationship between the mean of the longitudinal outcomes and the covariates, the variance–covariance of these outcomes needs to be taken into account. Marginal models for longitudinal data, which specify model structures for both the marginal mean and the variance–covariance, have been discussed widely in the literature (e.g., Diggle et al., 2002; Song, 2007). Estimation and inference for marginal models may be based on, for example, the quasi-likelihood approach (Wedderburn, 1974; Heyde, 1997) or the estimating functions approach (Godambe, 1960, 1991). Liang & Zeger (1986) proposed the popular generalized estimating equations (GEE) method, which assumes a working model for the within-subject

* *Author to whom correspondence may be addressed.*
E-mail: peisonghan@uwaterloo.ca

correlation of the longitudinal outcomes. When the correlation is incorrectly modelled, the GEE estimator preserves consistency, but may lose substantial estimation efficiency (Wang & Carey, 2003). To improve the efficiency, Qu, Lindsay, & Li (2000) proposed the quadratic inference functions (QIF) estimator, which has been shown to be equally efficient to the GEE estimator when the correlation is correctly modelled, and more efficient when the correlation is incorrectly modelled. Wang & Lin (2005) pointed out that, to avoid losing efficiency, not only the correlation, but also the marginal variance of the longitudinal outcomes needs to be correctly modelled. Some strategies for modelling the second moments can be found in Pan & MacKenzie (2003) and Ye & Pan (2006). However, the introduction of more models beyond that for the marginal mean presents more risk of model misspecification, which can lead to poor efficiency and misleading conclusions. Therefore, many researchers proposed to model the variance–covariance nonparametrically; see, for example, Jiang, Luan, & Wang (2007) and Li (2011).

Conditional empirical likelihood (CEL) (Zhang & Gijbels, 2003; Kitamura, Tripathi, & Ahn, 2004) is a generalization of empirical likelihood (EL) method (Owen, 1988, 1990, 2001; Qin & Lawless, 1994) to the setting where the model is defined by conditional moment restrictions. Despite the success of EL method in various research areas, its application in longitudinal data analysis has not been studied adequately. Some existing works include You, Chen, & Zhou (2006) and Xue & Zhu (2007), who studied partially linear models for longitudinal data. However, those authors did not account for the within-subject correlation. For data with continuous outcomes, Wang, Qian, & Carroll (2010) proposed a generalized empirical likelihood (GEL) method, which requires a pre-estimated variance–covariance matrix. To the best of our knowledge, all existing applications of EL method to longitudinal data analysis follow the setting of Qin & Lawless (1994), where the model is defined by unconditional moment restrictions.

In this paper, we propose to model longitudinal data using the CEL method, which only requires a model for the marginal mean of the outcomes, and thus eliminates the dependence of data analysis on modelling the variance–covariance. Possible unbalanced follow-up visits are dealt with via stratification according to distinctive follow-up patterns. We show that the CEL method is closely connected to the GEE method, in the sense that the proposed estimator can be regarded as the solution to a set of estimating equations analogous to GEE. The difference is that, the CEL method implicitly incorporates a consistently estimated variance–covariance matrix, which, in theory, endows the proposed estimator the same efficiency as that of the GEE estimator employing the true variance–covariance. We also derive the asymptotic distribution of the proposed estimator, and provide two ways to estimate the asymptotic variance.

This article is organized as follows. In Section 2, we discuss longitudinal data analysis using the CEL method. Section 3 contains the large sample properties. Section 4 presents numerical implementation and some discussion on computational issues. In Section 5, we conduct simulation experiments to study the finite sample performance of the CEL method. Section 6 contains a data application. Section 7 consists of concluding remarks. Technical assumptions and proofs, together with some extra simulation results, are provided in the online Supplementary Materials.

2. METHODOLOGY

2.1. Model and Estimation

For each subject i, $i = 1, \ldots, N$, let Y_{it} denote the outcome and $\boldsymbol{X}_{it}$ denote a p-element vector of covariates measured at time $t = 1, \ldots, n_i$. Write $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^\top, \ldots, \boldsymbol{X}_{in_i}^\top)^\top$. We only specify a model for the marginal mean of Y_{it} given $\boldsymbol{X}_i$. Specifically, we assume the marginal mean regression model

$$E(Y_{it} \mid \boldsymbol{X}_i) = \mu(\boldsymbol{X}_{it}^\top \boldsymbol{\beta}) \quad \text{for some } \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathcal{R}^p,$$

where $\mu(\cdot)$ is a known link function and $\boldsymbol{\beta}$ is the vector of the regression coefficients. This model assumption is also adopted by the GEE and the QIF methods, and implies that the time-dependent covariates are of Type I according to Lai & Small (2007). See also Pepe & Anderson (1994). Define the residual vector to be $\boldsymbol{g}_i(\boldsymbol{\beta}) = \{Y_{i1} - \mu(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}), \ldots, Y_{in_i} - \mu(\boldsymbol{X}_{in_i}^\top \boldsymbol{\beta})\}^\top$. Our model assumption becomes

$$E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}, \quad i = 1, \ldots, N. \tag{1}$$

The above setting accommodates unbalanced longitudinal data, which arise in many practical studies. For example, in a clinical trial study using a balanced design, where all subjects are scheduled to visit the clinic at the same pre-fixed time points, unbalanced data may result due to subjects' missing visits. Therefore, unbalanced data can be treated within the framework of missing data analysis. In this case, similar to the GEE method, our formulation implicitly makes the missing completely at random (MCAR) assumption (Little & Rubin, 2002). This assumption is also widely adopted in longitudinal observational studies where the data are collected at available visits with no clear missingness mechanism. Following the GEE method, we focus our discussion on available-data analysis without modelling the missingness mechanism. Such an approach is the default for many popular statistical software in analyzing unbalanced longitudinal data. Note that the visit time is not informative with respect to $\boldsymbol{\beta}$.

To deal with the unbalanced data structure, we stratify the subjects according to their follow-up patterns. Let $\tau_i = \{1, \ldots, n_i\}$ denote the (intermittent) follow-up pattern for subject i, and let $S_i = \{j : 1 \le j \le N \text{ and } \tau_j = \tau_i\}$ be the stratum where subject i belongs. Let Ω denote the collection of all distinctive follow-up patterns observed in the data, and for $\omega \in \Omega$, let S^ω denote the set of indices for the subjects having pattern ω. Thus, $\{S^\omega\}_{\omega\in\Omega}$ constitutes the collection of all distinctive strata in the data. The model defined by (1) has a stratum-wise interpretation: for any $\omega \in \Omega$, $E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}$ for $i \in S^\omega$. Clearly, when the whole data are balanced, all subjects belong to a single stratum.

Given each subject i, we consider the empirical probabilities p_{ij} defined by a discrete distribution that has support on $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$. Intuitively, p_{ij} may be interpreted as the (discrete) probability of observing the value of the residual from subject j conditional on the value of the covariates from subject i, where j and i are in the same stratum. Based on these conditional empirical probabilities, we construct a localized (or subject-wise) empirical log-likelihood $\sum_{j\in S_i} w_{ij} \log p_{ij}$ for subject i, where the localization is carried out by certain non-negative weights w_{ij}. A "smoothed" empirical log-likelihood can then be obtained as $\sum_{i=1}^N \sum_{j\in S_i} w_{ij} \log p_{ij}$. On the other hand, the empirical version of the model assumption (1) employing the conditional empirical probabilities is given by $\sum_{j\in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0}$ for certain values of $\boldsymbol{\beta}$. Therefore, an analogue of the (maximum) likelihood function of $\boldsymbol{\beta}$ can be obtained through the following constrained maximization:

$$\begin{aligned} &\max_{p_{ij}} \sum_{i=1}^N \sum_{j\in S_i} w_{ij} \log p_{ij} \quad \text{subject to} \\ &p_{ij} \ge 0 \ (j \in S_i), \ \sum_{j\in S_i} p_{ij} = 1, \ \text{and} \ \sum_{j\in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0} \qquad \text{for all } i = 1, \cdots, N. \end{aligned} \tag{2}$$

Here, the first two constraints make sure that, conditional on each $i = 1, \ldots, N$, p_{ij} are well-defined empirical probabilities. Our CEL estimator of $\boldsymbol{\beta}_0$ is defined to be the value of $\boldsymbol{\beta}$ that maximizes the above maximum empirical likelihood:

$$\hat{\boldsymbol{\beta}}_{\text{CEL}} = \arg\max_{\boldsymbol{\beta}} \max_{p_{ij}} \sum_{i=1}^N \sum_{j\in S_i} w_{ij} \log p_{ij} \quad \text{subject to (2).}$$

A technique to achieve localization is the nonparametric kernel method. Write $\boldsymbol{X}_i = (\boldsymbol{X}_i^c, \boldsymbol{X}_i^d)$, where $\boldsymbol{X}_i^c$ consists of the continuous components with dimension q_i and $\boldsymbol{X}_i^d$ consists of the categorical components, respectively. Time itself as a covariate may be excluded from the calculation of the weights, since subjects from the same stratum have a common set of visit times. The weights w_{ij} are calculated as

$$w_{ij} = \frac{\mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j \in S_i} \mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \quad j \in S_i, \tag{3}$$

where $\mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\} = \prod_{l=1}^{q_i} K\{(X_i^{(l)} - X_j^{(l)})/b_N\}$, $K(\cdot)$ is a second order kernel function and symmetric around zero (e.g. standard Gaussian kernel), b_N is the bandwidth parameter, and $X_i^{(l)}$ denotes the lth component of $\boldsymbol{X}_i^c$. Note that $\sum_{j \in S_i} w_{ij} = 1$ for each i.

According to Owen (1988), one has to be aware of the "empty set problem" (e.g., Grendár & Judge 2009), which refers to the situation where $\mathbf{0}$ may not be in the convex hull spanned by $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$. In this case, the profile CEL in (2), and thus $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$, may be ill-defined because constraints in (2) may not be satisfied. For the scenario of estimating a population mean, Owen (1988) proved that the profile empirical likelihood is well-defined at the true parameter value with probability one under mild conditions. In our setting, since $E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}$ for any $\omega \in \Omega$ and $i \in S^\omega$, $\mathbf{0}$ should be inside the convex hull of $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$ for $i = 1, \ldots, N$, at least when N is large, the joint distribution of $(\boldsymbol{Y}^\top, \boldsymbol{X}^\top)$ is smooth enough and $\boldsymbol{\beta}$ is inside a small neighborhood of $\boldsymbol{\beta}_0$. However, specification of exact conditions under which the "empty set problem" disappears asymptotically for a general model defined by conditional moment restrictions is challenging. We will formally investigate this issue as future research topics. In this paper, for theoretical derivation, we will assume that $\mathbf{0}$ is inside the convex hull of $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$ so that $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is well-defined. For numerical implementation when this is not the case, we discuss some possible solutions in Section 4.

The use of a common bandwidth parameter b_N for different strata as in (3) is reasonable, since each stratum can be assumed to have size proportional to N without loss of generality; see Section 3 for more discussions. Results from Smith (2007) may serve as a rule of thumb for the order of b_N. That is, $b_N \to 0$, $N^{1-2\nu-2/\delta} b_N^{2q} \to \infty$ and $N^{1-2\nu} b_N^{5q/2} \to \infty$ as $N \to \infty$, where $\nu \in (0, 1/2)$, $\delta \geq 8$ and $q = \max_i q_i$. In practice, a data-driven bandwidth selection procedure is often desired. In this paper we implement a cross-validation criterion suggested by Newey (1993) for models defined by conditional moment restrictions, with some modifications to suit for the context of (possibly) unbalanced longitudinal data. Specifically, define

$$CV(b_N) = \mathrm{tr}\left\{\sum_{i=1}^{N} \hat{\boldsymbol{G}}_i^\top \hat{\boldsymbol{V}}_{-i}^{-1} \left(\boldsymbol{g}_i \boldsymbol{g}_i^\top - \hat{\boldsymbol{V}}_{-i}\right) \hat{\boldsymbol{V}}_{-i}^{-1} \left(\boldsymbol{g}_i \boldsymbol{g}_i^\top - \hat{\boldsymbol{V}}_{-i}\right) \hat{\boldsymbol{V}}_{-i}^{-1} \hat{\boldsymbol{G}}_i\right\}, \tag{4}$$

where $\hat{\boldsymbol{G}}_i = \partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}})/\partial \boldsymbol{\beta}$, $\hat{\boldsymbol{V}}_{-i} = \sum_{j \in S_i} \hat{w}_{ij} \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}})^\top$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(b_N)$ is the CEL estimator obtained with a given b_N, and

$$\hat{w}_{ii} = 0, \quad \hat{w}_{ij} = \frac{\mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j \in S_i, j \neq i} \mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \quad \text{for } j \in S_i \text{ and } j \neq i.$$

The optimal bandwidth b_N is chosen as the minimizer of $CV(b_N)$.

2.2. Reformulation and Connection to GEE

Calculating $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ requires us to solve a constrained optimization problem. Using the Lagrange multipliers method, define the Lagrangian as

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{j \in S_i} w_{ij} \log p_{ij} \right) - \sum_{i=1}^{N} \varpi_i \left(\sum_{j \in S_i} p_{ij} - 1 \right) - \sum_{i=1}^{N} \boldsymbol{\lambda}_i^{\top} \left\{ \sum_{j \in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\},$$

where scalar $\varpi_i \in \mathcal{R}$ and vector $\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}$ are the Lagrange multipliers associated with the second and third constraints in (2), respectively. Taking $\partial \mathcal{L} / \partial p_{ij} = 0$, together with the constraints in (2), it can be easily shown that, for a fixed $\boldsymbol{\beta}$,

$$p_{ij}(\boldsymbol{\beta}) = \frac{w_{ij}}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^{\top} \boldsymbol{g}_j(\boldsymbol{\beta})}, \quad j \in S_i, \quad i = 1, \ldots, N, \tag{5}$$

where $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ is the root of

$$\sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0}. \tag{6}$$

Note that the validity of (5) and (6) depends on the correctness of our model assumption (1).

It is easy to see that

$$\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \arg \min_{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}} \left[- \sum_{j \in S_i} w_{ij} \log \left\{ 1 + \boldsymbol{\lambda}_i^{\top} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\} \right] \stackrel{\mathrm{def}}{=} \arg \min_{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}} \Lambda_i(\boldsymbol{\lambda}_i, \boldsymbol{\beta}), \tag{7}$$

where $\Lambda_i(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) = - \sum_{j \in S_i} w_{ij} \log \left\{ 1 + \boldsymbol{\lambda}_i^{\top} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\}$. Thus, the objective function in (2) can be rewritten as a function of $\boldsymbol{\beta}$ only, namely

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j \in S_i} w_{ij} \log p_{ij} = \sum_{i=1}^{N} L_i(\boldsymbol{\beta}) + \sum_{i=1}^{N} \sum_{j \in S_i} w_{ij} \log w_{ij}, \tag{8}$$

where $L_i(\boldsymbol{\beta}) = \Lambda_i\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\}$. Therefore, the CEL estimator can be equivalently defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{CEL}} = \arg \max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\{ \min_{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}} \Lambda_i(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) \right\}. \tag{9}$$

Equation (9) provides a way to implement the CEL method, which will be detailed in Section 4.

As seen, the CEL method does not require to explicitly model the variance–covariance matrix $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. However, this matrix is taken into account implicitly in the optimization procedure (7) through the use of probabilities p_{ij}. To elaborate, note that the third constraint in (2) implies

$$\mathbf{0} = \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \left\{ 1 - \frac{w_{ij}}{p_{ij}(\boldsymbol{\beta})} \right\} \boldsymbol{g}_j(\boldsymbol{\beta}) + \sum_{j \in S_i} w_{ij} \boldsymbol{g}_j(\boldsymbol{\beta})$$

and from (5) we have

$$1 - \frac{w_{ij}}{p_{ij}(\boldsymbol{\beta})} = - \boldsymbol{g}_j^{\top}(\boldsymbol{\beta}) \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \quad j \in S_i.$$

Therefore, solving for $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ leads to

$$\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \left\{ \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta}) \boldsymbol{g}_j^\top(\boldsymbol{\beta}) \right\}^{-1} \left\{ \sum_{j \in S_i} w_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\}. \tag{10}$$

From the proof of Theorem 3 in Section 3, the term $\sum_{j \in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top$ is a consistent estimator of $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. This implies that a consistent estimator of $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ is involved in (10), which is the optimal solution to (7). Due to the positive definiteness of $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$, $\sum_{j \in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top$ is positive definite, at least when N is large. Note that $\sum_{j \in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top$ is a data-driven estimate of the variance–covariance matrix, and thus the CEL method has the flexibility of allowing the variance–covariance to be different across individuals, unlike the GEE method, which requires a common within-subject correlation structure, such as AR-1 or compound symmetry.

Since $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ maximizes $L(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is the solution to $\partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$. From (8) and the fact that $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ is the root of $\sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is the solution to

$$\sum_{i=1}^{N} \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{G}_j(\boldsymbol{\beta})^\top \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \mathbf{0},$$

where $\boldsymbol{G}_j(\boldsymbol{\beta}) = \partial \boldsymbol{g}_j(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Plugging $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ given by (10) into the above equation leads to

$$\sum_{i=1}^{N} \left\{ \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{G}_j(\boldsymbol{\beta})^\top \right\} \left\{ \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta})^\top \right\}^{-1} \left\{ \sum_{j \in S_i} w_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\} = \mathbf{0}.$$

Being the solution to the above estimating equation, $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ can be regarded as a GEE-type estimator, where the matrix $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ is estimated nonparametrically by using the empirical probabilities p_{ij} rather than parametrically by assuming a working model. Because $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is obtained under a consistently estimated $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$, it in theory will have improved efficiency over the GEE estimator employing an incorrectly modelled variance–covariance matrix.

3. LARGE SAMPLE PROPERTIES

To facilitate the presentation, we introduce some extra notation. Noting that $\boldsymbol{G}_i(\boldsymbol{\beta})$ is a function of covariates only, we have $\boldsymbol{G}_i(\boldsymbol{\beta}) = E\{\partial \boldsymbol{g}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \mid \boldsymbol{X}_i\}$. Denote $\boldsymbol{V}_i(\boldsymbol{\beta}) = E\left\{\boldsymbol{g}_i(\boldsymbol{\beta}) \boldsymbol{g}_i(\boldsymbol{\beta})^\top \mid \boldsymbol{X}_i\right\}$. For any set A, let $|A|$ denote the cardinality of A. Define $r_\omega = \lim_{N \to \infty} |S^\omega|/N$. Without loss of generality, we assume that $r_\omega > 0$ for all $\omega \in \Omega$, as a stratum with $r_\omega = 0$ should be asymptotically negligible compared to the strata that have size proportional to N, and this will be formally illustrated in the proofs of the following theorems. Technical assumptions and proofs for the theoretical results presented in this section are provided in the online Supplementary Materials.

The consistency of the proposed CEL estimator is established in the following Theorem 1.

Theorem 1. *Under Assumptions (i)–(vii) in the online Supplementary Materials, we have* $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$.

The following Theorem 2 gives the asymptotic distribution of the CEL estimator.

Theorem 2. *Under Assumptions (i)–(viii) in the online Supplementary Materials, we have* $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_0)^{-1}\}$ *as* $N \to \infty$, *where* $\boldsymbol{J}(\boldsymbol{\beta}_0) = \sum_{\omega\in\Omega} r_\omega \boldsymbol{J}_\omega(\boldsymbol{\beta}_0)$, *and for all* $i \in S^\omega$, $\boldsymbol{J}_\omega(\boldsymbol{\beta}_0) = E\left\{\boldsymbol{G}_i(\boldsymbol{\beta}_0)^\top \boldsymbol{V}_i(\boldsymbol{\beta}_0)^{-1}\boldsymbol{G}_i(\boldsymbol{\beta}_0)\right\}$.

In the case of balanced longitudinal data, all subjects belong to a unique stratum, and thus we have $\boldsymbol{J}(\boldsymbol{\beta}_0) = E\left\{\boldsymbol{G}(\boldsymbol{\beta}_0)^\top \boldsymbol{V}(\boldsymbol{\beta}_0)^{-1}\boldsymbol{G}(\boldsymbol{\beta}_0)\right\}$, which is the semiparametric efficiency bound for the model defined by (1) (Chamberlain, 1987). Therefore, $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ has optimal efficiency amongst all regular and asymptotically linear estimators under model (1) with balanced data.

To consistently estimate $\boldsymbol{J}(\boldsymbol{\beta}_0)$, we consider two approaches. The first approach is based on Lemma C1 in Kitamura, Tripathi, & Ahn (2004). As one can show that, for any $\omega \in \Omega$,

$$-\frac{1}{|S^\omega|}\frac{\partial^2 L^\omega}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \xrightarrow{p} \boldsymbol{J}_\omega(\boldsymbol{\beta}_0)$$

as $N \to \infty$, where $L^\omega(\boldsymbol{\beta}) = \sum_{i\in S^\omega} L_i(\boldsymbol{\beta})$, a consistent estimator of $\boldsymbol{J}(\boldsymbol{\beta}_0)$ takes the following form that is similar to an observed information matrix:

$$-\frac{1}{N}\frac{\partial^2 L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}). \tag{11}$$

The derivatives in (11) are taken with respect to $\boldsymbol{\beta}$ holding $\boldsymbol{\lambda}_i$ as constant. The calculation of (11) is straightforward, and is a byproduct of the Newton–Raphson algorithm searching for $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$. See more discussions in Section 4.

The second approach to estimating $\boldsymbol{J}(\boldsymbol{\beta}_0)$ is given by the following Theorem 3.

Theorem 3. *Under Assumptions (i)–(ix) in the online Supplementary Materials, as* $N \to \infty$, *we have*

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top\left\{\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top\right\}^{-1}\boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}) \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\beta}_0). \tag{12}$$

The above estimator is a simplification of the well-known "sandwich" estimator (e.g., Liang & Zeger, 1986)

$$\left\{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i^\top\boldsymbol{V}_{\mathrm{w},i}^{-1}\boldsymbol{G}_i\right\}^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i^\top\boldsymbol{V}_{\mathrm{w},i}^{-1}\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)\boldsymbol{V}_{\mathrm{w},i}^{-1}\boldsymbol{G}_i\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i^\top\boldsymbol{V}_{\mathrm{w},i}^{-1}\boldsymbol{G}_i\right\}^{-1},$$

where $\boldsymbol{V}_{\mathrm{w}}$ is a working variance–covariance matrix. For the CEL method, $\boldsymbol{V}_{\mathrm{w},i} = \sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}})^\top$ consistently estimates $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. Therefore, the variability matrix and the sensitivity matrix in the "sandwich" estimator cancel each other asymptotically, yielding the estimator given in (12).

Although both estimators (11) and (12) are consistent, they may perform differently under finite sample size. A numerical comparison and some practical suggestions will be given in the simulation studies in Section 5.

Finally, the optimality of $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is established by the following Theorem 4.

Theorem 4. *Under Assumptions (i)–(viii) in the online Supplementary Materials, we have* $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{CEL}} - \hat{\boldsymbol{\beta}}_{\mathrm{opt}}) = o_p(1)$ *as* $N \to \infty$, *where* $\hat{\boldsymbol{\beta}}_{\mathrm{opt}}$ *is the most efficient estimator amongst the*

class of estimators that solve estimating equations of the form $\sum_{i=1}^{N} \boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{0}$*, with* $\boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})$ *being a* $p \times n_i$ *matrix whose dimension and structure may vary across different subjects due to different follow-up visits.*

This result can be easily justified by noting that $\hat{\boldsymbol{\beta}}_{\text{opt}}$ should solve the equation $\sum_{i=1}^{N} \boldsymbol{G}_i(\boldsymbol{\beta})^\top \boldsymbol{V}_i(\boldsymbol{\beta})^{-1}\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{0}$ based on the theory of inference functions (e.g., Godambe, 1991; Heyde, 1997). One special case of Theorem 4 is when the longitudinal outcomes follow a normal distribution, in which the equation $\sum_{i=1}^{N} \boldsymbol{G}_i(\boldsymbol{\beta})^\top \boldsymbol{V}_i(\boldsymbol{\beta})^{-1}\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{0}$ is actually the score equation. In such a case, $\hat{\boldsymbol{\beta}}_{\text{CEL}}$ is asymptotically equivalent to the maximum likelihood estimator.

4. NUMERICAL IMPLEMENTATION

Numerical implementation of the proposed CEL method can be accomplished based on Equation (9) with the invocation of nested optimization routines. Two optimization loops are required. The inner loop updates the Lagrange multipliers $\boldsymbol{\lambda}_i(\boldsymbol{\beta})$, $i = 1, \ldots, N$, for a fixed $\boldsymbol{\beta}$, and the outer loop updates $\boldsymbol{\beta}$. We use Newton–Raphson algorithm for both loops.

Inner loop:

For a fixed $\boldsymbol{\beta}$, given $\boldsymbol{\lambda}_i^{\text{old}}$, $\boldsymbol{\lambda}_i$ is updated as

$$\boldsymbol{\lambda}_i^{\text{new}} = \boldsymbol{\lambda}_i^{\text{old}} - \boldsymbol{\Lambda}_{i,\lambda\lambda}^{-1}\left(\boldsymbol{\lambda}_i^{\text{old}}, \boldsymbol{\beta}\right)\boldsymbol{\Lambda}_{i,\lambda}\left(\boldsymbol{\lambda}_i^{\text{old}}, \boldsymbol{\beta}\right),$$

where

$$\boldsymbol{\Lambda}_{i,\lambda}(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) = -\sum_{j\in S_i} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})}, \quad \boldsymbol{\Lambda}_{i,\lambda\lambda}(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) = \sum_{j\in S_i} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})^\top}{\left\{1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})\right\}^2}.$$

The algorithm may start with an initial value $\boldsymbol{\lambda}_i = 0$. The converged value gives the estimated $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ for a fixed $\boldsymbol{\beta}$ as the solution to (7).

Outer loop:

For a given $\boldsymbol{\beta}^{\text{old}}$, $\boldsymbol{\beta}$ is updated as

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left\{\sum_{i=1}^{N} \boldsymbol{L}_{i,\beta\beta}(\boldsymbol{\beta}^{\text{old}})\right\}^{-1}\left\{\sum_{i=1}^{N} \boldsymbol{L}_{i,\beta}(\boldsymbol{\beta}^{\text{old}})\right\},$$

where

$$\boldsymbol{L}_{i,\beta}(\boldsymbol{\beta}) = -\sum_{j\in S_i} w_{ij}\frac{\boldsymbol{G}_j(\boldsymbol{\beta})^\top}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})}\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}),$$

$$\boldsymbol{L}_{i,\beta\beta}(\boldsymbol{\beta}) = -\boldsymbol{\Lambda}_{i,\lambda\beta}^\top\left\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\}\boldsymbol{\Lambda}_{i,\lambda\lambda}^{-1}\left\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\}\boldsymbol{\Lambda}_{i,\lambda\beta}\left\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\},$$

$$\boldsymbol{\Lambda}_{i,\lambda\beta}\left\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\} = \sum_{j\in S_i} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{G}_j(\boldsymbol{\beta})}{\left\{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})\right\}^2} - \sum_{j\in S_i} w_{ij}\frac{\boldsymbol{G}_j(\boldsymbol{\beta})}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})}.$$

Iterate the inner and outer loops until certain convergence criterion is satisfied. At convergence, the algorithm produces the CEL estimator $\hat{\boldsymbol{\beta}}_{\text{CEL}}$.

A complication for the above numerical implementation may exist due to the empty set problem. This problem exists for both the EL and the CEL methods, and may become particularly

serious for the latter due to the subject-wise nature of (6). There are many possible reasons leading to the empty set problem; for example, when the model assumption (1) is incorrect, or the number of follow-up visits is large, or the size of a stratum is small, or the initial value of $\boldsymbol{\beta}$ is chosen far from the true $\boldsymbol{\beta}_0$, or combinations of the above. Although this problem will likely disappear when the sample size goes to infinity if model (1) is correctly specified, it could occasionally cause numerical difficulties in practice. Some researchers have proposed different ways to mitigate this issue, including Owen (2001), Chen, Variyath, & Abraham (2008), Emerson & Owen (2009) and Grendár & Judge (2010). So far, all these works were developed under the standard EL setting. Their adaption to the CEL setting is beyond the scope of this paper, and will be formally studied as future topics.

For now, our suggestion is to, in the inner loop of the numerical implementation, restrict each updated value of $\boldsymbol{\lambda}_i$ to be in the legitimate region $\{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i} : 1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta}) \geq w_{ij},\ j \in S_i\}$ and make $\Lambda_i(\boldsymbol{\lambda}_i, \boldsymbol{\beta})$ in (7) decrease. This approach has also been suggested and applied by other researchers (e.g., Kitamura, 2007; Hansen, 2014) to mitigate the impact of the empty set problem, and "appears to work reasonably well in practice" (Kitamura, 2007). It is easy to see that (7) is a convex minimization problem. Therefore, if the empty set problem does not exist, the inner loop estimate almost always converges to the global minimizer, which satisfies (6). A proof of this convergence can be given by following Chen, Sitter, & Wu (2002). When the empty set problem does exist, we may still try to minimize $\Lambda_i(\boldsymbol{\lambda}_i, \boldsymbol{\beta})$, although the final minimizer in this situation, if the algorithm indeed produces a finite minimizer, does not satisfy (6). However, this minimizer (if it exists) should eventually become the solution to (6) when the empty set problem disappears as the sample size increases, and thus using this minimizer in the calculation should not affect the asymptotic results. This is the strategy we took in our simulation studies. The maximization in the outer loop is more complicated, and the convergence of the Newton–Raphson algorithm may not be guaranteed. See Owen (2001) for some detailed discussion on related issues in the setting of unconditional moment restrictions. Nonetheless, the nested optimization is widely used by many researchers to implement the EL (CEL) method; see, for example, Owen (2001), Kitamura (2007) and Hansen (2014). According to Kitamura (2007), the nested optimization appears to be "the most stable way to compute the EL estimator."

For unbalanced longitudinal data, although we only need to consider the strata with $r_\omega > 0$ when deriving the asymptotic results, in practice any follow-up patterns could exist, and the values of r_ω are unknown due to a finite sample size. When there are many different patterns, to avoid potential numerical difficulties caused by small stratum sizes, we consider another strategy of data augmentation by creating pseudo-subjects from those who have longer and more complete follow-up visits. Specifically, for subject i, we redefine $S_i = \{j : 1 \leq j \leq N \text{ and } \tau_j \supseteq \tau_i\}$ and

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \min_{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}} \left[-\sum_{j \in S_i} w_{ij} \log \left\{ 1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j^{\dagger_i}(\boldsymbol{\beta}) \right\} \right] + \sum_{i=1}^{N} \sum_{j \in S_i} w_{ij} \log w_{ij},$$

where the superscript $\dagger_i$ means selecting components from the residual vector $\boldsymbol{g}_j(\boldsymbol{\beta})$ according to the set τ_i. Therefore, $\boldsymbol{g}_j^{\dagger_i}(\boldsymbol{\beta})$ may be regarded as the residual vector of a pseudo-subject created from subject j whose visit times contain subject i's visit times as a subset. As N increases, the small-stratum-size problem is unlikely to occur with the inclusion of pseudo-subjects and with our assumption that each stratum has size proportional to N. In the extreme case where a small stratum still exists, that stratum may be discarded in the analysis. In our simulation studies we employ this strategy, which improves the numerical performance of $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ in general.

In the case of binary outcome data, we have $\mathrm{var}(Y_{it} \mid \boldsymbol{X}_i) = \mu_{it}(1 - \mu_{it})$ where $\mu_{it} = \mu(\boldsymbol{X}_{it}^\top \boldsymbol{\beta})$. This information could be used to improve the finite sample performance of the CEL estimator

by redefining the residual vector to be

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \{Y_{i1} - \mu_{i1}, \ldots, Y_{in_i} - \mu_{in_i}, \times (Y_{i1} - \mu_{i1})^2 - \mu_{i1}(1 - \mu_{i1}), \ldots, \left(Y_{in_i} - \mu_{in_i}\right)^2 - \mu_{in_i}(1 - \mu_{in_i})\}^\top.$$

Model assumption (1) is still satisfied. Estimation and inference remain the same, except that the dimension of the Lagrange multiplier $\boldsymbol{\lambda}_i$ increases to $2n_i$.

5. SIMULATION EXPERIMENTS

The setup of our simulation experiments mimics a typical longitudinal study that involves three covariates: a continuous baseline covariate b_i, a binary treatment indicator d_i, and the visit time t. Assuming that there are five designed follow-up visits, we generate the longitudinal outcomes Y_{it} from the following model:

$$Y_{it} = \beta_1 + \beta_2 b_i + \beta_3 d_i + \beta_4 t + \beta_5 d_i \times t + \epsilon_{it}, \quad t = 1, \ldots, 5, \quad i = 1, \ldots, N,$$

where $b_i \sim \mathcal{N}(0, 3^2)$ and is truncated between -7 and 7, $d_i \sim$ Bernoulli(0.5), $d_i \times t$ is the treatment and time interaction, $(\epsilon_{i1}, \ldots, \epsilon_{i5})^\top \sim \mathcal{N}_5(\mathbf{0}, \boldsymbol{V}_i^{1/2}\boldsymbol{R}\boldsymbol{V}_i^{1/2})$, $\boldsymbol{V}_i$ is a 5×5 diagonal matrix with the tth diagonal element $\sigma_{it}^2 = \exp(\alpha_1 + \alpha_2|b_i| + \alpha_3 t)$, and $\boldsymbol{R}$ takes the first-order autoregressive structure with correlation coefficient ρ. The true values are set as $(\beta_1, \ldots, \beta_5)^\top = (0.5, 0.5, 1.0, 0.3, 0.3)^\top$, $(\alpha_1, \alpha_2, \alpha_3)^\top = (0.4, 0.6, 0.4)^\top$, and $\rho = 0.5$.

We use the following mechanism to generate the unbalanced data:

$$P(Y_{it} \text{ is deleted}) = \frac{1}{1 + \exp(3 - 0.1|b_i| - 0.01t)}.$$

Here, the unbalancedness depends on the covariates. This mechanism is not the MCAR by Little & Rubin (2002), but the MCAR by Diggle & Kenward (1994) or the covariate-dependent MCAR by Little (1995). Our proposed method, viewed as an available-data analysis approach, is still valid under this missingness mechanism. In the generated data, there are occasionally a few subjects with no more than two visits (on average 0.22%). To ease the computational burden of the CEL estimator, we discard this small proportion of subjects in its calculation. It can be verified that the final observed residuals satisfy model assumption (1). On average, approximately 30% of subjects have incomplete visits in the generated data, and their follow-up patterns are irregular. We conduct 500 replications to calculate the summary results. The Gaussian kernel is employed to calculate the weights, where the baseline covariate is standardized to have mean 0 and variance 1 in the bandwidth selection. We compare the CEL estimator with the GEE estimator under different working correlation structures, including the true one. In addition, we compare the CEL estimator with Wang, Qian, & Carroll's (2010) GEL estimator, for which the variance–covariance matrix is estimated separately using various models. In effect, Wang, Qian, & Carroll's GEL estimator may be regarded as either a GEE2 (Prentice & Zhao, 1991) estimator, or an EL estimator based on a set of estimating equations specified by the GEE2. Therefore, the comparison between the CEL and the GEL may be referred to either as the comparison between the CEL and the GEE2 or as that between the CEL and the GEE2-based EL.

Table 1 presents the summarized results of the comparisons based on $N = 100$ and $N = 500$. To speed up the calculation, a prefixed bandwidth $b_N = 2N^{-1/10}$ is used, and this rate of b_N converging to 0 is selected following the rule of thumb described in Section 2.1. When σ_{it}^2 is not correctly modelled, Table 1 clearly illustrates that $\hat{\boldsymbol{\beta}}_{\text{CEL}}$ outperforms both the GEE estimators and the GEL estimators in terms of the total mean square error (tMSE), regardless of whether the

TABLE 1: Comparison of different estimators.

		cel	*gee.ar*	*gee.cs*	*gel*1.*ar*	*gel*1.*cs*	*gel*2.*ar*	*gel*2.*cs*	*gel*3.*ar*	*gel*3.*cs*
		0.0321	−0.0223	−0.0190	−0.0087	−0.0083	−0.0102	−0.0094	0.0208	0.0187
	β_1	(0.6441)	(0.8189)	(0.8377)	(0.7193)	(0.7275)	(0.7188)	(0.7317)	(0.4741)	(0.4887)
		[0.4159]	[0.6710]	[0.7022]	[0.5175]	[0.5293]	[0.5167]	[0.5355]	[0.2252]	[0.2392]
		−0.0102	−0.0106	−0.0102	−0.0118	−0.0110	−0.0110	−0.0106	−0.0078	−0.0076
	β_2	(0.1873)	(0.2366)	(0.2341)	(0.1943)	(0.1990)	(0.1945)	(0.2015)	(0.1382)	(0.1428)
		[0.0352]	[0.0561]	[0.0549]	[0.0379]	[0.0397]	[0.0380]	[0.0407]	[0.0192]	[0.0204]
N = 100		−0.0472	−0.0124	−0.0056	−0.0230	−0.0170	−0.0191	−0.0140	−0.0421	−0.0368
	β_3	(0.9243)	(1.1334)	(1.1835)	(0.9958)	(1.0146)	(0.9961)	(1.0221)	(0.6471)	(0.6712)
		[0.8565]	[1.2848]	[1.4007]	[0.9921]	[1.0297]	[0.9926]	[1.0448]	[0.4205]	[0.4519]
		−0.0056	0.0092	0.0087	0.0055	0.0059	0.0059	0.0062	−0.0112	−0.0089
	β_4	(0.2630)	(0.3007)	(0.2947)	(0.2709)	(0.2716)	(0.2720)	(0.2736)	(0.1870)	(0.1916)
		[0.0692]	[0.0905]	[0.0869]	[0.0734]	[0.0738]	[0.0740]	[0.0749]	[0.0351]	[0.0368]
		0.0133	0.0068	0.0028	0.0079	0.0035	0.0058	0.0017	0.0267	0.0225
	β_5	(0.3713)	(0.4392)	(0.4310)	(0.3943)	(0.3930)	(0.3936)	(0.3936)	(0.2501)	(0.2527)
		[0.1381]	[0.1929]	[0.1858]	[0.1556]	[0.1545]	[0.1550]	[0.1549]	[0.0633]	[0.0644]
		0.0394	0.0205	0.0247	0.0349	0.0320	0.0343	0.0317	0.0185	0.0166
	β_1	(0.2810)	(0.3573)	(0.3785)	(0.3135)	(0.3303)	(0.3148)	(0.3335)	(0.2076)	(0.2162)
		[0.0805]	[0.1281]	[0.1439]	[0.0995]	[0.1101]	[0.1003]	[0.1122]	[0.0434]	[0.0470]
		−0.0048	−0.0065	−0.0064	−0.0056	−0.0053	−0.0054	−0.0052	−0.0062	−0.0049
	β_2	(0.0775)	(0.1054)	(0.1050)	(0.0897)	(0.0924)	(0.0895)	(0.0933)	(0.0589)	(0.0615)
		[0.0060]	[0.0111]	[0.0111]	[0.0081]	[0.0086]	[0.0080]	[0.0087]	[0.0035]	[0.0038]
N = 500		−0.0673	−0.0426	−0.0455	−0.0639	−0.0596	−0.0633	−0.0589	−0.0255	−0.0210
	β_3	(0.3891)	(0.5095)	(0.5458)	(0.4449)	(0.4712)	(0.4471)	(0.4758)	(0.2982)	(0.3186)
		[0.1559]	[0.2614]	[0.2999]	[0.2020]	[0.2256]	[0.2039]	[0.2299]	[0.0896]	[0.1019]
		−0.0101	−0.0031	−0.0051	−0.0077	−0.0076	−0.0079	−0.0079	−0.0069	−0.0064
	β_4	(0.1108)	(0.1436)	(0.1438)	(0.1272)	(0.1318)	(0.1275)	(0.1322)	(0.0811)	(0.0842)
		[0.0124]	[0.0206]	[0.0207]	[0.0162]	[0.0174]	[0.0163]	[0.0175]	[0.0066]	[0.0071]
		0.0269	0.0177	0.0197	0.0243	0.0244	0.0247	0.0248	0.0147	0.0135
	β_5	(0.1574)	(0.2066)	(0.2082)	(0.1810)	(0.1879)	(0.1808)	(0.1880)	(0.1146)	(0.1209)
		[0.0255]	[0.0430]	[0.0437]	[0.0334]	[0.0359]	[0.0333]	[0.0360]	[0.0134]	[0.0148]

cel, CEL; *gee*, GEE; *gel*1, *gel*2, *gel*3, GEL with σ_{it}^2 being modelled by $(\alpha_1 + \alpha_2 t)^2$, $\exp(\alpha_1 + \alpha_2 t)$ and the truth, respectively; *ar*, first-order autoregressive; *cs*, compound symmetry. Three summary statistics are calculated: bias, empirical standard error (number in parentheses), and mean square error (number in brackets).

correlation structure is correctly specified. More specifically, when $N = 100$, the CEL estimator reduces the tMSE by 34%, 38%, 15%, 17%, 15% and 18% compared to the 2nd to the 7th estimators, respectively; and when $N = 500$, the amount of reduction becomes even higher as 40%, 46%, 22%, 30%, 23% and 31%, respectively. Such robustness against incorrect modelling of the variance–covariance matrix is a clear advantage of our proposed method. Since the data

TABLE 2: Effect of bandwidth on the CEL estimator.

	$c = 0.6$	$c = 0.9$	$c = 1.2$	$c = 1.5$	$c = 1.8$	$c = 2.1$	$c = 2.4$	$c = 2.7$	CV
	0.0265	0.0285	0.0362	0.0368	0.0381	0.0391	0.0400	0.0410	0.0394
β_1	(0.3361)	(0.3315)	(0.3399)	(0.3547)	(0.3644)	(0.3741)	(0.3830)	(0.3914)	(0.3781)
	[0.1137]	[0.1107]	[0.1168]	[0.1271]	[0.1342]	[0.1415]	[0.1483]	[0.1549]	[0.1445]
	−0.0040	−0.0083	−0.0078	−0.0069	−0.0057	−0.0052	−0.0050	−0.0049	−0.0049
β_2	(0.1048)	(0.0983)	(0.0987)	(0.1017)	(0.1017)	(0.1041)	(0.1065)	(0.1087)	(0.1055)
	[0.0110]	[0.0097]	[0.0098]	[0.0104]	[0.0104]	[0.0109]	[0.0114]	[0.0118]	[0.0112]
	−0.0372	−0.0400	−0.0491	−0.0513	−0.0529	−0.0547	−0.0565	−0.0582	−0.0571
β_3	(0.4799)	(0.4696)	(0.4755)	(0.4873)	(0.5002)	(0.5120)	(0.5237)	(0.5354)	(0.5159)
	[0.2317]	[0.2221]	[0.2285]	[0.2401]	[0.2530]	[0.2652]	[0.2775]	[0.2900]	[0.2694]
	−0.0089	−0.0079	−0.0099	−0.0100	−0.0103	−0.0105	−0.0106	−0.0107	−0.0108
β_4	(0.1335)	(0.1310)	(0.1335)	(0.1388)	(0.1429)	(0.1462)	(0.1493)	(0.1523)	(0.1475)
	[0.0179]	[0.0172]	[0.0179]	[0.0194]	[0.0205]	[0.0215]	[0.0224]	[0.0233]	[0.0219]
	0.0171	0.0178	0.0200	0.0205	0.0203	0.0205	0.0207	0.0209	0.0210
β_5	(0.1886)	(0.1843)	(0.1857)	(0.1923)	(0.1984)	(0.2033)	(0.2081)	(0.2128)	(0.2046)
	[0.0358]	[0.0343]	[0.0349]	[0.0374]	[0.0398]	[0.0418]	[0.0437]	[0.0457]	[0.0423]

$N = 300$; $b_N = cN^{-1/10}$; CV: cross-validation. Three summary statistics are calculated: bias, empirical standard error (number in parentheses), and mean square error (number in brackets).

are generated from a normal distribution, $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is asymptotically equivalent to the maximum likelihood estimator, which is essentially the one given by the second last column. Due to the nonparametric nature of $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$, its finite-sample tMSE is larger than that of the maximum likelihood estimator. However, this difference becomes smaller as the sample size increases. Specifically, the tMSE of $\hat{\boldsymbol{\beta}}_{\mathrm{CEL}}$ is 98% higher than that of the maximum likelihood estimator when $N = 100$, and this number drops to 79% when $N = 500$. A similar trend is observed by comparing the CEL estimator with the estimator given by the last column, in which case the former has 86% higher tMSE than the latter when $N = 100$, and this number drops to 61% when $N = 500$. From Table 1, an additional observation is that, under the same model for σ_{it}^2, misspecification of the correlation structure causes only marginal efficiency loss. On the other hand, different models for σ_{it}^2 lead to significant efficiency variation. This is in a full agreement with Wang & Lin (2005).

Table 2 reports results concerning the influence of the bandwidth on the performance of the CEL estimator. With $N = 300$, the bandwidth is given by $b_N = cN^{-1/10}$, which varies with $c = 0.6, 0.9, \ldots, 2.7$. Table 2 also includes results based on the cross-validation-selected bandwidth. The value of c that gives the smallest tMSE is 0.9. When c becomes larger (toward oversmoothing), both the bias (except for β_2) and the empirical standard error increase. Overall, the impact of different bandwidths does not appear to be dramatic. The tMSE when $c = 0.9$ is 25% smaller than that when $c = 2.7$. This empirical property has also been reported in Kitamura, Tripathi, & Ahn (2004), and is of practical importance, as it could save substantial computing cost on the search for the optimal bandwidth. Taking our simulation study as an example, for smaller values of c, such as 0.6 and 0.9, the algorithm requires much longer time to converge in comparison to cases with a larger value of c.

Table 3 reports the numerical performance of the two asymptotic variance estimators, (11) and (12), under different sample sizes. The bandwidth is prefixed as $b_N = 2N^{-1/10}$. The empirical

TABLE 3: Comparison of the asymptotic variance estimators.

	$N = 100$	$N = 300$	$N = 500$	$N = 1000$
$\text{var}(\beta_1)_{(11)}$	−0.0328	−0.0079	−0.0020	−0.0001
	(0.1344)	(0.0272)	(0.0137)	(0.0043)
$\text{var}(\beta_1)_{(12)}$	−0.0709	−0.0147	−0.0046	−0.0011
	(0.1025)	(0.0216)	(0.0106)	(0.0036)
$\text{var}(\beta_2)_{(11)}$	−0.0240	−0.0064	−0.0034	−0.0017
	(0.0032)	(0.0009)	(0.0005)	(0.0002)
$\text{var}(\beta_2)_{(12)}$	−0.0250	−0.0065	−0.0034	−0.0017
	(0.0024)	(0.0006)	(0.0003)	(0.0001)
$\text{var}(\beta_3)_{(11)}$	−0.0844	0.0026	0.0025	0.0016
	(0.2399)	(0.0448)	(0.0214)	(0.0066)
$\text{var}(\beta_3)_{(12)}$	−0.1639	−0.0124	−0.0031	−0.0003
	(0.1785)	(0.0348)	(0.0165)	(0.0057)
$\text{var}(\beta_4)_{(11)}$	−0.0080	−0.0004	0.0000	0.0001
	(0.0232)	(0.0044)	(0.0022)	(0.0007)
$\text{var}(\beta_4)_{(12)}$	−0.0141	−0.0014	−0.0004	0.0000
	(0.0172)	(0.0035)	(0.0017)	(0.0006)
$\text{var}(\beta_5)_{(11)}$	−0.0142	0.0010	−0.0002	0.0001
	(0.0397)	(0.0072)	(0.0034)	(0.0011)
$\text{var}(\beta_5)_{(12)}$	−0.0270	−0.0013	−0.0011	−0.0002
	(0.0297)	(0.0056)	(0.0026)	(0.0009)

$\text{var}()_{(11)}$, estimator based on formula (11); $\text{var}()_{(12)}$, estimator based on formula (12). Two summary statistics are calculated: bias and empirical standard error (number in parentheses).

variance of the CEL estimator based on 500 replications is used as the true value of the variance when calculating the bias. As shown in Table 3, both estimators appear to be consistent, and (11) has smaller bias but slightly larger variance. Both formulas tend to underestimate the variance of $\hat{\beta}_{\text{CEL},2}$, the CEL estimator of the regression coefficient for the continuous baseline covariate. The issue that the sandwich estimator may underestimate the asymptotic variance has been observed by many researchers; see, for example, Kauermann & Carroll (2001) and references therein.

Table 4 reports the influence of the bandwidth on the two asymptotic variance estimators when $N = 300$ and $b_N = cN^{-1/10}$, with $c = 0.6, 0.9, \ldots, 2.7$. We can see that small bandwidth leads to underestimation for both estimators. As the bandwidth increases, such underestimation disappears, except for the variance of $\hat{\beta}_{\text{CEL},2}$. For a larger bandwidth, the bias of (11) is smaller than that of (12). When the bandwidth is selected by cross-validation, both estimators perform reasonably well, and (11) has smaller bias.

To further compare (11) and (12), we consider a Wald-type test statistic to test the following hypothesis of the overall treatment effect:

$$\text{H}_0: \quad \beta_3 = 1 \text{ and } \beta_5 = 0.3.$$

TABLE 4: Effect of bandwidth on the asymptotic variance estimators.

	$c = 0.6$	$c = 0.9$	$c = 1.2$	$c = 1.5$	$c = 1.8$	$c = 2.1$	$c = 2.4$	$c = 2.7$	CV
$\text{var}(\beta_1)_{(11)}$	−0.0524	−0.0407	−0.0250	−0.0278	−0.0127	−0.0078	−0.0101	−0.0131	−0.0118
	(0.0201)	(0.0295)	(0.0352)	(0.0366)	(0.0300)	(0.0268)	(0.0265)	(0.0264)	(0.0296)
$\text{var}(\beta_1)_{(12)}$	−0.0390	−0.0203	−0.0139	−0.0148	−0.0142	−0.0151	−0.0166	−0.0186	−0.0164
	(0.0125)	(0.0156)	(0.0188)	(0.0205)	(0.0212)	(0.0218)	(0.0223)	(0.0228)	(0.0228)
$\text{var}(\beta_2)_{(11)}$	−0.0093	−0.0082	−0.0074	−0.0078	−0.0066	−0.0065	−0.0071	−0.0076	−0.0071
	(0.0011)	(0.0010)	(0.0013)	(0.0013)	(0.0011)	(0.0008)	(0.0007)	(0.0007)	(0.0009)
$\text{var}(\beta_2)_{(12)}$	−0.0073	−0.0053	−0.0054	−0.0060	−0.0061	−0.0067	−0.0073	−0.0078	−0.0070
	(0.0008)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
$\text{var}(\beta_3)_{(11)}$	−0.1087	−0.0827	−0.0466	−0.0404	−0.0079	0.0031	−0.0005	−0.0061	−0.0026
	(0.0317)	(0.0465)	(0.0586)	(0.0616)	(0.0482)	(0.0447)	(0.0458)	(0.0472)	(0.0501)
$\text{var}(\beta_3)_{(12)}$	−0.0817	−0.0408	−0.0222	−0.0150	−0.0129	−0.0124	−0.0140	−0.0174	−0.0130
	(0.0182)	(0.0225)	(0.0271)	(0.0304)	(0.0331)	(0.0357)	(0.0379)	(0.0399)	(0.0383)
$\text{var}(\beta_4)_{(11)}$	−0.0084	−0.0065	−0.0039	−0.0040	−0.0015	−0.0003	−0.0005	−0.0008	−0.0009
	(0.0036)	(0.0050)	(0.0061)	(0.0064)	(0.0051)	(0.0043)	(0.0043)	(0.0043)	(0.0049)
$\text{var}(\beta_4)_{(12)}$	−0.0060	−0.0029	−0.0016	−0.0016	−0.0015	−0.0014	−0.0015	−0.0017	−0.0016
	(0.0022)	(0.0026)	(0.0031)	(0.0033)	(0.0035)	(0.0036)	(0.0037)	(0.0038)	(0.0037)
$\text{var}(\beta_5)_{(11)}$	−0.0168	−0.0126	−0.0066	−0.0062	−0.0010	0.0012	0.0007	−0.0002	0.0002
	(0.0053)	(0.0077)	(0.0097)	(0.0103)	(0.0084)	(0.0072)	(0.0074)	(0.0077)	(0.0084)
$\text{var}(\beta_5)_{(12)}$	−0.0119	−0.0052	−0.0019	−0.0014	−0.0013	−0.0013	−0.0016	−0.0021	−0.0013
	(0.0030)	(0.0036)	(0.0043)	(0.0048)	(0.0053)	(0.0058)	(0.0062)	(0.0065)	(0.0062)

$N = 300$; $b_N = cN^{-1/10}$; CV, cross-validation; $\text{var}()_{(11)}$, estimator based on formula (11); $\text{var}()_{(12)}$, estimator based on formula (12). Two summary statistics are calculated: bias and empirical standard error (number in parentheses).

Here, the Wald statistic is given by

$$(\hat{\beta}_{\text{CEL},3} - 1, \hat{\beta}_{\text{CEL},5} - 0.3)\{\text{aVar}(\hat{\beta}_{\text{CEL},3}, \hat{\beta}_{\text{CEL},5})\}^{-1}(\hat{\beta}_{\text{CEL},3} - 1, \hat{\beta}_{\text{CEL},5} - 0.3)^{\top},$$

where $\text{aVar}(\hat{\beta}_{\text{CEL},3}, \hat{\beta}_{\text{CEL},5})$ is the asymptotic variance of $(\hat{\beta}_{\text{CEL},3}, \hat{\beta}_{\text{CEL},5})$. Under H_0, both Wald statistics constructed from (11) and (12) follow the χ^2_2 distribution asymptotically. In the comparison, we take $N = 300$ and b_N is selected by cross-validation. Figure 1 presents a QQ-plot of these two Wald statistics. It can be seen that although both statistics approximately follow the χ^2_2 distribution, the one based on (11) turns out to be closer to the 45^o line. This agrees with the conclusion drawn from Table 4 that the variance estimate given by (11) has smaller bias than that given by (12).

In summary, we recommend using cross-validation to select the bandwidth, and using (11) to estimate the asymptotic variance.

6. DATA ANALYSIS

We now illustrate the proposed CEL method by analyzing the data collected from the Kenya primary school nutritional intervention study (Neumann et al., 2003). This is a randomized controlled

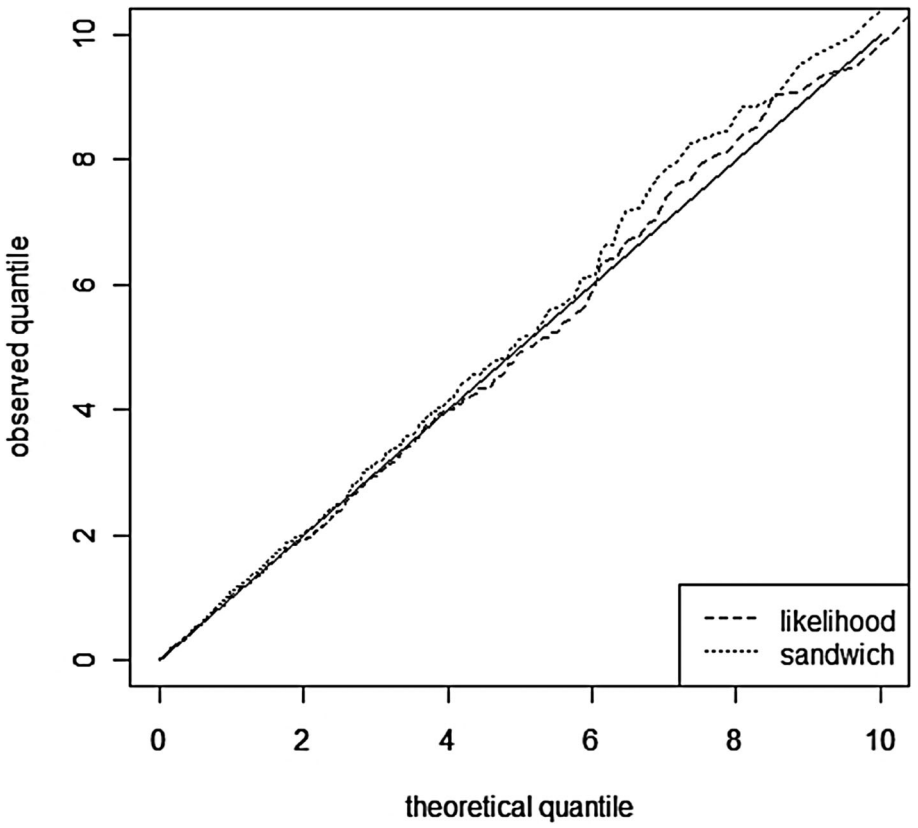

FIGURE 1: QQ-plot of the two Wald statistics based on the "likelihood" (formula (11)) and the "sandwich" (formula (12)) asymptotic variance estimators, respectively.

trial designed to examine the relationship of growth, cognitive development and physical activity with the intake of animal source foods, adjusted by other covariates. A total of 554 school children from 12 schools in rural Embu District, Kenya, are randomized to four nutritional intervention arms: Meat, milk, energy and control. Meat, milk and energy in the form of extra vegetable oil are added into the local plant-based dish Githeri for the Meat, Milk and Energy arms, respectively. The control arm has no feeding intervention. Lasting for over 2 years (from July 1998 to December 2000), this study includes a baseline visit during the period of July to August, 1998, and the delivery of intervention and data collection over the period of September 1998 to December 2000. The cognitive function is measured at the baseline.

We analyse a sub dataset of this study that concerns the cognitive outcomes. The data are downloaded from http://rem.ph.ucla.edu/rob/mld/data/tabdelimiteddata/cognitive.txt. More details about the data can be found in Weiss (2005). Cognitive ability is measured by the Raven's score determined on the Raven's coloured progressive matrices. Each subject has five follow-up visits, and the first (baseline) visit takes place prior to the randomization. The randomization is marked as time zero in our analysis.

Since all subjects effectively belong to the control arm at the baseline visit, the baseline Raven's score is treated as a covariate (*braven*) in the model. Other covariates include the baseline age (*age*), baseline social economic status (*ses*) determined by an extensive survey, gender (*boy*), and visit time (*time*). For the intervention, three dummy variables, *meat*, *milk* and *energy* are created.

TABLE 5: Analysis of the Kenya school nutritional intervention data ($N = 524$).

	cel			gee.ar			gee.cs			gee.un		
	Est.	SE	*P*-value	Est	SE	*P*-value	Est	SE	*P*-value	Est	SE	*P*-value
Intercept	11.53	0.955	0.00	11.67	1.055	0.00	11.44	1.072	0.00	11.47	1.061	0.00
Age	0.119	0.075	0.11	0.111	0.082	0.18	0.149	0.082	0.07	0.137	0.082	0.09
SEs	0.009	0.004	0.02	0.006	0.004	0.12	0.007	0.004	0.09	0.006	0.004	0.08
Braven	0.241	0.037	0.00	0.251	0.044	0.00	0.247	0.045	0.00	0.250	0.044	0.00
Boy	0.511	0.178	0.00	0.636	0.178	0.00	0.567	0.180	0.00	0.596	0.179	0.00
Time	0.885	0.149	0.00	1.010	0.142	0.00	0.954	0.142	0.00	0.958	0.141	0.00
Energy × *time*	0.119	0.193	0.54	−0.127	0.189	0.50	−0.060	0.186	0.75	−0.089	0.186	0.63
Meat × *time*	0.538	0.199	0.01	0.354	0.203	0.08	0.405	0.204	0.05	0.392	0.201	0.05
Milk × *time*	−0.019	0.191	0.92	−0.273	0.188	0.15	−0.227	0.187	0.23	−0.234	0.186	0.21

cel, CEL; *gee*, GEE; *ar*, first-order autoregressive; *cs*, compound symmetry; *un*, unstructured; est, estimated value; se, estimated standard error, where for the CEL method, it is based on formula (11).

After removing the subjects with unmeasured covariates and the subjects who only have baseline visit, our analysis is based on 524 subjects and the data are unbalanced. We use the average time of each visit in our analysis, and the following model is fitted:

$$E(raven_{it}) = \beta_1 + \beta_2 age_i + \beta_3 ses_i + \beta_4 braven_i + \beta_5 boy_i + (\beta_6 + \beta_7 energy_i + \beta_8 meat_i + \beta_9 milk_i) \times time_t.$$

Our residual analysis does not suggest any clear dependence of the marginal variance on any covariate. Therefore, we only compare the results obtained from the CEL method and the GEE method (Table 5). Since no clear structure of the within-subject correlation has been found in the residual analysis, we also include the GEE estimator employing the unstructured correlation.

Both the CEL and the GEE methods find that the cognitive ability significantly improves over time, and is significantly higher for boys than for girls. It is interesting to note that our CEL method has identified social economic status as a significant factor for the cognitive ability, while the GEE method fails to detect this significance. As for the growth rate under the three different nutritional interventions compared to the control arm, our CEL method indicates that meat significantly improves the development of the cognitive ability, whereas the GEE method shows only a similar trend with no statistical significance. Neither the CEL nor the GEE finds significance of milk or energy intervention in helping the cognitive growth.

7. DISCUSSION

In this paper we propose to perform inference for longitudinal data using the CEL method. No explicit modelling of the variance–covariance matrix of the longitudinal outcomes is required. Some additional insights from our simulation experiments and data application are worth mentioning. First, the CEL method enjoys high estimation efficiency when moderate to high level of heteroscedasticity exists, especially when it is difficult to model this heteroscedasticity. When homoscedasticity is a more reasonable assumption, the CEL estimator may not outperform some of the existing estimators, such as the GEE estimator, due to the nonparametric weight calculation. Second, the cross-validation criterion (4) for bandwidth selection has not been theoretically justified yet, and it tends to lead to over-smoothing. Third, when the number of covariates is large and

the covariates vary in scales and/or types, the kernel-based weight calculation is challenging. It is worthwhile to explore more flexible methods to calculate the weights. Fourth, our stratification strategy dealing with the possible unbalancedness is more appropriate for data collected from balanced study designs. Therefore, studies in which subjects visit at irregular follow-up times call for some adjustments. For example, in analyzing the longitudinal nutrition data, the average time of each visit is used. Fifth, the stratification strategy may not perform well if one or more strata have small sizes. Although the creation of pseudo-subjects is effective to improve the numerical performance, some additional attention is needed, especially when the total number of follow-up visits and the number of distinctive visit patterns are both large. Last, but not least, a major hurdle for applying the CEL method is its computational burden, which is a well-known open problem in EL methodology. The development of fast algorithms is of great interest.

In addition to the CEL method, there are others available for estimation under models defined by conditional moment restrictions, including Newey (1993) and Donald, Imbens, & Newey (2003). Newey's (1993) method requires a preliminary estimator of the parameter of interest, and needs to explicitly estimate the variance–covariance matrix in a nonparametric way. However, it is often difficult to find a well-behaved estimate of this matrix in practice (Kitamura, Tripathi, & Ahn, 2004). The CEL method, on the contrary, achieves the same efficiency automatically without estimating the variance–covariance. Donald, Imbens, & Newey (2003) proposed to construct an increasing number of unconditional moment restrictions to "span" the conditional moment restrictions that define the model, and then use the EL method for estimation. As the sample size and the number of unconditional moment restrictions both go to infinity, Donald, Imbens, & Newey's (2003) estimator achieves the same efficiency as that of the CEL estimator. However, no numerical study of this method has been reported, and thus its finite sample performance is unknown.

ACKNOWLEDGEMENTS

We wish to thank the Editor, the Associate Editor and two reviewers for their valuable comments, which have helped significantly improve the quality of our work.

BIBLIOGRAPHY

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34, 305–334.

Chen, J., Sitter, R. R. & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230–237.

Chen, J., Variyath, A. M. & Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17, 426–443.

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

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43, 49–73.

Donald, S. G., Imbens, G. W. & Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117, 55–93.

Emerson, S. & Owen, A. (2009). Calibration of the empirical likelihood method for a vector mean. *Electronic Journal of Statistics*, 3, 1161–1192.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208–1212.

Godambe, V. P. (1991). *Estimating Functions*, Oxford University Press, Oxford.

Grendár, M. & Judge, G. (2009). Empty set problem of maximum empirical likelihood methods. *Electronic Journal of Statistics*, 3, 1542–1555.

Grendár, M. & Judge, G. Revised empirical likelihood. *CUDARE Working Papers*, page http://ageconsearch.umn.edu/bitstream/91799/2/CUDARE2010.

Hansen, B. E. (2014). *Econometrics*. Draft graduate textbook.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*, Springer-Verlag, New York.

Jiang, J., Luan, Y., & Wang, Y-G. (2007). Iterative estimating equations: Linear convergence and asymptotic properties. *Annals of Statistics*, 35, 2233–2260.

Kauermann, G. & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387–1396.

Kitamura, Y. (2007). Empirical likelihood methods in econometrics: Theory and practice. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Vol. 3, Blundell, R., Newey, W. K., and Persson, T., editors. Cambridge University Press, Cambridge.

Kitamura, Y., Tripathi, G., & Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72, 1667–1714.

Lai, T. L. & Small, D. S. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B*, 69, 79–99.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, 98, 355–370.

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

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., Wiley, New York.

Neumann, C. G., Bwibo, N. O., Murphy, S. P., Sigman, M., Whaley, S., Allen, L. H., Guthrie, D., Weiss, R. E., & Demment, M. W. (2003). Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in Kenyan School Children: Background, study design and baseline findings. *The Journal of Nutrition*, 133, 3941S–3949S.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Handbook of Statistics*, Vol. 11. North-Holland, Amsterdam.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18, 90–120.

Owen, A. (2001). *Empirical Likelihood*, Chapman & Hall/CRC Press, New York.

Pan, J. & MacKenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, 90, 239–244.

Pepe, M. S. & Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation*, 23, 939–951.

Prentice, R. L. & Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825–839.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300–325.

Qu, A., Lindsay, B. G., & Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823–836.

Smith, R. J. (2007). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*, 138, 430–460.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modelling, Analytics, and Applications*. Springer, New York.

Wang, S., Qian, L., & Carroll, R. J. (2010). Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97, 79–93.

Wang, Y. G. & Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika*, 90, 29–41.

Wang, Y. G., & Lin, X. (2005). Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, 61, 413–421.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439–447.

Weiss R. E. (2005). *Modelling Longitudinal Data*, Springer, New York.

Xue, L., & Zhu, L. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, 94, 921–937.

Ye, H. & Pan, J. (2006). Modelling of covariance structures in generalized estimating equations for longitudinal data. *Biometrika*, 93, 927–941.

You, J., Chen, G., & Zhou, Y. (2006). Block empirical likelihood for longitudinal partially linear regression models. *Canadian Journal of Statistics*, 34, 79–96.

Zhang, J. & Gijbels, I. (2003). Sieve empirical likelihood and extensions of the generalized least squares, *Scandinavian Journal of Statistics*, 30, 1–24.

Received 21 June 2013
Accepted 8 May 2014

.