Journal of Multivariate Analysis 130 (2014) 409–424

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Variable selection and estimation for longitudinal survey data

Li Wang [a,*], Suojin Wang [b], Guannan Wang [a]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, United States

[b] Department of Statistics, Texas A&M University, College Station, TX 77843, United States

HIGHLIGHTS

- We develop a general strategy for model selection in longitudinal surveys.
- We propose a survey weighted penalized GEE to select significant variables.
- We apply the EF-bootstrap method to obtain standard errors for complex surveys.
- We find that survey weights should be accounted for informative sampling designs.

ARTICLE INFO

Article history:
Received 10 October 2012
Available online 20 May 2014

AMS 2000 subject classifications:
primary 62G08

Keywords:
Bootstrap
Generalized estimating equations
Penalty
Superpopulation
Sampling weights

ABSTRACT

There is wide interest in studying longitudinal surveys where sample subjects are observed successively over time. Longitudinal surveys have been used in many areas today, for example, in the health and social sciences, to explore relationships or to identify significant variables in regression settings. This paper develops a general strategy for the model selection problem in longitudinal sample surveys. A survey weighted penalized estimating equation approach is proposed to select significant variables and estimate the coefficients simultaneously. The proposed estimators are design consistent and perform as well as the oracle procedure when the correct submodel was known. The estimating function bootstrap is applied to obtain the standard errors of the estimated parameters with good accuracy. A fast and efficient variable selection algorithm is developed to identify significant variables for complex longitudinal survey data. Simulated examples are illustrated to show the usefulness of the proposed methodology under various model settings and sampling designs.

1. Introduction

In the past two decades, various longitudinal surveys have been undertaken, where sample subjects are observed successively over time. Some examples are the US National Compensation Survey, the International Price Program, the Survey of Income and Program Participation, the US Longitudinal Studies of Aging and a range of more specialized studies. These represent a very substantial investment in longitudinal resources, producing a diverse portfolio of research materials, and a vibrant national research culture that has a strong international visibility. Although many studies of these surveys focus on estimating means, totals, proportions or ratios for certain populations, longitudinal survey data are frequently used for the modeling and estimation of the relationship in regression analysis. For example, longitudinal social surveys are conducted in many countries to identify factors that have effects on unemployment status or income; many health surveys are aimed to gain insight of health determinants rather than estimating population totals or proportions.

* Corresponding author.
E-mail address: lilywang@uga.edu (L. Wang).

http://dx.doi.org/10.1016/j.jmva.2014.05.006
0047-259X/

Longitudinal surveys are usually stratified and often multistage with unequal probabilities of selection at certain stages. If some parts of the population are sampled more intensively than others and the survey sampling design is ignored in the model selection, statistical inferences drawn from the sample can be remarkably different from those drawn from the population.

In this paper, marginal models for longitudinal survey data are considered to tackle the design and longitudinal features simultaneously. Generalized estimating equations (GEE) proposed by [11] is a popular method for these models. [13] first introduced GEE for longitudinal survey data. [17] adapted the GEE approach of [13] to the analysis of ordinal longitudinal survey responses. [3] developed a pseudo-GEE approach for longitudinal surveys under a joint randomization framework and established the consistency of the resulting estimators.

In many surveys, a large number of auxiliary variables may be collected, and we may want to determine the "best" subset of the variables. Longitudinal survey data with a large number of covariates have become increasingly more common in many scientific disciplines. One representative example is the Canadian National Population Health Survey where the researchers are interested in linking common risk factors with the possibility of loss of independence among seniors. In this study, many variables, such as age, gender, smoking status, weight, height, chronic conditions, area of residence, etc., were measured over the years to describe the seniors' health status and lifestyles. In some other examples of longitudinal data, the number of variables measured on each individual or sampling unit may not be many, but if we consider various interaction effects, the number of predictors in the statistical model can still be large. In addition, knowing which variables are relevant gives insight into the nature of the survey design problem. For example, variable selection can be adopted to find stratification variables in the primary sampling unit selection process for many surveys.

Variable selection is an essential part of many statistical methods, yet has been less studied in survey sampling compared with other areas of applied statistics. This is partly due to the challenges created in joint consideration of the sampling scheme, multilevel correlation and variable selection. The problem of selecting auxiliary variables was considered by [16] in the model-assisted framework while [4,5] in the prediction framework. [21] proposed a Bayesian information criterion based method to select the auxiliary variables for use in the additive model-assisted framework. Although they are practically useful, these traditional selection procedures ignore stochastic errors inherited in the stages of variable selections [6]. The well studied shrinkage methods such as LASSO [18,19,7] are developed under non-survey settings and are inappropriate to select variables for data collected through complex sampling designs.

In this paper, we propose a consistent variable selection and estimation method for the marginal mean models for survey sampling based on the penalized estimating equation approach. To the best of our knowledge, this is the first attempt to consider this approach in sample surveys. We demonstrate that the proposed method performs as well as the oracle procedure that assumes the true submodel to be known.

The rest of the paper is organized as follows. Section 2 introduces the models for longitudinal survey data, discusses the main ideas of the penalized estimating equation approach, and provides the asymptotic properties of the penalized estimators. Section 3 discusses some implementation issues and provides the estimating function bootstrap variance estimators. The performance of the proposed variable selection method is studied via simulated data in Section 4. Section 5 summarizes the main results along with areas for future research. The proofs of the theorems along with technical lemmas are provided in the Appendix.

2. Methodology

2.1. General setting

Suppose that the finite population U_N consists of N individuals. In a longitudinal survey, a sample s of size n is selected at wave one using a specified sampling scheme, and observed over a specified number of time points. Let w_i be longitudinal weights attached to the ith sample. We assume that the w_is are first adjusted for unit nonresponse, then subjected to post stratification adjustment to ensure consistency.

Suppose that the ith respondent is observed for m_i occasions ($1 \leq m_i \leq m$). The data for the ith sample consist of $\left\{y_{ij}, \mathbf{x}_{ij}\right\}_{j=1}^{m_i}$, where y_{ij} is the response on occasion j, $\mathbf{x}_{ij} = \left(x_{ij1}, \ldots, x_{ijd}\right)^{\mathrm{T}}$ is a d-vector of covariates for each sampled individual. The marginal model assumes that the mean response $\mu_{ij} = E(y_{ij}|\mathbf{x}_{ij})$ is a function of $\mathbf{x}_{ij}$. In this paper, we assume that μ_{ij} depends on $\mathbf{x}_{ij}$ through a known monotonic and differentiable link function $g(\cdot)$, so that we get the generalized linear superpopulation model

$$\eta_{ij} \equiv g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1}$$

which holds for the whole population, where $\boldsymbol{\beta}$ is a d-dimensional regression parameter. To avoid confusion, in the following let $\boldsymbol{\beta}_0 \equiv (\beta_{01}, \ldots, \beta_{0d})^{\mathrm{T}}$ be the true value of $\boldsymbol{\beta}$. Denote $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^{\mathrm{T}}, \boldsymbol{\beta}_{02}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{01}$ is $d_a \times 1$ vector of the active superpopulation coefficients, and $\boldsymbol{\beta}_{02} \equiv \mathbf{0}$ is a $(d - d_a) \times 1$ vector of the inactive coefficients. Our main goal is to identify the d_a significant variables in model (1) and provide an accurate estimation for the non-zero coefficients. Our estimated d_a is the number of the remaining non-zero β_js from the iterative algorithm presented in Section 3.1.

The GEE approach is a class of estimating equations which take into account the correlation arising due to a longitudinal study design, resulting in the increased efficiency of standard error estimates. For simplicity, denote the sampled response

$\mathbf{y}_i = \left\{\left(y_{i1}, \ldots, y_{im_i}\right)^{\mathrm{T}}\right\}_{m_i\times 1}$ and $\boldsymbol{\eta}_i = \left\{\left(\eta_{i1}, \ldots, \eta_{im_i}\right)^{\mathrm{T}}\right\}_{m_i\times 1}$, where $1 \le m_i \le m$, $1 \le i \le n$. Similarly, let $\mathbf{x}_i = \left\{\left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}\right)^{\mathrm{T}}\right\}_{m_i\times d}$. The mean function in model (1) can be written in matrix notation as $\boldsymbol{\eta}_i = \mathbf{x}_i\boldsymbol{\beta}$. Let $\mu(\cdot) = g^{-1}(\cdot)$ be the inverse of the link function.

Let $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}_i(\mathbf{x}_i) = \mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i)$ be the true covariance of $\mathbf{y}_i$ and let $\mathbf{G}_i = \mathbf{G}(\mathbf{x}_i)$ be the assumed "working" covariance of $\mathbf{y}_i$, where $\mathbf{G}_i = \phi\mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}$, $\mathbf{A}_i$ denotes an $m_i \times m_i$ diagonal matrix that contains the marginal variances of y_{ij}, and $\mathbf{R}_i$ is an invertible working correlation matrix. Throughout, we assume that $\mathbf{R}_i$ can depend on a nuisance finite dimensional parameter vector $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is distinct from $\boldsymbol{\beta}$. The main advantage of the GEE approach is that it yields a consistent estimator even if the working correlation matrix is misspecified.

2.2. Census GEE and survey-weighted GEE

Let $\boldsymbol{\mu}_i = \mu(\mathbf{x}_i\boldsymbol{\beta})$ and $\boldsymbol{\Delta}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\eta}_i^{\mathrm{T}} = \mathrm{diag}\left\{\left(\partial\mu_{ij}/\partial\eta_{ij}\right)_{j=1}^{m_i}\right\}$. If the entire population is observed, we define the population parameter of $\boldsymbol{\beta}$ as the solution of

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{i\in U_N} d(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i\in U_N} \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i(\boldsymbol{\beta})\mathbf{G}_i^{-1}(\boldsymbol{\beta})\left\{\mathbf{y}_i - \mu(\mathbf{x}_i\boldsymbol{\beta})\right\} = \mathbf{0}. \tag{2}$$

In practice, we generally do not observe the values for the whole population but only for those in a sample drawn from the population. As in [3], we define the survey weighted GEE by

$$\widehat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{i\in s} w_i d(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i\in s} w_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i(\boldsymbol{\beta})\mathbf{G}_i^{-1}(\boldsymbol{\beta})\left\{\mathbf{y}_i - \mu(\mathbf{x}_i\boldsymbol{\beta})\right\} = \mathbf{0}, \tag{3}$$

by introducing the survey weights in (3). As pointed out in Subsection 6.3.1 of [8], when the inclusion probability and the model error are correlated, the unweighted version of the estimators is generally biased. The proposed survey-weighted estimator is robust to misspecifications of the superpopulation estimating equations (2).

2.3. Penalized survey-weighted GEE

We are interested in variable selection for the marginal mean models based on the sample drawn from the finite population. [10] proposed a consistent variable selection method based on the following penalized GEE,

$$\mathbf{S}_N^{\mathcal{P}}(\boldsymbol{\beta}) = \mathbf{S}_N(\boldsymbol{\beta}) - N\mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\,\mathrm{sgn}(\boldsymbol{\beta}) = \mathbf{0} \tag{4}$$

for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$, where $\mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \mathrm{diag}\left\{p'_{\lambda_1}(|\beta_1|), \ldots, p'_{\lambda_d}(|\beta_d|)\right\}$ for some penalty function p_{λ_j} with a regularization parameter λ_j. Various penalty functions have been used in the literature of variable selection for regression models. For example, the hard thresholding (HARD) penalty, $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$; the LASSO penalty [18,19], $p_\lambda(|\beta|) = \lambda|\beta|$; the adaptive LASSO (ALASSO) penalty [22], $p_{\lambda_l}(|\beta|) = \lambda w_l^*|\beta_l|$, for a known data-driven weight w_l^*; and the SCAD penalty [6], $p'_\lambda(\beta) = \lambda\left[I(\beta < \lambda) + (c\lambda - \beta)_+/\{(c-1)\lambda\}I(\beta \ge \lambda)\right]$ for some $c > 2$ and $\beta > 0$. The selection of $\boldsymbol{\lambda}$ will be discussed in Section 3.1.

Define $\boldsymbol{\beta}_N$ to be the solution of (4) based on the whole population. Under general conditions, [10] showed the sparsity of the penalized GEE estimator $\boldsymbol{\beta}_N$. In addition, [10] also showed that with appropriate penalty functions $\boldsymbol{\beta}_N$ behaves asymptotically as if the true model is known a priori, i.e., the "oracle" property given in (9).

We define the survey penalized GEE as

$$\widehat{\mathbf{S}}^{\mathcal{P}}(\boldsymbol{\beta}) = \widehat{\mathbf{S}}(\boldsymbol{\beta}) - N\mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\,\mathrm{sgn}(\boldsymbol{\beta}) = \mathbf{0}. \tag{5}$$

Recently, [3] has shown that the usual GEE software procedures are not appropriate for analyzing longitudinal survey data, even if one specifies the weight variable as the survey weights in (3). Consequently, the direct variable selection procedure in [10] (with survey weights being the weight variable) is not valid for longitudinal surveys, either. Therefore, the theory and practice of the penalized survey-weighted estimator must be re-examined. In the following we study the asymptotic behavior for the proposed estimators.

2.4. Asymptotic results

Let $\mathcal{F}_N = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ be the N-th finite population in the sequence of finite populations $\{\mathcal{F}_N\}$. Let π_i be the inclusion probability for $i \in U_N$ and π_{ik} be the inclusion probability for both elements $i, k \in U_N$. Without loss of generality, we take the sampling weight w_i in (3) and (5) as $1/\pi_i$.

Let $h(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{G}_i^{-1}(\boldsymbol{\beta}) \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{x}_i$. We assume the following conditions:

(C1) the sampling fraction $\lim_{N\to\infty} n/N = f$ for some $0 < f \leq 1$;
(C2) the inclusion probabilities satisfy that $\min_{i\in U_N} \pi_i \geq a > 0$, $\min_{i\neq k} \pi_{ik} \geq a^* > 0$ and $\limsup_{N\to\infty} n \max_{i\neq k} |\pi_{ik} - \pi_i\pi_k| \leq C$ for some constant C;
(C3) the sequence of functions $d(\mathbf{x}_i, \boldsymbol{\beta})$ and $h(\mathbf{x}_i, \boldsymbol{\beta})$ are both continuous in $\boldsymbol{\beta}$ for all $\mathbf{x}_i$ and $\boldsymbol{\beta}$ in a closed set $\mathcal{B}$ containing $\boldsymbol{\beta}_0$;
(C4) there exist positive constants M, M_1 and M_2, such that $\|d(\mathbf{x}_i, \boldsymbol{\beta}_0)\| \leq M$, $M_1 \leq \rho_{\min}\{h(\mathbf{x}_i, \boldsymbol{\beta}_0) h^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}_0)\} \leq \rho_{\max}\{h(\mathbf{x}_i, \boldsymbol{\beta}_0) h^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}_0)\} \leq M_2$, for any $i \in U_N$, where $\rho_{\min}$ and $\rho_{\max}$ denote the minimum and maximum eigenvalues of a matrix;
(C5) for all $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_N$, there exists a nonsingular matrix $\Omega(\boldsymbol{\beta})$ such that

$$\lim_{N\to\infty} \frac{n}{N^2} \sum_{i,k\in U_N} \left(\frac{\pi_{ik}}{\pi_i \pi_k} - 1 \right) d(\mathbf{x}_i, \boldsymbol{\beta}) d^{\mathrm{T}}(\mathbf{x}_k, \boldsymbol{\beta}) = \Omega(\boldsymbol{\beta}).$$

Conditions (C1) and (C2) are not unusual in the survey sampling; see, for example, [8,21]. Conditions (C3) and (C4) are common ones to show the consistency of the GEE estimators; see [20]. Condition (C5) is similar to Condition 5 in [1]. We now present the following main theorems whose proofs are given in the Appendix.

Theorem 1. *Under Conditions* (C1)–(C5)*, the survey weighted estimating equation* (3) *has a root* $\widetilde{\boldsymbol{\beta}}$*, such that*

(i) $\widetilde{\boldsymbol{\beta}}$ *is design consistent for the finite population estimator* $\boldsymbol{\beta}_N$ *in the following sense that for any* $\varepsilon > 0$*,*

$$\lim_{N\to\infty} \mathrm{pr}\left\{ \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| > \varepsilon | \mathcal{F}_N \right\} = 0 \quad a.s.,$$

and further

$$\lim_{C\to\infty} \limsup_{N\to\infty} \mathrm{pr}\left\{ \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| > C n^{-1/2} | \mathcal{F}_N \right\} = 0 \quad a.s.;$$

(ii) *As* $N \to \infty$*,*

$$n^{1/2} \mathbf{V}_1^{-1/2} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) | \mathcal{F}_N \to N(\mathbf{0}, \mathbf{I})$$

in distribution, and the asymptotic covariance

$$\mathbf{V}_1 = \lim_{N\to\infty} \mathbf{H}_N^{-1}(\boldsymbol{\beta}_N) \boldsymbol{\Omega}(\boldsymbol{\beta}_N) \mathbf{H}_N^{-1}(\boldsymbol{\beta}_N) \tag{6}$$

with

$$\mathbf{H}_N(\boldsymbol{\beta}) = N^{-1} \sum_{i\in U_N} h(\mathbf{x}_i, \boldsymbol{\beta}) = N^{-1} \sum_{i\in U_N} \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{G}_i^{-1}(\boldsymbol{\beta}) \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{x}_i; \tag{7}$$

(iii) *As* $N \to \infty$*,*

$$\sqrt{n} (\mathbf{V}_1 + f\mathbf{V}_2)^{-1/2} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \mathbf{I})$$

in distribution, where

$$\mathbf{V}_2 = \lim_{N\to\infty} \mathbf{H}_N^{-1}(\boldsymbol{\beta}_0) \mathbf{Q}_N(\boldsymbol{\beta}_0) \mathbf{H}_N^{-1}(\boldsymbol{\beta}_0) \tag{8}$$

with $\mathbf{Q}_N(\boldsymbol{\beta}) = N^{-1} \sum_{i\in U_N} \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{G}_i^{-1}(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i \mathbf{G}_i^{-1}(\boldsymbol{\beta}) \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{x}_i$.

From Theorem 1, one sees that the variance of the survey-weighted estimator consists of two parts: a sampling variance component $\mathbf{V}_1$ and a model variance component $\mathbf{V}_2$. In most applications, $0 < f < 1$, and thus both $\mathbf{V}_1$ and $\mathbf{V}_2$ are positive definite. If $N - n \to 0$ as N increases, for example, we take a complete census so that $n = N$, then the sampling variance component $\mathbf{V}_1 = \mathbf{0}$. Conversely, if the sampling rate is very small, then the sampling variance component $\mathbf{V}_1$ will be the leading term in the joint variance.

Denote

$$\mathbf{q}_{01} = -\left(p'_{\lambda_1}(|\beta_{01}|) \operatorname{sgn}(\beta_{01}), \ldots, p'_{\lambda_{d_a}}\left(|\beta_{0d_a}|\right) \operatorname{sgn}\left(\beta_{0d_a}\right) \right)^{\mathrm{T}},$$

$$\mathbf{J}_{01} = \operatorname{diag}\left\{ p''_{\lambda_1}(|\beta_{01}|), \ldots, p''_{\lambda_{d_a}}\left(|\beta_{0d_a}|\right) \right\}.$$

Write $\boldsymbol{\beta}_N = (\boldsymbol{\beta}_{N1}^{\mathrm{T}}, \boldsymbol{\beta}_{N2}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{N1}^{\mathrm{T}} = (\beta_{1N}, \ldots, \beta_{d_aN})$, $\boldsymbol{\beta}_{N2}^{\mathrm{T}} = (\beta_{(1+d_a)N}, \ldots, \beta_{dN})$. Under some regularity conditions, Theorem 1 of [10] implies that

$$\sqrt{N} \mathbf{V}_{22}^{-1/2} \left\{ \boldsymbol{\beta}_{N1} - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1} \mathbf{q}_{01} \right\} \to N(0, \mathbf{I}) \tag{9}$$

in distribution, where the asymptotic covariance matrix

$$\mathbf{V}_{22} = \lim_{N\to\infty} (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1} \mathbf{Q}_{N1} (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1}$$

with $\mathbf{H}_{N1}$ and $\mathbf{Q}_{N1}$ being the first $d_a \times d_a$ submatrices of $\mathbf{H}_N$ and $\mathbf{Q}_N$.

We write $\mathbf{S}^{\mathcal{P}}(\boldsymbol{\beta}) = (S_1^{\mathcal{P}}(\boldsymbol{\beta}), \ldots, S_d^{\mathcal{P}}(\boldsymbol{\beta}))^{\mathrm{T}}$, where $S_j^{\mathcal{P}}(\boldsymbol{\beta}) = \mathbf{e}_j^{\mathrm{T}} \mathbf{S}^{\mathcal{P}}(\boldsymbol{\beta})$ and $\mathbf{e}_j$ be a d-dimensional vector with a "1" in the jth position and a "0" elsewhere. In the following we denote λ by λ_n to indicate its dependence on the sample size n. We introduce one more condition that is on the penalty.

(C6) The penalty function, $p_{\lambda_n}(\cdot)$, has the following properties: (i) for nonzero fixed θ, $\lim_{n\to\infty} n^{1/2} p'_{\lambda_n}(|\theta|) = 0$ and $\lim_{n\to\infty} n^{1/2} p''_{\lambda_n}(|\theta|) = 0$; (ii) for any positive constant C, $\liminf_{|\theta|\le Cn^{-1/2},\, n\to\infty} \sqrt{n} p'_{\lambda_n}(|\theta|) = \infty$.

Condition (C6) is about the penalty function and regularization parameter, which is also required in [10].

Theorem 2. *Under Conditions* (C1)–(C6)*, there exists an approximate solution to the survey penalized GEE (5),* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_2^{\mathrm{T}})^{\mathrm{T}}$*, such that*

(i) $\widehat{\boldsymbol{\beta}}$ *is design consistent for the finite population estimator* $\boldsymbol{\beta}_{N0} = (\overline{\boldsymbol{\beta}}_{N1}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$*, where* $\overline{\boldsymbol{\beta}}_{N1}$ *is the exact solution to the Census GEE (2) using the first* d_a *auxiliary variables, and* $\widehat{\boldsymbol{\beta}}$ *is a root-n consistent estimator of* $\boldsymbol{\beta}_0$*;*

(ii) $\lim_{N\to\infty} \mathrm{pr}(|\widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}})| = 0,\ j = 1, \ldots, d_a) = 1$*;*

(iii) $\lim_{N\to\infty} \mathrm{pr}(\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}) = 1$*;*

(iv) $\widehat{\boldsymbol{\beta}}$ *is an approximate zero-crossing of* $\widehat{\mathbf{S}}^{\mathcal{P}}$ *in the sense that for any* $j = d_a + 1, \ldots, d$

$$\limsup_{N\to\infty} \limsup_{\epsilon\to 0+} N^{-1} \widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} + \epsilon \mathbf{e}_j) \widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} - \epsilon \mathbf{e}_j) \le 0;$$

(v) *As* $N \to \infty$*,*

$$\sqrt{n}(\mathbf{V}_{11} + f\mathbf{V}_{22})^{-1/2} \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1} \mathbf{q}_{01} \right\} \to N(\mathbf{0}, \mathbf{I})$$

in distribution, where $\mathbf{V}_{11}$ *and* $\mathbf{V}_{22}$ *are the first* $d_a \times d_a$ *submatrix of* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *in (6) and (8), respectively.*

Theorem 2 implies that the penalized estimators with the penalty functions satisfying Condition (C6), such as the HARD, SCAD and ALASSO penalty, have the oracle property.

3. Implementation

3.1. Algorithm and choice of tuning parameters

To estimate the penalized regression coefficients, we consider the Majorize–minorize (MM) algorithm in [9]. Suppose that we have an initial value $\widehat{\boldsymbol{\beta}}^{(0)}$. For example, we can take $\widehat{\boldsymbol{\beta}}^{(0)}$ to be the solution of the survey weighted estimating equation in (3). Denote

$$\widehat{\mathbf{H}}(\boldsymbol{\beta}) = N^{-1} \sum_{i\in s} w_i h(\mathbf{x}_i, \boldsymbol{\beta}) = N^{-1} \sum_{i\in s} w_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{G}_i^{-1}(\boldsymbol{\beta}) \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \mathbf{x}_i. \tag{10}$$

By the local quadratic approximations for penalty functions [6], the MM algorithm can be implemented as follows:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \left\{ \widehat{\mathbf{H}}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) \right\}^{-1} \left\{ N^{-1} \sum_{i\in s} w_i d\left(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^{(k)}\right) - \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) \widehat{\boldsymbol{\beta}}^{(k)} \right\},$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \mathrm{diag}\left\{ \frac{p'_{\lambda_1}(|\beta_1|)}{\epsilon + |\beta_1|}, \ldots, \frac{p'_{\lambda_d}(|\beta_d|)}{\epsilon + |\beta_d|} \right\} \tag{11}$$

for a small ϵ ($\epsilon = 10^{-6}$ in our simulation studies) and tuning penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$. One can repeat the above iteration process until convergence is reached. An estimate $\widehat{\beta}_l^{(k)}$ is treated as zero if its absolute value is smaller than ϵ_0 (a pre-specified value, for example, 10^{-6}), and we delete the l-th component of $\mathbf{x}_i$ from the iteration. We use the convergence criterion such that $\|\widehat{\boldsymbol{\beta}}^{(k+1)} - \widehat{\boldsymbol{\beta}}^{(k)}\| \le \tau$ ($\tau = 10^{-6}$ in our simulation studies). From our experience, the algorithm is quite stable and fast to compute. It usually reaches a reasonable convergence tolerance within a few iterations.

Let $\hat{N} = \sum_{i\in s} w_i$ and $\hat{N}_m = \sum_{i\in s} w_i m_i$. We can estimate the dispersion parameter ϕ by

$$\hat{\phi} = (\hat{N}_m - p)^{-1} \sum_{i\in s} w_i \sum_{j=1}^{m_i} \hat{r}_{ij}^2,$$

where $\hat{r}_{ij} = \{V(\hat{\mu}_{ij})\}^{-1/2}(y_{ij} - \hat{\mu}_{ij})$ and $\hat{\mu}_{ij} = g^{-1}(\mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$. Next, for the working correlation matrix $\mathbf{R} = (\alpha_{jj'})$, we estimate $\alpha_{jj'}$ by

$$\hat{\alpha}_{jj'} = (\hat{N} - p)^{-1}\hat{\phi}^{-1}\sum_{i\in s} w_i \hat{r}_{ij}\hat{r}_{ij'}.$$

If the correlation matrix is assumed to have some special structures, we can estimate it using the method proposed by [11] but with the sampling weights. Taking the exchangeable correlation structure for example, $\mathbf{R} = \mathbf{R}(\alpha) = (1-\alpha)\mathbf{I} + \alpha\mathbf{1}\mathbf{1}^{\mathrm{T}}$ in which $\mathbf{1}$ is a vector with all "1", and we can estimate α by

$$\hat{\alpha} = \hat{\phi}\sum_{i\in s}\sum_{j>j'} w_i \hat{r}_{ij}\hat{r}_{ij'} \Big/ \left\{\frac{1}{2}\sum_{i\in s} w_i m_i(m_i - 1) - p\right\}.$$

We use the Bayesian information criterion to select the tuning penalty parameters. Let

$$e(\boldsymbol{\lambda}) = \mathrm{tr}\left[\left\{\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right\}^{-1}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right]$$

be the effective number of parameters in the last step of the Newton–Raphson iteration. We define

$$\mathrm{BIC}(\boldsymbol{\lambda}) = \log\left\{\hat{N}_m^{-1}\sum_{i\in s}\sum_{j=1}^{m_i} w_i \hat{r}_{ij}^2\right\} + \hat{N}_m^{-1}\log(\hat{N}_m)e(\boldsymbol{\lambda}).$$

We select $\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\mathrm{BIC}(\boldsymbol{\lambda})$.

3.2. Standard error estimation

The standard errors for the estimated parameters can be obtained as follows. For a given sample design with inclusion probabilities π_i and π_{ik}, we estimate $\boldsymbol{\Omega}$ by

$$\widehat{\boldsymbol{\Omega}} = \frac{n}{N^2}\sum_{i,k\in s}\frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}\pi_i^{-1}\pi_k^{-1}d(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})d^{\mathrm{T}}(\mathbf{x}_k, \widehat{\boldsymbol{\beta}}),$$

a design-consistent estimator of $\boldsymbol{\Omega}$. Let $\widehat{\mathbf{J}}_1 = \mathrm{diag}\left\{p''_{\lambda_1}(|\widehat{\beta}_1|), \ldots, p''_{\lambda_{d_a}}(|\widehat{\beta}_{d_a}|)\right\}$. Then $\mathbf{V}_{11}$ can be estimated by

$$\widehat{\mathbf{V}}_{11} = (\widehat{\mathbf{H}}_1 + \widehat{\mathbf{J}}_1)^{-1}\widehat{\boldsymbol{\Omega}}_1(\widehat{\mathbf{H}}_1 + \widehat{\mathbf{J}}_1)^{-1},$$

where $\widehat{\mathbf{H}}_1$ and $\widehat{\boldsymbol{\Omega}}_1$ are the first $d_a \times d_a$ submatrix of $\widehat{\mathbf{H}}$ and $\widehat{\boldsymbol{\Omega}}$, respectively. Similarly, let

$$\widehat{\mathbf{Q}}_N = \sum_{i\in s} w_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}_i^{-1}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_i\mathbf{G}_i^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Delta}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_i.$$

A consistent estimator of $\mathbf{V}_{22}$ is given by

$$\widehat{\mathbf{V}}_{22} = (\widehat{\mathbf{H}}_1 + \widehat{\mathbf{J}}_1)^{-1}\widehat{\mathbf{Q}}_1(\widehat{\mathbf{H}}_1 + \widehat{\mathbf{J}}_1)^{-1},$$

where $\widehat{\mathbf{Q}}_1$ is the first $d_a \times d_a$ submatrix of $\widehat{\mathbf{Q}}_N$.

In complex surveys, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is often too complicated to calculate directly. The delete-one-cluster Jackknife variance estimator was proposed by [15]. It is possible to have many sets of customary jackknife weights for which the above estimation algorithm would not converge due to ill-conditioned matrices that are not invertible. To avoid the inversion of possibly ill-conditioned matrices, one can use the estimating function (EF) bootstrap approach proposed by [2,14]. Let $\left\{w_i^{(b)}\right\}_{i\in s}$ be the weights for the bootstrap sample s_b^*, $b = 1, \ldots, B$. For example, in cluster sampling, the survey units forming the b-th bootstrap replicate are obtained by sampling $l_c - 1$ clusters independently with replacement from the l_c sampled cluster. For the b-th bootstrap sample, each weight $w_i^{(b)}$ is created by adjusting the survey weight variable for the ith unit to account for the results of the replicate sampling and any other adjustments done to the survey weight; see Eq. (14) below. The one-step bootstrap estimators from the bootstrap sample b are then given by

$$\widehat{\boldsymbol{\beta}}^{(b)} = \widehat{\boldsymbol{\beta}} + (\widehat{\mathbf{H}} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}})^{-1} \times \left\{\frac{1}{N}\sum_{i\in s} w_i^{(b)} d(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}}\widehat{\boldsymbol{\beta}}\right\},$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})$ for $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\cdot)$ given in (11). The EF-bootstrap estimator of $\mathbf{V}_1$ in (6) is given by

$$\widehat{\mathbf{V}}_1^{\mathrm{EF}} = \frac{n}{B}\sum_{b=1}^{B}\left(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}. \tag{12}$$

Our limited simulation results indicate that this variance estimator is useful even in some complex sampling context.

Table 1
Example 1. Model selection results.

n	P_λ	WI				EC				AR-1				AR-2				Time (s)
		TP	FP	C	ME	TP	FP	C	ME	TP	FP	C	ME	TP	FP	C	ME	
(a) Sample without weights																		
100	scad	4.79	0.0	80	133.0	4.84	0.0	86	8.7	4.84	0.0	86	20.9	3.20	0.0	23	16.9	3.0
	hard	4.99	0.0	99	92.4	4.88	0.0	89	8.1	4.92	0.0	93	9.5	3.31	0.0	30	15.0	3.3
	alasso	5.00	0.0	100	91.6	5.00	0.0	100	4.6	5.00	0.0	100	5.8	4.74	0.0	81	9.2	1.9
	oracle	5.00	0.0	100	90.5	5.00	0.0	100	3.8	5.00	0.0	100	5.1	5.00	0.0	100	4.3	0.02
200	scad	4.88	0.0	89	58.3	4.93	0.0	95	4.5	4.65	0.0	72	8.2	4.19	0.0	61	11.9	9.0
	hard	5.00	0.0	100	44.4	5.00	0.0	100	3.8	4.97	0.0	97	5.0	4.26	0.0	66	7.2	5.2
	alasso	5.00	0.0	100	44.4	5.00	0.0	100	2.7	5.00	0.0	100	3.0	4.99	0.0	99	3.8	2.3
	oracle	5.00	0.0	100	44.4	5.00	0.0	100	1.8	5.00	0.0	100	2.5	5.00	0.0	100	2.1	0.04
(b) Sample with weights																		
100	scad	4.78	0.0	81	124.8	4.99	0.0	99	8.1	4.32	0.0	53	13.1	4.43	0.0	71	22.8	3.9
	hard	5.00	0.0	100	82.8	5.00	0.0	100	7.3	4.85	0.0	88	9.5	4.51	0.0	76	12.6	5.6
	alasso	5.00	0.0	100	82.9	5.00	0.0	100	5.9	5.00	0.0	100	5.9	4.82	0.0	87	8.4	2.7
	oracle	5.00	0.0	100	82.8	5.00	0.0	100	4.3	5.00	0.0	100	5.8	5.00	0.0	100	6.7	0.02
200	scad	4.86	0.0	87	58.9	5.00	0.0	100	4.3	4.67	0.0	73	8.1	4.66	0.0	83	10.2	7.0
	hard	5.00	0.0	100	45.8	5.00	0.0	100	3.8	4.98	0.0	98	5.2	4.69	0.0	85	5.2	9.2
	alasso	5.00	0.0	100	45.8	5.00	0.0	100	3.2	5.00	0.0	100	3.2	4.95	0.0	96	4.0	5.1
	oracle	5.00	0.0	100	45.7	5.00	0.0	100	2.0	5.00	0.0	100	2.8	5.00	0.0	100	2.7	0.04

4. Simulation studies

In this section, we study the numerical performance of the proposed selection method via some simulation studies.

4.1. Example 1: simple random sampling design

In this example, we consider the simple random sampling design. We simulate $N = 5000$ individuals through the following marginal mean model

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + 0.5\epsilon_{ij}, \quad i = 1, \ldots, 5000,\ j = 1, 2, \ldots, 5$$

where the coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$. There are eight predictors in this example. The first six predictors are generated independently from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_X)$, where $\boldsymbol{\Sigma}_X = (1-\rho_x)\mathbf{I} + \rho_x\mathbf{1}\mathbf{1}^{\mathrm{T}}$ with $\rho_x = 0.5$. The last two predictors are generated by $x_{ij7} = 0.3x_{ij1} + 0.7u_{ij}$ and $x_{ij8} = 0.3x_{ij2} + 0.7v_{ij}$, where u_{ij} and v_{ij} are generated from $N(0, 1)$. The errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{i5})^{\mathrm{T}}$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma}_E)$, where the covariance matrix is exchangeable and $\boldsymbol{\Sigma}_E = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^{\mathrm{T}}$ with $\rho = 0.95$. Note that, unbalanced data are quite common in longitudinal studies, we make 1% of the data missing on purpose, and the missing data are missing completely at random. We select 1000 Monte Carlo simple random samples (SRS) of sizes 100 and 200 from the same population as described above.

To evaluate how the sampling weights affect the selection and estimation results, we consider two penalized GEE methods: one with weights and one without weights. Since we consider SRS in this example, the sampling weights for all the individuals are identical, and $w_i = N/n$ for $i \in s$. We consider the following correlation structures: working independence (WI), exchangeable (EC), AR(1) (AR-1) and AR(2) (AR-2). We compare the model error

$$\mathrm{ME}(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \left\{ \sum_{i \in U_N} \sum_{j=1}^{m} \mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}} \right\} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{13}$$

of the estimators $\widehat{\boldsymbol{\beta}}$ under different correlation structures.

Table 1 reports the selection results with weights and without weights. The columns labeled with "TP" give the average number of the five zero coefficients correctly set to 0; the columns labeled with "FP" give the average number of the three nonzero coefficients incorrectly set to 0; the columns labeled with "C" represent the percentage of times the true model is exactly selected; and the columns labeled with "ME" give the median of the model errors defined in (13). The rows with "scad", "hard" and "alasso" stand respectively for the penalized least squares with the SCAD, HARD and adaptive LASSO penalty. For the SCAD penalty, we did not tune the parameter c. Following [6]'s suggestion, we set $c = 3.7$ to reduce the computation burden. The "oracle" method always identifies the five zero coefficients and three nonzero coefficients correctly since we know the true submodel in this case. From Table 1, one sees that the proposed penalized estimators all behave closer to the oracle estimator as the sample size increases. Table 1 also shows that the regular penalized GEE without using weights and the survey weighted GEE yield very similar results. This is not surprising since the weights associated with individuals are all the same for SRS.

Table 2
Example 1. Bias ratios (%) of the estimated coefficients.

Size	P_λ	WI			EC			AR-1			AR-2		
		β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5
(a) Population													
5000	scad	−0.31	0.39	−0.33	−0.04	−0.22	−0.20	−0.13	−0.29	−0.31	−0.04	−0.23	−0.20
	hard	−0.31	0.39	−0.33	−0.03	−0.12	−0.01	−0.05	−0.14	0.00	−0.04	−0.12	0.00
	alasso	−0.31	0.39	−0.33	−0.02	−0.00	−0.04	−0.04	0.00	−0.06	−0.03	0.03	−0.03
	oracle	−0.31	0.39	−0.33	−0.02	−0.13	0.00	−0.03	0.14	−0.02	−0.01	0.09	−0.02
(b) Sample without weights													
100	scad	0.29	−1.46	−0.49	−0.49	−1.47	−0.14	−0.54	−1.51	−0.20	−0.93	−2.14	−0.06
	hard	0.31	−0.04	−0.24	−0.15	−0.61	−0.34	−0.16	−0.54	−0.33	−0.86	−2.02	−0.09
	alasso	0.31	−0.02	−0.24	−0.04	−0.17	−0.05	−0.06	−0.13	−0.06	−0.38	−0.91	−0.03
	oracle	0.31	0.04	−0.24	−0.01	−0.01	0.01	−0.01	0.05	0.03	−0.01	0.00	0.04
200	scad	0.22	−0.72	−0.51	−0.35	−0.76	−0.52	−0.31	−0.54	−0.40	−0.23	−0.55	−0.32
	hard	0.25	−0.03	−0.22	−0.10	−0.55	−0.27	−0.13	−0.46	−0.27	−0.07	−0.43	−0.17
	alasso	0.25	−0.04	−0.22	−0.04	−0.24	−0.08	−0.06	−0.13	−0.06	−0.03	−0.19	−0.04
	oracle	0.25	−0.03	−0.22	0.00	−0.01	0.01	−0.01	0.03	0.02	−0.01	0.00	−0.04
(c) Sample with weights													
100	scad	0.08	−1.17	−0.81	−0.45	−1.02	−0.13	−0.34	−0.57	−0.45	−0.37	−0.83	−0.53
	hard	0.27	0.20	−0.10	−0.09	−0.66	−0.34	−0.16	−0.56	−0.36	−0.10	−0.60	−0.27
	alasso	0.27	0.19	−0.10	−0.05	−0.35	−0.14	−0.05	−0.09	−0.06	−0.05	−0.27	−0.07
	oracle	0.28	0.20	−0.10	−0.01	0.01	0.01	−0.03	0.08	0.01	−0.01	−0.03	0.02
200	scad	0.22	−0.67	−0.51	−0.28	−0.69	−0.49	−0.30	−0.51	−0.42	−0.27	−0.66	−0.43
	hard	0.26	0.03	−0.21	−0.06	−0.57	−0.28	−0.13	−0.45	−0.30	−0.07	−0.52	−0.24
	alasso	0.26	0.02	−0.21	−0.05	−0.35	−0.15	−0.05	−0.11	−0.08	−0.04	−0.30	−0.10
	oracle	0.26	0.03	−0.21	−0.01	0.00	0.00	−0.03	0.06	−0.01	0.00	−0.03	0.01

It has been pointed out in [8] that if the model error is independent of the inclusion probability, π_i, the estimator remains unbiased; while if the error is correlated with π_i, the estimator will be biased. Table 2 illustrates the relative bias of the estimated coefficients: panel (a) shows $(\beta_{Nj} - \beta_{0j})/\beta_{0j} \times 100\%,\ j = 1, 2, 5$, where β_{Nj} is obtained from the penalized GEE based on population; panels (b) and (c) show the relative design bias $(E_p(\hat{\beta}_j) - \beta_{0j})/\beta_{0j} \times 100\%,\ j = 1, 2, 5$, where $\hat{\beta}_j$ is obtained from the penalized GEE based on sample without sampling weights and with weights, respectively. Comparing panels (b) and (c), one sees that the survey weighted estimators have similar bias to those without using weights.

We now test the accuracy of the asymptotic standard error formula given in Theorem 2 and the EF-bootstrap based standard error in (12). To obtain the EF-bootstrap standard error, we produce 500 bootstrap replicates by taking an SRS of $n - 1$ individuals with replacement from the n sampled individuals. The bootstrap weights in the b-th bootstrap replicate were then obtained by

$$w_i^{(b)} = \frac{n}{n-1} q_i^{(b)} w_i = \frac{N}{n-1} q_i^{(b)}$$

where $q_i^{(b)}$ is the number of repetitions of the ith cluster in the b-th bootstrap replicate.

Table 3 summarizes the standard error estimation results under the correct "exchangeable" working correlation. The median absolute deviation divided by 0.6745, denoted by SD in Table 3, of 1000 estimated coefficients in the 1000 simulations can be regarded as the true standard error. The column labeled "$\mathrm{SD}_{\mathrm{SF}}$" represents the asymptotic standard error shown in Theorem 2. The column labeled "$\mathrm{SD}_{\mathrm{EF}}$" represents the standard error calculated using the EF-bootstrap method. From Table 3, one sees that the SD, $\mathrm{SD}_{\mathrm{SF}}$ and $\mathrm{SD}_{\mathrm{EF}}$ are very close to each other.

4.2. Example 2: stratified cluster sampling design

In this example, we consider a stratified one-stage cluster sampling design. Similarly to Example 1, we generate $N = 5000$ individuals through the following model:

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + 0.5\epsilon_{ij}, \quad i = 1, \ldots, 5000,\ j = 1, 2, \ldots, 5$$

where the coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$ and we include eight predictors in the model. The first seven predictors are time-varying generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_X)$, where $\boldsymbol{\Sigma}_X = (1-\rho_x)\mathbf{I} + \rho_x\mathbf{1}\mathbf{1}^{\mathrm{T}}$ with $\rho_x = 0.5$. The last predictor is the stratification variable, which is time independent, and we consider four equal sized strata, that is, $\mathbf{x}_{ij8} = 1, 2, 3, 4$. The errors $\boldsymbol{\epsilon}_i$ are generated using the same way as in Example 1, and we let 1% of the data be missing completely at random.

Next, we describe the details of our stratified one-stage cluster sampling plan. Within each stratum, we randomly select 5 or 10 clusters and we consider unequal probability sampling without replacement. Let π_{hci} and $\boldsymbol{\epsilon}_{hci}$ be the inclusion

Table 3
Example 1. Standard errors of the estimated coefficients (exchangeable covariance).

	P_λ	β_1			β_2			β_5		
		SD	SD_{SF}	SD_{EF}	SD	SD_{SF}	SD_{EF}	SD	SD_{SF}	SD_{EF}
(a) Population										
5000	scad	–	0.0012	–	–	0.0012	–	–	0.0012	–
	hard	–	0.0012	–	–	0.0012	–	–	0.0012	–
	alasso	–	0.0012	–	–	0.0012	–	–	0.0012	–
(a) Sample with weights										
$n = 100$	scad	0.0093	0.0084	0.0085	0.0084	0.0084	0.0085	0.0095	0.0084	0.0084
	hard	0.0093	0.0084	0.0084	0.0107	0.0084	0.0085	0.0092	0.0083	0.0084
	alasso	0.0086	0.0083	0.0098	0.0101	0.0083	0.0097	0.0098	0.0082	0.0095
$n = 200$	scad	0.0063	0.0060	0.0061	0.0064	0.0060	0.0061	0.0066	0.0059	0.0060
	hard	0.0068	0.0059	0.0060	0.0074	0.0060	0.0061	0.0067	0.0059	0.0060
	alasso	0.0062	0.0059	0.0070	0.0079	0.0059	0.0069	0.0068	0.0058	0.0068

Table 4
Example 2. Model selection results.

n	P_λ	WI				EC				AR-1				AR-2				Time (s)
		TP	FP	C	ME	TP	FP	C	ME	TP	FP	C	ME	TP	FP	C	ME	
(a) Sample without weights																		
100	scad	4.41	0.0	52	209.0	4.05	0.0	36	16.0	4.38	0.0	54	21.4	3.03	0.0	11	39.9	2.0
	hard	4.87	0.0	88	115.8	4.71	0.0	75	12.2	4.83	0.0	85	14.9	3.04	0.0	13	41.0	2.9
	alasso	4.98	0.0	98	98.9	4.99	0.0	99	5.5	4.99	0.0	99	7.5	4.53	0.0	64	17.1	2.0
	oracle	5.00	0.0	100	95.2	5.00	0.0	100	4.6	5.00	0.0	100	5.7	5.00	0.0	100	4.7	0.02
200	scad	4.50	0.0	55	95.2	4.73	0.0	79	13.8	4.62	0.0	70	11.7	3.63	0.0	39	22.4	5.3
	hard	4.95	0.0	95	56.3	4.97	0.0	97	6.7	4.94	0.0	94	7.5	3.73	0.0	45	18.9	5.2
	alasso	4.99	0.0	99	51.1	5.00	0.0	100	5.2	5.00	0.0	100	5.9	4.95	0.0	96	9.8	2.4
	oracle	5.00	0.0	100	51.4	5.00	0.0	100	2.8	5.00	0.0	100	3.5	5.00	0.0	100	3.1	0.03
(b) Sample with weights																		
100	scad	4.83	0.0	83	97.1	4.99	0.0	99	9.5	4.98	0.0	98	21.9	4.36	0.0	71	21.2	6.6
	hard	4.98	0.0	98	78.5	5.00	0.0	100	6.3	5.00	0.0	100	9.7	4.38	0.0	73	11.9	7.4
	alasso	5.00	0.0	100	77.9	5.00	0.0	100	8.2	5.00	0.0	100	9.5	4.82	0.0	86	13.8	4.3
	oracle	5.00	0.0	100	74.0	5.00	0.0	100	4.6	5.00	0.0	100	6.1	5.00	0.0	100	6.8	0.06
200	scad	5.00	0.0	100	51.1	5.00	0.0	100	4.7	5.00	0.0	100	10.0	4.92	0.0	93	8.7	7.8
	hard	4.98	0.0	98	41.4	4.99	0.0	99	3.5	4.99	0.0	099	5.4	4.52	0.0	81	5.3	9.6
	alasso	5.00	0.0	100	40.9	5.00	0.0	100	4.4	5.00	0.0	100	5.2	4.92	0.0	93	5.9	7.8
	oracle	5.00	0.0	100	40.8	5.00	0.0	100	2.2	5.00	0.0	100	3.0	5.00	0.0	100	2.8	0.09

probability and associated measurement error for the ith individual from cluster c in stratum h. Unlike Example 1, in this simulation we consider an informative design, and specifically, we let π_{hci} be proportional to $\sum_{i=1}^{5} \|\epsilon_{hci}\|$. We select all the individuals in the selected clusters. The final sample size is $n = 100$ and $n = 200$. We generate 1000 Monte Carlo samples using the above design from the same population.

To evaluate how the sampling weights affect the selection and estimation results, we consider two penalized GEE methods: one with weights $w_{hci} = \pi_{hci}^{-1}$ and one without weights. Table 4 reports the selection results with weights and without weights. From Table 4, one sees that the proposed penalized estimators all behave closer to the oracle estimator as the sample size increases. Table 4 also implies that the survey weighted penalized GEE improves upon those without weights, regardless of the sampling fractions and the penalization functions, which shows that the sampling weights are not redundant.

Table 5 illustrates the relative design bias of the coefficients estimated without sampling weights and with weights, respectively. From Table 2, one sees that the survey weighted estimators have smaller bias than those without using weights, which confirms the findings in [8] that the unweighted estimators are biased if the model error is correlated with the inclusion probability.

We now test the accuracy of the EF-bootstrap based standard error formula given in (12). To handle the multistage aspect of the sampling within stratum, we produce 500 bootstrap replicates by taking a simple random sample of $l_c - 1$ clusters with replacement from the l_c sample clusters. The bootstrap weights in the b-th bootstrap replicate were then obtained by

$$w_{hci}^{(b)} = \frac{l_c}{l_c - 1} q_c^{(b)} w_{hci}, \tag{14}$$

Table 5
Example 2. Bias ratios (%) of the estimated coefficients.

Size	P_λ	WI			EC			AR-1			AR-2		
		β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5
(a) Population													
5000	scad	0.26	−0.33	0.06	0.03	−0.09	−0.05	0.01	−0.16	−0.01	0.03	−0.09	−0.06
	hard	0.26	−0.33	0.06	0.03	−0.14	−0.04	0.01	−0.37	−0.08	0.03	−0.13	−0.06
	alasso	0.26	−0.33	0.06	0.02	−0.10	−0.08	0.00	−0.15	0.00	0.02	−0.09	−0.09
	oracle	0.26	−0.33	0.06	0.02	−0.01	−0.06	0.03	−0.06	−0.02	0.02	−0.02	−0.04
(b) Sample without weights													
100	scad	0.08	−2.15	−1.31	−0.25	−0.51	−0.43	−0.32	−0.70	−0.58	−0.03	−0.03	−0.11
	hard	0.37	−0.44	−0.25	−0.17	−0.68	−0.47	−0.16	−0.76	−0.52	−0.02	−0.05	−0.11
	alasso	0.38	−0.26	−0.25	−0.03	−0.17	−0.14	−0.02	−0.22	−0.17	−0.02	−0.02	−0.13
	oracle	0.27	−0.63	−0.01	0.03	−0.01	−0.05	−0.05	−0.06	0.01	0.04	−0.02	−0.03
200	scad	0.23	−1.31	−0.69	−0.35	−0.71	−0.59	−0.29	−0.59	−0.52	−0.14	−0.23	−0.28
	hard	0.30	−0.42	−0.19	−0.14	−0.68	−0.45	−0.13	−0.59	−0.42	−0.06	−0.22	−0.23
	alasso	0.30	−0.28	−0.20	−0.04	−0.24	−0.18	−0.03	−0.19	−0.17	−0.04	−0.11	−0.15
	oracle	0.30	−0.58	0.06	−0.03	0.025	−0.05	−0.03	0.03	−0.02	−0.03	0.02	−0.02
(c) Sample with weights													
100	scad	0.21	−0.45	−0.10	−0.13	−0.88	−0.53	−0.39	−1.17	−0.81	−0.16	−0.65	−0.50
	hard	0.21	−0.07	−0.07	−0.01	−0.62	−0.17	−0.03	−0.75	−0.31	−0.02	−0.41	−0.20
	alasso	0.21	−0.11	−0.08	−0.11	−0.73	−0.37	−0.06	−0.59	−0.28	−0.09	−0.46	−0.30
	oracle	0.17	−0.40	−0.05	0.02	−0.02	−0.04	0.02	−0.06	−0.04	0.02	−0.01	−0.07
200	scad	0.22	−0.29	−0.02	−0.08	−0.57	−0.32	−0.06	−0.53	−0.28	−0.07	−0.44	−0.28
	hard	0.22	−0.25	−0.01	−0.01	−0.45	−0.11	−0.01	−0.61	−0.26	−0.02	−0.40	−0.18
	alasso	0.22	−0.29	−0.02	−0.08	−0.57	−0.32	−0.06	−0.53	−0.28	−0.07	−0.44	−0.28
	oracle	0.22	−0.25	0.01	0.00	0.03	−0.02	0.01	−0.04	0.05	0.03	0.01	−0.06

Table 6
Example 2. Standard errors of the estimated coefficients (exchangeable covariance).

Size	P_λ	β_1		β_2		β_5	
		SD	SD_{EF}	SD	SD_{EF}	SD	SD_{EF}
100	scad	0.0111	0.0139	0.0090	0.0092	0.0093	0.0094
	hard	0.0089	0.0103	0.0090	0.0091	0.0093	0.0093
	alasso	0.0099	0.0123	0.0090	0.0090	0.0117	0.0119
200	scad	0.0067	0.0093	0.0068	0.0067	0.0086	0.0086
	hard	0.0062	0.0063	0.0068	0.0068	0.0069	0.0070
	alasso	0.0067	0.0093	0.0068	0.0067	0.0086	0.0086

where w_{hci} is the original sampling weight of the ith individual from the c-th cluster in the h-th stratum, $q_c^{(b)}$ is the number of repetitions of the c-th cluster in the b-th bootstrap replicate.

Table 6 summarizes the estimation results under the correct "exchangeable" working correlation. The column labeled "SD" represents the median absolute deviation of 1000 estimated coefficients in the 1000 simulations divided by 0.6745, which can be regarded as the true standard error. The column labeled "SD_{EF}" represents the standard error calculated using the EF-bootstrap method. These numerical results suggest that the proposed EF-bootstrap procedure in Section 3.2 yields reasonable standard error estimates.

4.3. Example 3: Canadian National Population Health Survey

The Canadian National Population Health Survey (NPHS) is a longitudinal survey on the health of Canadians. [12] fit a logistic model to the Canadian NPHS data, which explains the relationship between the loss of independence among seniors (LOSS) and individual factors associated with their health status, living arrangements and habits. In this example, we simulate two cycles of health survey data similar to the Canadian NPHS using the following model:

$$\text{logit}\,(\text{LOSS}) = -2.5 + 2 * \text{SEX} + 1.2 * \text{AGE} + 0.5 * \text{BMI} + 1.5 * \text{CHRDS} + 1.5 * \text{SMOK}. \tag{15}$$

In the above model, the response variable LOSS is binary: it is 1 if a person has lost independence within the last two years, and it is 0 if remains independent. We generate the LOSS variable using the exchangeable structure with a correlation parameter of 0.1. The covariates in this model are: SEX (binary, 0 for women, 1 for men), AGE (continuous, $N(0, 1)$ after standardization), BMI (continuous, $N(0, 1)$ after standardization), CHRDS (binary, 1 if at least one of 10 chronic conditions is present, 0 otherwise) and SMOK (binary, 1 if presently smokes daily or if quit recently, 0 otherwise). Other variables listed in the NPHS but not included in model (15) are URBRUR (0 if an area of residence is urban, and 1 otherwise), EDU (0 if

Table 7
Example 3. Four ACC (Accessibility) score classes.

Class	Accessibility	pr(ACC)
1	Low	$p < 0.015$
2	Medium low	$0.015 \leq p < 0.048$
3	Medium high	$0.048 \leq p < 0.15$
4	High	$p \geq 0.15$

Table 8
Example 3. Model selection results.

P_λ	WI			EC			AR-1		
	TP	FP	C	TP	FP	C	TP	FP	C
scad	5.962	0.000	97	5.962	0.000	97	5.962	0.000	97
hard	5.940	0.000	94	5.940	0.000	94	5.940	0.000	94
alasso	6.000	0.000	100	6.000	0.000	100	6.000	0.000	100
oracle	6.000	0.000	100	6.000	0.000	100	6.000	0.000	100

Table 9
Example 3. Estimation results of the coefficients (exchangeable covariance).

Est.	P_λ	BIAS	SD_{EF}	RMSE	P_λ	BIAS	SD_{EF}	RMSE
Intercept	oracle	0.0166	0.1105	0.1065	hard	0.0166	0.1102	0.1094
SEX		−0.0264	0.0821	0.0922		−0.0264	0.0825	0.0922
AGE		0.0040	0.0503	0.0504		0.0040	0.0500	0.0504
BMI		−0.0113	0.0427	0.0447		−0.0113	0.0428	0.0447
CHRDS		−0.0506	0.0914	0.1072		−0.0506	0.0922	0.1072
SMOK		0.0047	0.0931	0.0915		0.0047	0.0938	0.0917
Intercept	scad	0.0143	0.1126	0.1092	alasso	0.2687	0.2780	0.2790
SEX		−0.0265	0.0823	0.0923		−0.1890	−0.1951	−0.1957
AGE		0.0041	0.0502	0.0505		−0.0839	−0.0872	−0.0876
BMI		−0.0124	0.0428	0.0489		−0.1032	−0.1029	−0.1029
CHRDS		−0.0511	0.0921	0.1073		−0.2226	−0.2285	−0.2291
SMOK		0.0048	0.0932	0.0916		−0.1605	−0.1671	−0.1678

education is less than postsecondary and 1 otherwise), INC (0 for low income and 1 for medium/high income), LIVING (0 if living alone, 1 if living with spouse or others), ACTIVITY (1 if active and 0 otherwise) and ACH (1 if regular drinker, 0 if occasional or non-drinker). These variables are simulated using some conditional probabilities within age–sex groups. A similar simulation design is used in [2] to compare the properties of some variance estimators.

We generate a finite population of 25,000 individuals, which has some of the characteristics of the Canadian NPHS subpopulation of elderly people, aged 65 and more. Several variables, not used in model (15), URBRUR, EDU and INC, are used to define a "natural" cluster structure of the simulated finite population. Following [2], a fictional propensity score of ACC (accessibility) to medical care facilities is defined as

$$\text{pr}\,(\text{ACC}) = \{1 + \exp\,(2 + 2.5 * \text{URBRUR} - 0.8 * \text{EDU} - 0.75 * \text{INC} + 1.0 * \text{LOSS})\}^{-1}$$

for each individual. The individual records are placed into four classes according to the value of their accessibility score, as shown in Table 7.

The records are sorted according to the accessibility class, but with random order within the class. We then arrange the finite population into clusters whose sizes are generated as random integers between 30 and 50, using the uniform distribution $U(30, 50)$. Clustering resulted in 628 clusters. In this way, individuals with similar values of "accessibility" are placed in the same cluster.

We consider one-stage cluster sampling. A sample of clusters is selected without replacement with selection probability proportional to the cluster size using the "Sampford" method in SAS procedure `PROC SURVEYSELECT`. We select 50 clusters from the above population, and repeat the selection 500 times to get 500 samples of 50 clusters.

We apply the proposed variable selection procedures to the selected sample. The covariates in the full model include all the 11 variables described above: SEX, AGE, BMI, CHRDS, SMOK, URBRUR, EDU, INC, LIVING, ACTIVITY and ACH. The true submodel only contains the first five variables. The selection results are summarized in Table 8, and the estimation results, including the design bias (BIAS), the EF-bootstrap standard error (SD_{EF}) and the root mean squared error (RMSE), are listed in Table 9. We do not observe big difference among three correlation structures in Tables 8 and 9. This is not unexpected because only two cycles of the data are simulated.

Following a referee's suggestion, we have also carried out simulations for longitudinal surveys with a large number of covariates, for example, $d = 50$ for $n = 100$ sampled individuals and $d = 100$ for $n = 200$ sampled individuals. We find that our proposed variable selection method still performs well for high dimensional longitudinal surveys. Due to the space limitation, the detailed results are omitted.

5. Discussion

In this paper, we show, both theoretically and empirically, the properties of the survey-weighted penalized GEE estimators for longitudinal surveys. The results can be extended to other penalized estimators defined as the solution to a system of equations based on complex survey designs. To obtain the oracle property, we need to make a number of technical assumptions [1], for example, the assumptions of existence of moments for the superpopulation, assumptions about the functions defining the estimator and assumptions about the design. Our results show that survey weights should be accounted in our variable selection and estimation approach for the analysis of longitudinal surveys when the sampling design is informative, which is parallel to the findings in [3]. Furthermore, our simulation studies indicate that the estimating function bootstrap works well for moderate sample sizes. While the weights used in the main theorem are the design weights, alternative weights are also possible. How to weight the survey data wisely in order to select the best model would be an interesting future research topic. In addition, we find that the working covariance matrix is less important for model selection than for the estimation of the standard error for the coefficients and the simultaneous consideration of correlation with high dimensionality and the characteristics of sampling design presents great challenges in our survey analysis. As a result, we would like to leave the covariance matrix selection problem for future research.

Acknowledgments

The research of L. Wang was partially supported by NSF grants DMS-0905730, DMS-1106816, DMS-1309800 and the ASA/NSF/BLS research fellow program. The research of S. Wang was partially supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the US Bureau of Labor Statistics.

Appendix

A.1. Technical details

In this appendix, we derive the asymptotic properties of $\widetilde{\boldsymbol{\beta}}$. Denote by $E_p[\cdot]$ the expectation with respect to the sampling design. Let $\mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \mu(\mathbf{x}_i\boldsymbol{\beta})$ and $z_i = w_i I_i - 1 = I_i/\pi_i - 1$. Let $\nabla\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\mathbf{S}(\boldsymbol{\beta})$.

Lemma A.1. *If Conditions* (C1)–(C4) *hold, there exists a positive constant* C_1 *such that,*

$$\frac{n}{N^2}E_p\left\|\widehat{\mathbf{S}}\left(\boldsymbol{\beta}_N\right)\right\|^2 \leq C_1.$$

Proof. Noting that $\mathbf{S}_N\left(\boldsymbol{\beta}_N\right) = \mathbf{0}$, we can write

$$\left\|\widehat{\mathbf{S}}\left(\boldsymbol{\beta}_N\right)\right\|^2 = \left\|\widehat{\mathbf{S}}\left(\boldsymbol{\beta}_N\right) - \mathbf{S}_N\left(\boldsymbol{\beta}_N\right)\right\|^2 = \sum_{i,k\in U_N} z_i z_k d^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}_N) d(\mathbf{x}_k, \boldsymbol{\beta}_N),$$

by Conditions (C1)–(C3), we obtain

$$\begin{aligned}
\frac{n}{N^2}E_p\left\|\widehat{\mathbf{S}}\left(\boldsymbol{\beta}_N\right) - \mathbf{S}_N\left(\boldsymbol{\beta}_N\right)\right\|^2 &= \frac{n}{N^2}E_p\left\{\sum_{i,k\in U_N} z_i z_k d^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}_N) d(\mathbf{x}_k, \boldsymbol{\beta}_N)\right\} \\
&\leq \frac{n}{N^2}\sum_{i,k\in U_N}\left|\frac{\pi_{ik}}{\pi_i\pi_k} - 1\right| d^{\mathrm{T}}(\mathbf{x}_i, \boldsymbol{\beta}_N) d(\mathbf{x}_k, \boldsymbol{\beta}_N) \\
&\leq \left[\frac{n}{N}\frac{1}{a} + \frac{n}{a^2}\max_{i\neq k}|\pi_{ik} - \pi_i\pi_k|\right]\frac{1}{N}\sum_{i\in U_N}\|d(\mathbf{x}_i, \boldsymbol{\beta}_N)\|^2 \\
&\leq \left[\frac{n}{N}\frac{1}{a} + \frac{n}{a^2}\max_{i\neq k}|\pi_{ik} - \pi_i\pi_k|\right]\left\{\frac{1}{N}\sum_{i\in U_N}\|d(\mathbf{x}_i, \boldsymbol{\beta}_0)\|^2 + o(1)\right\}.
\end{aligned}$$

The desired result holds from Condition (C4). ■

Lemma A.2. *If Conditions* (C1)–(C4) *hold, there exists a positive constant* C_2 *such that, for any d-dimensional vector* $\mathbf{a}$ *with* $\|\mathbf{a}\| = 1$,

$$nE_p\left\|\mathbf{a}^{\mathrm{T}}\{\widehat{\mathbf{H}}\left(\boldsymbol{\beta}_N\right) - \mathbf{H}_N\left(\boldsymbol{\beta}_N\right)\}\mathbf{a}\right\|^2 \leq C_2,$$

where $\mathbf{H}_N(\cdot)$ *and* $\widehat{\mathbf{H}}(\cdot)$ *are given in* (7) *and* (10)*, respectively.*

Proof. Note that

$$\mathbf{a}^{\mathrm{T}}\{\widehat{\mathbf{H}}\left(\boldsymbol{\beta}_N\right)-\mathbf{H}_N\left(\boldsymbol{\beta}_N\right)\}\mathbf{a}=\sum_{i\in U_N}\frac{1}{N}z_i\mathbf{a}^{\mathrm{T}}h(\mathbf{x}_i,\boldsymbol{\beta})\mathbf{a}.$$

Using similar arguments in the proof of Lemma A.1, under Conditions (C1) and (C2) we have

$$\begin{aligned}nE_p|\mathbf{a}^{\mathrm{T}}\{\widehat{\mathbf{H}}\left(\boldsymbol{\beta}_N\right)-\mathbf{H}_N\left(\boldsymbol{\beta}_N\right)\}\mathbf{a}|^2&=\frac{n}{N^2}E_p\left|\sum_{i\in U_N}z_i\mathbf{a}^{\mathrm{T}}h(\mathbf{x}_i,\boldsymbol{\beta}_N)\mathbf{a}\right|^2\\&\le\left[\frac{n}{N}\frac{1}{a}+\frac{n}{a^2}\max_{i\neq k}|\pi_{ik}-\pi_i\pi_k|\right]\frac{1}{N}\sum_{i\in U_N}\left\|h(\mathbf{x}_i,\boldsymbol{\beta}_N)\right\|^2.\end{aligned}$$

The desired result follows from Conditions (C3) and (C4). ■

A.2. Proof of Theorem 1

We first evaluate the sign of $(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\widehat{\mathbf{S}}(\boldsymbol{\beta})$ on $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|=C_3n^{-1/2}\}$. We start with the expansion

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\widehat{\mathbf{S}}(\boldsymbol{\beta})=(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)+(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\nabla\widehat{\mathbf{S}}\left(\boldsymbol{\beta}^*\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_N),\tag{A.1}$$

where $\boldsymbol{\beta}^*=t\boldsymbol{\beta}+(1-t)\boldsymbol{\beta}_N$, for some $t\in[0,1]$. By Lemma A.1, $|(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)|\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|\times O_p(N/\sqrt{n})=C_3O_p(N/n)$.

For the second term in (A.1), similarly to [20], we can show that

$$\begin{aligned}&|(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\{-\nabla\widehat{\mathbf{S}}(\boldsymbol{\beta})-N\widehat{\mathbf{H}}(\boldsymbol{\beta})\}(\boldsymbol{\beta}-\boldsymbol{\beta}_N)|\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|^2O_p(\sqrt{N})=C_3^2O_p(\sqrt{N}/n),\\&|(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\{\widehat{\mathbf{H}}(\boldsymbol{\beta})-\widehat{\mathbf{H}}\left(\boldsymbol{\beta}_N\right)\}(\boldsymbol{\beta}-\boldsymbol{\beta}_N)|\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|^2O_p(n^{-1/2})=C_3^2O_p(n^{-1/2}).\end{aligned}$$

By Lemma A.2,

$$|(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\{\widehat{\mathbf{H}}\left(\boldsymbol{\beta}_N\right)-\mathbf{H}_N\left(\boldsymbol{\beta}_N\right)\}(\boldsymbol{\beta}-\boldsymbol{\beta}_N)|\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|^2O_p(n^{-1/2})=C_3^2O_p(n^{-3/2}).$$

By the definition of $\mathbf{H}_N\left(\boldsymbol{\beta}_N\right)$ and Conditions (C3)–(C4),

$$M_1C_3^2n^{-1}\le|(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\mathbf{H}_N\left(\boldsymbol{\beta}_N\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_N)|\le M_2C_3^2n^{-1}.$$

Therefore,

$$M_1C_3^2N/n\le|-(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\nabla\widehat{\mathbf{S}}\left(\boldsymbol{\beta}^*\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_N)|\le M_2C_3^2N/n.$$

Thus, the first term in (A.1) is asymptotically dominated in probability by the second term on $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|=C_3n^{-1/2}\}$. For $\varepsilon>0$, there exists a constant $C_3>0$ such that for all n sufficiently large,

$$\mathrm{pr}\left\{\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_N\|=C_3n^{-1/2}}(\boldsymbol{\beta}-\boldsymbol{\beta}_N)^{\mathrm{T}}\widehat{\mathbf{S}}(\boldsymbol{\beta})<0\,\middle|\,\mathcal{F}_N\right\}\ge1-\varepsilon,\quad\text{a.s.},$$

which is sufficient to ensure the existence of a sequence of roots $\widetilde{\boldsymbol{\beta}}$ of (3); see [20]. Hence, we have obtained (i):

$$\lim_{C_3\to\infty}\limsup_{N\to\infty}\mathrm{pr}\left\{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N\|>C_3n^{-1/2}|\mathcal{F}_N\right\}=0\quad\text{a.s.}$$

Next, we establish the asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ in (ii). Note that

$$\mathbf{0}=\frac{\sqrt{n}}{N}\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)+\sqrt{n}\mathbf{H}_N(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)+o_p(1),$$

which implies that the asymptotic distribution of $n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)$ is the same as the asymptotic distribution of $n^{1/2}N^{-1}\mathbf{H}_N^{-1}\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)$. We obtain

$$n^{1/2}N^{-1}\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)=n^{1/2}N^{-1}\left\{\widehat{\mathbf{S}}(\boldsymbol{\beta}_N)-\mathbf{S}_N(\boldsymbol{\beta}_N)\right\}=n^{1/2}N^{-1}\sum_{i\in U_N}z_id(\mathbf{x}_i,\boldsymbol{\beta}_N),$$

in which $E\left[n^{1/2}N^{-1}\sum_{i\in U_N}z_id(\mathbf{x}_i,\boldsymbol{\beta}_N)\,\middle|\,\mathcal{F}_N\right]=\mathbf{0}$, and

$$\mathrm{Var}\left[\frac{n^{1/2}}{N}\sum_{i\in U_N}z_id(\mathbf{x}_i,\boldsymbol{\beta})\,\middle|\,\mathcal{F}_N\right]=\frac{n}{N^2}\sum_{i,k\in U_N}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k}d(\mathbf{x}_i,\boldsymbol{\beta})d^{\mathrm{T}}(\mathbf{x}_k,\boldsymbol{\beta}).$$

By Condition (C5),

$$\lim_{N\to\infty}\frac{n}{N^2}\sum_{i,k\in U_N}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k}d(\mathbf{x}_i,\boldsymbol{\beta})d^{\mathrm{T}}(\mathbf{x}_k,\boldsymbol{\beta}_N)=\Omega(\boldsymbol{\beta}),$$

where $\Omega(\boldsymbol{\beta})$ is positive definite. By the Central Limit Theorem,

$$\frac{n^{1/2}}{N}\left\{\widehat{\mathbf{S}}(\boldsymbol{\beta})-\mathbf{S}_N(\boldsymbol{\beta})\right\}\Big|\mathcal{F}_N\to N(\mathbf{0},\Omega(\boldsymbol{\beta}))$$

in distribution.

Finally, we prove (iii). According to (9), for $f>0$, as $N\to\infty$

$$n^{1/2}(\boldsymbol{\beta}_N-\boldsymbol{\beta}_0)\to N(\mathbf{0},f\mathbf{V}_2)$$

in distribution. By the above result (ii), we have

$$n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)|\mathcal{F}_N\to N(\mathbf{0},\mathbf{V}_1) \tag{A.2}$$

in distribution. Let $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ denote the normal cumulative distribution function with mean $\mathbf{0}$ and variance $\mathbf{V}_1$ and $f\mathbf{V}_2$, respectively. Further, let $\Phi_3(\cdot)$ denote the normal cumulative distribution function with mean $\mathbf{0}$ and variance $\mathbf{V}_1+f\mathbf{V}_2$, which is the convolution of $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$. We obtain

$$\left|\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\le\mathbf{t}\right\}-\Phi_3(\mathbf{t})\right|=\left|E\left[\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\le\mathbf{t}|\mathcal{F}_N\right\}\right]-\Phi_3(\mathbf{t})\right|+o(1). \tag{A.3}$$

Let $\mathbf{s}=\mathbf{t}-n^{1/2}(\boldsymbol{\beta}_N-\boldsymbol{\beta}_0)$, which is a random variable due to $\boldsymbol{\beta}_N$. Then we have

$$\begin{aligned}&\left|E\left[\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\le\mathbf{t}|\mathcal{F}_N\right\}\right]-\Phi_3(\mathbf{t})\right|=\left|E\left[\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)\le\mathbf{s}|\mathcal{F}_N\right\}\right]-\Phi_3(\mathbf{t})\right|\\&\le\left|E\left[\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)\le\mathbf{s}|\mathcal{F}_N\right\}\right]-E\{\Phi_1(\mathbf{s})\}\right|+|E\{\Phi_1(\mathbf{s})\}-\Phi_3(\mathbf{t})|.\end{aligned}$$

By the Dominated Convergence Theorem and (A.2),

$$\begin{aligned}&\lim_{N\to\infty}E\left|\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)\le\mathbf{s}\Big|\mathcal{F}_N\right\}-E\{\Phi_1(\mathbf{s})\}\right|\\&\le E\left\{\lim_{N\to\infty}\left[\sup_{\mathbf{s}}\left|\mathrm{pr}\left\{n^{1/2}(\widetilde{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_N)\le\mathbf{s}|\mathcal{F}_N\right\}-E\{\Phi_1(\mathbf{s})\}\right|\right]\right\}=0.\end{aligned} \tag{A.4}$$

On the other hand, using the Dominated Convergence Theorem again, we have

$$\lim_{N\to\infty}E\{\Phi_1(\mathbf{s})\}=E\left\{\lim_{N\to\infty}\Phi_1\left[\mathbf{t}-n^{1/2}(\boldsymbol{\beta}_N-\boldsymbol{\beta}_0)\right]\right\}.$$

Thus, by (9)

$$\lim_{N\to\infty}E\{\Phi_1(\mathbf{s})\}=\Phi_1*\Phi_2(\mathbf{t})=\Phi_3(\mathbf{t}). \tag{A.5}$$

The desired result then follows from (A.3)–(A.5).

A.3. Proof of Theorem 2

Let $\widehat{\boldsymbol{\beta}}_1$ be the exact solution to the survey-weighted GEE (3) using the first d_a auxiliary variables. Let $\widehat{\boldsymbol{\beta}}=(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{0}$ is a $(d-d_a)$-dim vector. In the following, we show that $\widehat{\boldsymbol{\beta}}$ possesses properties (i)–(v) in the theorem. Applying the same arguments as that in Theorem 1 to the model with first d_a variables only, one can show that $\widehat{\boldsymbol{\beta}}_1$ is a design consistent estimator of $\overline{\boldsymbol{\beta}}_{N1}$, and $\widehat{\boldsymbol{\beta}}_1$ is a root-n consistent estimator of $\boldsymbol{\beta}_{01}$. Therefore, (i) is seen to be valid.

Denote $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_1^{\mathcal{P}}$ the first d_a-components of $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{S}}^{\mathcal{P}}$, respectively. We consider the boundary of a ball around $\boldsymbol{\beta}_{01}$, $\{\boldsymbol{\beta}_1\in R^{d_a}:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|=C_4n^{-1/2}\}$ for a positive constant C_4. Then we have for any $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$,

$$\begin{aligned}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}\widehat{\mathbf{S}}_1^{\mathcal{P}}(\boldsymbol{\beta})&=(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}\widehat{\mathbf{S}}_1(\boldsymbol{\beta})-(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}N\mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_1)\mathrm{sgn}(\boldsymbol{\beta}_1)\\&=(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}\widehat{\mathbf{S}}_1((\boldsymbol{\beta}_1^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}})+(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}\nabla\widehat{\mathbf{S}}_1\left((\boldsymbol{\beta}^*,\mathbf{0})^{\mathrm{T}}\right)(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})\\&\quad-(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^{\mathrm{T}}N\mathrm{diag}\{q'_\lambda(|\beta_j^*|)\mathrm{sgn}(\beta_{0j})\}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})\\&=I_1+I_2+I_3,\end{aligned} \tag{A.6}$$

where β_j^* is between β_j and β_{0j}, for $j = 1, \ldots, d_a$. By Lemma A.1,

$$|I_1| = |(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})^{\mathrm{T}} \widehat{\mathbf{S}}_1(\boldsymbol{\beta}_0)| \leq C_4 n^{-1/2} \|\widehat{\mathbf{S}}_1(\boldsymbol{\beta}_0)\| \leq C_4 C_1 n^{-1} N.$$

Using arguments similar to that in the proof of Theorem 1(i), we have

$$|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})^{\mathrm{T}} \{-\nabla \widehat{\mathbf{S}}_1(\boldsymbol{\beta}) - \mathbf{H}_{N1}(\boldsymbol{\beta}_0)\}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})| = C_4^2 O_p(\sqrt{N}/n),$$

and by Conditions (C3)–(C4),

$$M_1 C_4^2 N/n \leq |(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})^{\mathrm{T}} \mathbf{H}_{N1}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})| \leq M_2 C_4^2 N/n.$$

Therefore,

$$M_1 C_4^2 N/n \leq |I_2| \leq M_2 C_4^2 N/n.$$

By Condition (C6), $\max_j p''(|\beta_j^*|) \to 0$, so $I_3 = o(N/n)$. Thus, for C_4 large enough, both I_1 and I_3 in (A.6) are asymptotically dominated in probability by I_2 on $\{\boldsymbol{\beta}_1 \in R^{d_a} : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = C_4 n^{-1/2}\}$. For $\varepsilon > 0$, there exists a constant $C_4 > 0$ such that for all n sufficiently large,

$$\mathrm{pr}\left\{ \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = C_4 n^{-1/2}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})^{\mathrm{T}} \widehat{\mathbf{S}}_1^{\mathcal{P}}(\boldsymbol{\beta}) < 0 \right\} \geq 1 - \varepsilon.$$

Hence, $\widehat{\mathbf{S}}_1^{\mathcal{P}}((\boldsymbol{\beta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) = \mathbf{0}$ has a solution within this ball. Note that $\widehat{\mathbf{S}}_1((\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) = \mathbf{0}$, and according to Condition (C6)(i), the penalty term is approximately 0. Furthermore, $\widehat{\boldsymbol{\beta}}$ satisfies that $\lim_{N\to\infty} \mathrm{pr}(\widehat{\mathbf{S}}_1^{\mathcal{P}}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}) = 1$, which is (ii). Result (iii) is true since $\widehat{\boldsymbol{\beta}}$ is constructed in that way.

Next, we show (iv). Note that

$$n^{-1/2} \widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} \pm \epsilon \mathbf{e}_j) = n^{-1/2} \widehat{S}_j(\widehat{\boldsymbol{\beta}} \pm \epsilon \mathbf{e}_j) - \frac{N}{\sqrt{n}} q_\lambda(|\widehat{\beta}_j \pm \epsilon|) \mathrm{sgn}(\widehat{\beta}_j \pm \epsilon).$$

According to Condition (C6), for $j = d_a + 1, \ldots, d$, the first term is dominated by the second term. Thus, the signs of $\widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} \pm \epsilon \mathbf{e}_j)$ depend on the signs of $\mp q_\lambda(\epsilon)$ when ϵ goes to zero. Thus, $\widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} + \epsilon \mathbf{e}_j)$ and $\widehat{S}_j^{\mathcal{P}}(\widehat{\boldsymbol{\beta}} - \epsilon \mathbf{e}_j)$ have different signs as ϵ goes to zero.

Finally, we show (v) on the asymptotic normality of the estimator $\widehat{\boldsymbol{\beta}}$. Let $\mathbf{J}_N = \mathrm{diag}\left\{ p''_{\lambda_1}(|\beta_{1N}|), \ldots, p''_{\lambda_d}(|\beta_{dN}|) \right\}$. We have

$$\mathbf{0} = \frac{\sqrt{n}}{N} \widehat{\mathbf{S}}(\boldsymbol{\beta}_N) - \sqrt{n} \mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_N) \mathrm{sgn}\left(\boldsymbol{\beta}_N\right) + \sqrt{n} \left\{ \mathbf{H}_N + \mathbf{J}_N + o_p(1) \right\} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N),$$

which implies that the asymptotic distribution of $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$ is the same as the asymptotic distribution of $(\mathbf{H}_N + \mathbf{J}_N)^{-1} n^{1/2} \left\{ \frac{1}{N} \widehat{\mathbf{S}}(\boldsymbol{\beta}_N) - \mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_N) \mathrm{sgn}\left(\boldsymbol{\beta}_N\right) \right\}$, as $N \to \infty$. Noting that $\mathbf{S}_N(\boldsymbol{\beta}_N) - N \mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_N) \mathrm{sgn}\left(\boldsymbol{\beta}_N\right) = \mathbf{0}$, we obtain

$$\begin{aligned} n^{1/2} \left\{ \frac{1}{N} \widehat{\mathbf{S}}(\boldsymbol{\beta}_N) - \mathbf{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_N) \mathrm{sgn}\left(\boldsymbol{\beta}_N\right) \right\} &= \frac{n^{1/2}}{N} \left\{ \widehat{\mathbf{S}}(\boldsymbol{\beta}_N) - \mathbf{S}_N(\boldsymbol{\beta}_N) \right\} \\ &= \frac{n^{1/2}}{N} \sum_{i \in U_N} z_i d(\mathbf{x}_i, \boldsymbol{\beta}_N). \end{aligned}$$

By the Central Limit Theorem,

$$n^{1/2} N^{-1} \left\{ \widehat{\mathbf{S}}(\boldsymbol{\beta}) - \mathbf{S}_N(\boldsymbol{\beta}) \right\} \Big| \mathcal{F}_N \to N(\mathbf{0}, \Omega(\boldsymbol{\beta}))$$

in distribution. Finally, we prove (iii). According to (9), for $f > 0$

$$n^{1/2} \left\{ \boldsymbol{\beta}_{N1} - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1} \mathbf{q}_{01} \right\} \to N(\mathbf{0}, f \mathbf{V}_{22})$$

in distribution. By Theorem 1, we have

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{N1}) | \mathcal{F}_N \to N(\mathbf{0}, \mathbf{V}_{11})$$

in distribution. Let $\Phi_{11}(\cdot)$ and $\Phi_{22}(\cdot)$ denote the normal cumulative distribution function with mean $\mathbf{0}$ and variance $\mathbf{V}_{11}$ and $f\mathbf{V}_{22}$, respectively. Further, let $\Phi_{33}(\cdot)$ denote the normal cumulative distribution function with mean $\mathbf{0}$ and variance $\mathbf{V}_{11} + f\mathbf{V}_{22}$, which is the convolution of $\Phi_{11}(\cdot)$ and $\Phi_{22}(\cdot)$. We obtain

$$\begin{aligned} &\left| \mathrm{pr}\left[n^{1/2} \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_1)^{-1} \mathbf{q}_{01} \right\} \leq \mathbf{t} \right] - \Phi_{33}(\mathbf{t}) \right| \\ &= \left| E\left(\mathrm{pr}\left[n^{1/2} \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_1)^{-1} \mathbf{q}_{01} \right\} \leq \mathbf{t} \Big| \mathcal{F}_N \right] \right) - \Phi_{33}(\mathbf{t}) \right| + o(1). \end{aligned} \tag{A.7}$$

Let $\mathbf{s} = \mathbf{t} - n^{1/2}\left\{\boldsymbol{\beta}_{N1} - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1}\mathbf{q}_{01}\right\}$, which is a random variable due to $\boldsymbol{\beta}_{N1}$. Then we have

$$\begin{aligned}\left|E\left(\mathrm{pr}\left[n^{1/2}\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_1)^{-1}\mathbf{q}_{01}\right\} \le \mathbf{t}\middle|\mathcal{F}_N\right]\right) - \Phi_{33}(\mathbf{t})\right| &= \left|E\left(\mathrm{pr}\left\{n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{N1}) \le \mathbf{s}|\mathcal{F}_N\right\}\right) - \Phi_{33}(\mathbf{t})\right| \\ &\le \left|E\left(\mathrm{pr}\left\{n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{N1}) \le \mathbf{s}|\mathcal{F}_N\right\}\right) - E\{\Phi_{11}(\mathbf{s})\}\right| + |E\{\Phi_{11}(\mathbf{s})\} - \Phi_{33}(\mathbf{t})|.\end{aligned}$$

By the Dominated Convergence Theorem and (A.2),

$$\begin{aligned}&\lim_{N\to\infty} E\left|\mathrm{pr}\left\{n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{N1}) \le \mathbf{s}|\mathcal{F}_N\right\} - E\{\Phi_{11}(\mathbf{s})\}\right| \\ &\le E\left\{\lim_{N\to\infty}\left[\sup_{\mathbf{s}}\left|\mathrm{pr}\left\{n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{N1}) \le \mathbf{s}|\mathcal{F}_N\right\} - E\{\Phi_{11}(\mathbf{s})\}\right|\right]\right\} = 0. \qquad \text{(A.8)}\end{aligned}$$

On the other hand, using the Dominated Convergence Theorem again, we have

$$\lim_{N\to\infty} E\{\Phi_{11}(\mathbf{s})\} = E\left\{\lim_{N\to\infty}\Phi_{11}\left[\mathbf{t} - n^{1/2}\left\{\boldsymbol{\beta}_{N1} - \boldsymbol{\beta}_{01} + (\mathbf{H}_{N1} + \mathbf{J}_{01})^{-1}\mathbf{q}_{01}\right\}\right]\right\}.$$

Thus, by (9) we obtain

$$\lim_{N\to\infty} E\{\Phi_{11}(\mathbf{s})\} = \Phi_{11} * \Phi_{22}(\mathbf{t}) = \Phi_{33}(\mathbf{t}). \qquad \text{(A.9)}$$

The desired result then follows from (A.7)–(A.9).

References

[1] D.A. Binder, On the variances of asymptotically normal estimators from complex surveys, Internat. Statist. Rev. 51 (1983) 279–292.
[2] D.A. Binder, M.S. Kovacevic, G. Roberts, Design-based methods for survey data: alternative uses of estimating functions, in: Proceedings of the Survey Research Methods Section, American Statistical Association, American Statistical Association, Washington, DC, 2004, pp. 3301–3312.
[3] I. Carrillo, J. Chen, C. Wu, The pseudo-GEE approach to the analysis of longitudinal survey, Canad. J. Statist. 38 (2010) 540–554.
[4] R.L. Chambers, C. Skinner, S. Wang, Intelligent calibration, Bull. Int. Statist. Inst. 58 (1999) 321–324.
[5] R.G. Clark, R.L. Chambers, Adaptive calibration for prediction of finite population totals, Surv. Methodol. 34 (2008) 163–172.
[6] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[7] W.J. Fu, Penalized estimating equations, Biometrics 59 (2003) 126–132.
[8] W.A. Fuller, Sampling Statistics, John Wiley and Sons, Hoboken, New Jersey, 2009.
[9] D.R. Hunter, R. Li, Variable selection using MM algorithms, Ann. Statist. 33 (2005) 1617–1642.
[10] B. Johnson, D.Y. Lin, D. Zeng, Penalized estimating functions and variable selection in semiparametric regression models, J. Amer. Statist. Assoc. 103 (2008) 672–680.
[11] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[12] L. Martel, A. Bélanger, J.M. Berthelot, Loss and recovery of independence among seniors, Health Rep. 13 (2002) 35–48.
[13] J.N.K. Rao, Marginal modeling for repeated observations: inference with survey data, in: Proceedings of the Survey Research Methods Section, American Statistical Association, American Statistical Association, Washington, DC, 1998, pp. 76–82.
[14] J.N.K. Rao, M. Tausi, Estimating function jackknife variance estimators under stratified multi-stage sampling, Commun. Stat.—Theory Methods 33 (2004) 2087–2095.
[15] J.N.K. Rao, C.F.J. Wu, K. Yue, Some recent work on resampling methods for complex surveys, Surv. Methodol. 18 (1992) 209–217.
[16] P.L.N. Silva, C. Skinner, Variable selection for regression estimation in finite populations, Surv. Methodol. 23 (1997) 23–32.
[17] B.C. Sutradhar, M. Kovacevic, Analysing ordinal longitudinal survey data: generalised estimating equations approach, Biometrika 87 (2000) 837–848.
[18] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[19] R. Tibshirani, The LASSO method for variable selection in Cox model, Stat. Med. 16 (1997) 385–395.
[20] L. Wang, GEE analysis of clustered binary data with diverging number of covariates, Ann. Statist. 39 (2011) 389–417.
[21] L. Wang, S. Wang, Nonparametric additive model-assisted estimation for survey data, J. Multivariate Anal. 102 (2011) 1126–1140.
[22] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.