Analysing Correlated Data from Survey with Complex Design

by

Wei Qian

A thesis submitted to

the Faculty of Graduate Studies and Postdoctoral

Affairs in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Mathematics

School of Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

Abstract

Correlated data collected from probability based sample surveys are often used in research studies in economics, health and social sciences. These surveys usually involve complex design such as stratification, clustering and unequal selection probability. Ignoring the correlations or the sampling design features may lead to erroneous inferences. In this thesis, we consider regression analysis of correlated survey data, taking account of both the correlation and sampling design.

In the non-survey context, marginal models and mixed effects models are two approaches commonly used for correlated data. The Generalized Estimating Equation (GEE) method is the main method for marginal models and likelihood based methods are often used for mixed effects models. Recent progresses have been made to both approaches. Qu, Lindsay and Li (2000) proposed a quadratic inference functions (QIF) approach for marginal models that improves the GEE in terms of efficiency under misspecification of the second moment, and also possesses other features that the GEE does not. Lindsay (1988) proposed composite likelihood (CL) approach for multi-level mixed effects models. The CL method has been developed to reduce high dimensional likelihood functions to low dimensional ones, which makes the computation simpler while still having many of the good inference properties of a full

likelihood function. In this thesis, we extend these methods to survey data with complex design.

A weighting technique is often used to account for the sampling design. Carrillo, Chen and Wu (2010) developed weighted GEE for longitudinal survey data. Following their work, we propose a weighted QIF method to improve the weighted GEE in parallel with the improvement in a non-survey context. We demonstrate that weighting should be used to remove the estimation bias due to informative sampling and that the weighted QIF is better than the weighted GEE when the working correlation is far from the true structure, and as good as weighted GEE when it is close. In addition, we derive its asymptotic properties related to regression parameter estimation and hypothesis testing, and also study the problem of variable selection under the QIF method. We demonstrate the importance of weighting and illustrate the similarity in performance of penalized QIF and GEE, with the former providing improved efficiency in some cases. For the CL method, Rao, Verret and Hidiroglou (2013) proposed a survey weighted pairwise CL approach for two-level survey data. Yi, Rao and Li (2016) further studied its properties for point estimation. We investigate its properties for analytical inference in this thesis, in particular composite likelihood ratio statistics. We show that the method possesses good coverage properties and significantly outperforms other approaches such as REML.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. J.N.K. Rao for his continuous support, patience, motivation, enthusiasm, and immense knowledge. It has been an honor to be his Ph.D. student. He provided the interesting research topics and guided me in all the time of research and writing of this thesis. Also, I want to thank my co-supervisor Prof. Patrick Farrell. He admitted me into this program, guided me at the the early stage of my Ph.D. study and helped me in the writing of this thesis. I wish to thank the members of my defence committee for their questions and comments.

I would like to acknowledge the support from Statistics Canada (STC) who paid my tuition fees in the first two years of my Ph.D. program and allowed me to attend the courses during working hours. I am especially grateful to my first supervisor in STC, Abdellatif Demnati, who encouraged me to pursue a Ph.D. and introduced me to Prof. Rao.

Besides my supervisors, my sincere thanks also go to Prof. Laura Dumitrescu from Victoria University of Wellington (New Zealand) and Prof. Song Cai from Carleton University for their time and efforts in reading this thesis and providing insightful comments.

I am grateful to Prof. Scott Leatherdale at the University of Waterloo for offering me a position in his research team and encouraging me to complete my degree.

Last but not the least, I would like to thank my family for all their support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Probability-sample surveys have been extensively used in economics, health and social sciences. Correlated data may arise from these studies due to repeated observation or clustering. For example, the National Longitudinal Survey of Children and Youth (NLSCY) collects information about factors influencing a child's social, emotional and behavioural development, and to monitor the impact of these factors on the child's development over time. Measurements of a child that are taken repeatedly are often correlated. The Survey of Household Spending (SHS) collects detailed information on household expenditures. The first stage sample is a sample of geographic areas and then a sample of dwellings within the selected areas is selected. The expenditure of the households within the same area can be correlated. In this thesis, we consider the use of correlated survey data for regression analysis.

Regression models are widely used to study the relationship between a response variable and one or more explanatory variables. The simplest are linear models (LM) for a continuous response variable. Generalized linear models (GLM) are a generalization of LM that allows for analyzing different types of response variables

such as continuous, binary, ordinal, and count data. The GLM assume independence between measurements. Therefore, a straightforward application of the GLM to correlated data is inappropriate due to the correlation between measurements. The GLM has been extended in many ways to handle correlated responses. Among them, marginal models and mixed effects models are two commonly used approaches for correlated data. Marginal models are mainly used to make inference on population means. The regression coefficients describe the effects of covariates on population mean response. In contrast, mixed effects models are based on individual responses, which allows for making inferences about individual, and the regression parameters describe the relationship between an individual response and covariates. The choice of methods often depends on subject matter. Mixed effects models are more flexible to accommodate the case of multi-level association, non response and measurement errors, but is computationally more difficult. When population-averaged effects are of interest, the interpretation of the coefficients in mixed effects models may need extra efforts. Likelihood based methods are often used for mixed effects models. In such a situation, a fully specified model is needed. On the other hand, the Generalized Estimating Equation (GEE) method (Liang and Zeger 1986) is the main method for marginal models, which only requires specifications of the first- (mean structure) and second- (covariance) moments of the response without any distributional assumption. Recent progresses have been made for both marginal models and mixed effects models approaches. Qu, Lindsay and Li (2000) proposed a quadratic inference functions (QIF) approach for marginal models. The QIF method has attracted more and more attention. It may improve the GEE in efficiency under the misspecification of the second moment, and is also possesses other features that the GEE does not. Lindsay

(1988) proposed composite likelihood (CL) approach for mixed effects models. The CL method has been developed to reduce high dimensional likelihood functions to low dimensional ones, which makes the computation simpler while still having many of the good inference properties of a full likelihood function. Also, the CL does not need full likelihood specification; as a result, it is more robust to model misspecification. The QIF and CL methods have been proposed and studied in non-survey contexts, but very few work has been done for survey data. In this thesis, we consider their extension to survey data.

Design-based inference and model-based inference are two approaches commonly used for inference from probability sample survey data. The design-based approach considers fixed values associated with the population units and finite population parameters, such as mean, total, or ratio, are of interest. See Cochran (1963), Godambe (1955, 1965, 1975). On the other hand, the model-based approach considers response values as the realization of a random variable under a super-population or model distribution. The objective is to infer about model parameter. See Royall (1971), Royall and Eberhardt (1975), Royall and Cumberland (1977). Model-assisted inference is a special case of the design-based approach for which the parameter of interest is motivated by an assumed model (Särndal, Swensson and Wretman 1992; Binder and Roberts 2003). In this thesis, we are interested in the joint model and design-based inference through a two-phase framework. Under this framework, a finite population is considered as a random sample from the model (super-population) and the survey sample is viewed as a sub-sample of the finite population. Therefore, assuming there is no non-response, model-based inference for data from probability sample surveys is subject to two sources of randomization: a random sample from the assumed model

(or super-population) and selection of the survey sample. If the finite population estimator is a consistent estimator of the model parameter and the survey sample estimator is consistent under sampling design, then the survey sample estimator is a consistent estimator of the model parameter under the joint framework. The validity of model-based inference relies on the correct model specification. The two-phase random framework provides some protection against the model mis-specification in the sense that the survey sample can still be used to make inference on the finite population parameter. There are many contributions on this topic, such as Hartley and Sielken (1975), Fuller (1975), Isaki and Fuller (1982), Godambe and Thompson (2009), Korn and Graubard (1999), Pfeffermann and Sverchkov (2003), Binder and Roberts (2009), Rubin-Bleuer and Schiopu-Kratina (2005).

Sampling surveys often involve complex design such as stratification, clustering and unequal selection probability. If the sampling design is informative(the sample selection probability depends on the response variable after conditioning on the covariates in the model), ignoring the feature of sampling design may lead to biased results. There are different ways to account for the feature of sampling design in survey data analysis. One is to add sample design variables or sampling weights as regression covariates. This method may be impractical either when the sampling design variables are not provided with the data due to confidentiality and disclosure control, or when the survey weight may not be enough precise to reflect the sampling design. Also, adding design variables into the regression models may complicate the model and change the research interest. For these reasons, an alternative method, weighting, is preferable. The pseudo-likelihood method (Binder 1983; Molina and Skinner 1992) and weighted estimating functions (Godambe and Thompson 1986)

method are two common practices of using survey weights in survey data analysis. Our research also follows these two directions when extending to complex survey data. Pfeffermann (1993) provided an excellent review of using sampling weights when modelling survey data.

The estimating equations approach is also commonly used for estimation in classical statistics, and does not require full distribution specification. Most model parameters can be expressed as a solution of a set of census estimating equations. Godambe and Thompson (1986) extended this approach to survey data by introducing a design-unbiased estimator; for example, an estimating function counterpart to the Narin-Horvitz-Thompson (NHT) estimator (Narain 1951; Horvitz and Thompson 1952) of estimating functions. Solving the survey weighted estimating equations leads to design consistent estimators of the census or finite population parameters which in turn estimate the associated model parameters. In this spirit, Carrillo, Chen and Wu (2010) developed weighted GEE for longitudinal survey data. Following their work, we would like to propose a weighted QIF method to improve the weighted GEE in parallel with the improvement in non-survey context.

The pseudo-likelihood approach is an extension of the conventional likelihood method to survey sample by incorporating survey weights. In regular cases where measurements between sample units are independent, the extension is straightforward. Defining a pseudo-likelihood function and taking the first derivative of the log pseudo-likelihood with respect to the parameters of interest, the survey sample estimator is obtained by solving estimating equations based on a weighted score function. However, in the case of complex designs such as multi-stage sampling, the observations may not be independent and multi-level models are often employed. The full

likelihood approach has computational challenges and the extension to the survey sample context requires high-order integration that is usually approximated in practice. The approximation can lead to biased estimates. On the other hand, the CL approach that uses only the pairwise likelihood and simplifies the computation. Rao, Verret and Hidiroglou (2013) proposed a survey weighted pairwise CL approach for two-level survey data. Yi, Rao and Li (2016) further studied its properties in point estimation. In our work, we would like to study its properties in analytical inference.

This thesis is organized as follows. In Chapter 2, we review the marginal models for longitudinal data and introduce the QIF method for non-survey data. In Chapters 3 and 4, we develop survey weighted QIF for longitudinal survey data and study its properties with regard to point estimation and inference. In Chapter 5, simulation studies are conducted to evaluate the performance of the proposed method. In Chapter 6, we further consider variable selection for high-dimensional longitudinal data and propose penalized QIF for survey data. In Chapter 7, we study the analytical properties of survey weighted pairwise CL proposed by Rao, Verret and Hidiroglou (2013). Finally, in Chapter 8, we summarize our results and provide suggestions for future research.

Chapter 2

Quadratic inference function for longitudinal data

2.1 Introduction

Longitudinal studies, also called panel studies in economics and sociology, are very popular in practice, in both observational and experimental studies. The defining feature of a longitudinal study is that repeated observations of the same variables are made for the same individuals at different points in time. Contrary to a longitudinal study, a cross-sectional (or one time) study is limited to a single point in time. In a longitudinal study, the repeated measurements on the same individual capture the within-individual changes, which allows one to study changes in the response over time. The main objective of a longitudinal study is to characterize changes over time and factors that influence changes. Since repeated measurements are taken on the same set of individuals, data from a longitudinal study are clustered and the observations on each individual are correlated (positively in most cases). The

within-individual correlation must be accounted for in the analysis.

Generalized linear models (GLM) introduced by Nelder and Wedderburn (1972) are widely used for regression analysis of independent observations of a discrete or continuous univariate response. For longitudinal data, direct application of the GLM may not be appropriate due to lack of independence. There are three common ways to extend the GLM to longitudinal data, differing on the approaches of incorporating the correlation; namely, the mixed effects models approach, marginal models approach and transitional models approach. Each modelling approach serves specific analytic purposes and answers relevant scientific questions. In mixed effects models, random effects are introduced for each cluster/individual to account for the correlation in clusters/individuals. A mixed effects model allows subject-specific inference, in addition to standard population-average inference. The specification of distributions are needed for the mixed effects models and the computation usually is very challenging. Marginal models describe the mean structure of the response. A marginal model approach requires only specifications of the first- (mean structure) and second- (covariance structure) moments of the response without assuming a particular form of the distribution. The marginal model only allows for population-average inference, not for subject-specific inference. The transitional models approach assumes a Markov structure for the longitudinal process and are used for multi-state data. In this thesis, we will focus on the marginal models approach.

Liang and Zeger (1986) proposed generalized estimating equations (GEE) for the marginal models approach. This method is closely related to the quasi-likelihood method by Wedderburn (1974) and McCullagh (1983). Using a working correlation matrix with fewer nuisance parameters, the GEE method avoids the estimation of

many nuisance parameters, and the risk of numerical error in the inversion of covariance matrix estimate. Under the weak assumption that a weighted average of the estimated correlation matrix converges to a fixed matrix, the GEE estimators of the regression parameters are consistent and asymptotically normally distributed, even if the working correlation is mis-specified. However, the regression parameter estimators are inefficient under the mis-specified working correlation. Moreover, in some worse cases where the nuisance correlation parameters are not consistently estimated, the GEE fails to produce consistent estimators for the regression parameter (Crowder 1995).

Qu, Lindsay and Li (2000) introduced the method of quadratic inference functions (QIF) to longitudinal data that also produces consistent regression parameter estimates. The QIF method does not involve direct estimation of the correlation parameters and remains optimal within the assumed family even if the working correlation structure is mis-specified. Thus, under a correctly modelling of the correlation structure, the QIF regression parameter estimator is as efficient as the GEE one; on the other hand, when the correlation structure is mis-specified, the QIF estimator is more efficient. The QIF method has other advantages over the GEE in terms of validation of assumptions, model selection and robustness against outliers or contamined data, see Song *et al.*(2009). Our research aims to apply the QIF method to longitudinal data collected from sample surveys.

2.2 General set-up for marginal models

Suppose that Y is a response variable (continuous, binary, ordinal, or a count) and that N subjects are measured repeatedly over time. Let Y_{it} denote the response variable for the i^{th} subject at the t^{th} measurement occasion. Assume that there are T_i repeated measurements of the response on the i^{th} subject and that each response Y_{it} is observed at time t. T_i may differ over subjects, which accommodates unbalanced data. The response variables for the i^{th} subject can be grouped into a vector of dimension T_i, such that, $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iT_i})$ for $i = 1, ..., N$. Response vectors for two different individuals are independent, i.e. $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_l$ are independent for $i \neq l$. Associated with each response, $\boldsymbol{Y}_i$, there is a $p \times 1$ vector of covariates $\boldsymbol{X}_{it}^T = (X_{it1}, ..., X_{itp})$ where $i = 1, ..., N$, $t = 1, ..., T_i$, and the superscript " ^{T}" denotes matrix (or vector) transpose. For longitudinal data, $\boldsymbol{X}_{it}$ includes two types of covariates: time-invariant covariates and time-varying covariates. The values of time-invariant covariates do not change throughout the duration of the study and those of time-varying covariates change over time. We assume $\boldsymbol{Y}_i$ are fully observed and covariates are always observed.

The marginal model approach has the following specifications.

1. $g(\mu_{it}) = \boldsymbol{X}_{it}^T\boldsymbol{\beta}$ where $g(\mu_{it})$ is a known monotone link function of μ_{it} such that $\mu_{it}(\boldsymbol{\beta}) = E[Y_{it}|\boldsymbol{X}_i]$ and $\boldsymbol{\beta}$ is an unknown p-dimensional regression vector; Inferences about the unknown regression parameters $\boldsymbol{\beta}$ are of interest.

2. $Var(Y_{it}|\boldsymbol{X}_i) = \phi\upsilon(\mu_{it})$ where $\upsilon(\mu_{it})$ is a known function of the mean, μ_{it}, and ϕ is a dispersion parameter that may be known, or may need to be estimated;

3. The pairwise within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of the means μ_{it}, the scale parameter ϕ, and an additional set of within-subject association parameters, $\boldsymbol{\alpha}$.

The first two assumptions correspond to standard generalized linear models and specify the marginal mean and variance form of Y_{it} respectively. Different from GLM, marginal models do not require full specification of the distribution for the observations, only a regression model for the mean response. The third assumption specifies the association among the repeated observations from the same subject. More details about marginal model specification can be found in Fitzmaurice, Laird and Ware (2011).

2.3 Generalized estimating equations (GEE)

When the response variable is discrete, there is no convenient specification of the joint multivariate distribution of the response vector $\boldsymbol{Y}_i$ for marginal models. Maximum likelihood estimation does not apply here. Based on the concept of "estimating equations", Liang and Zeger (1986) developed a GEE approach for the estimation of regression parameters in marginal models that requires the correct specification of the mean and only a working covariance structure of the response vector. The GEE provides a unified approach for analyzing correlated responses that can be discrete or continuous.

Let $\boldsymbol{R}_i(\boldsymbol{\alpha})$ be a $T_i \times T_i$ correlation matrix where $\boldsymbol{\alpha}$ is a nuisance parameter vector which fully characterize $\boldsymbol{R}_i(\boldsymbol{\alpha})$, for $i = 1, ..., N$. $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is called "working" correlation

matrix. Four commonly used correlation structures were studied: independent (IN), exchangeable (EX), autoregressive (AR) and unstructured (UN). The GEE estimator, $\hat{\boldsymbol{\beta}}^{GEE}$, for the regression parameter $\boldsymbol{\beta}$ is defined as the solution of estimation equations

$$\boldsymbol{S}_N(\boldsymbol{\beta}) \equiv \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0, \tag{1}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is a $T_i \times p$ matrix and $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$ with $\boldsymbol{A}_i = \text{diag}\{\phi\upsilon(\mu_{it}),\ t = 1, ..., T_i\}$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), ..., \mu_{iT_i}(\boldsymbol{\beta}))$. Generally, the GEE has no closed form solution, instead, the equations are solved using an iterative algorithm. Since the GEE depends on $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ, an iterative two-stage estimation procedure is used:

1. given an initial estimate of $\boldsymbol{\beta}$ that often obtained from a generalized linear model, update the estimates of $\boldsymbol{\alpha}$ and ϕ.

2. with the estimates of $\boldsymbol{\alpha}$ and ϕ, one may estimate $\boldsymbol{V}_i$ and update the estimate of $\boldsymbol{\beta}$ by solving equation (1).

To estimate ϕ and $\boldsymbol{\alpha}$, one calculates the standardized residuals

$$e_{it} = (Y_{it} - \hat{\mu}_{it})/\sqrt{\upsilon(\hat{\mu}_{it})},$$

where $\hat{\mu}_{it} = \mu_{it}(\hat{\boldsymbol{\beta}})$. Then, ϕ can be estimated by

$$\hat{\phi} = \left(\sum_{i=1}^{N} T_i - p\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T_i} e_{it}^2.$$

The estimation of correlation parameters $\boldsymbol{\alpha}$ depends on the working correlation structure, for example, we only need to estimate one correlation parameter

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} \sum_{t>t'} e_{it} e_{it'}}{(0.5 \sum_{i=1}^{N} T_i(T_i - 1) - p)\hat{\phi}}$$

for EX correlation structure,

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} \sum_{t \leq T_i - 1} e_{it} e_{i,t+1}}{(\sum_{i=1}^{N} (T_i - 1) - p)\hat{\phi}}$$

for AR1 structure. However, for an unstructured correlation structure, we need estimate all elements, such that

$$\hat{\alpha}_{tt'} = \frac{\sum_{i=1}^{N} e_{it} e_{it'}}{(\sum_{i=1}^{N} T_i - p)\hat{\phi}}$$

if $t \neq t'$, and 1 otherwise.

Under mild regularity conditions and provided that $\boldsymbol{\alpha}$ and ϕ are properly estimated, the GEE has the following large sample properties.

1. $\hat{\boldsymbol{\beta}}^{GEE}$ is a consistent estimator of $\boldsymbol{\beta}$. The consistency requires that the response mean model is correctly specified and the nuisance parameters, $\boldsymbol{\alpha}$ and ϕ, are properly estimated. The consistency holds regardless of the correct specification of $\boldsymbol{R}_i$; however, $\hat{\boldsymbol{\beta}}^{GEE}$ is efficient only if $\boldsymbol{R}_i$ is the true correlation matrix.

2. In large samples, given $\sqrt{N}-$consistent estimators of $\boldsymbol{\alpha}$ and ϕ, $\hat{\boldsymbol{\beta}}^{GEE}$ asymptotically follows a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance

matrix $\boldsymbol{V}_G$ where $\boldsymbol{V}_G = \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$, with

$$\boldsymbol{B} = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i,$$

$$\boldsymbol{M} = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{COV}(\boldsymbol{Y}_i|\boldsymbol{X}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i.$$

If the "working" correlation structure is correct, $\boldsymbol{V}_G = (\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i)^{-1}$, the GEE estimator has the Crowder optimality (Song, 2007) and performs similarly to the MLE estimator under a correct correlation specification. Yuan and Jennrich (1998) and Shao (2003) discus general conditions that lead to the consistency and normality of $\hat{\boldsymbol{\beta}}^{GEE}$ for large samples.

Rotnitzky and Jewell (1990) constructed generalized score test statistics for the regression parameter $\boldsymbol{\beta}$ in marginal models. Under a correct specification of the working correlation, the test statistic follows an asymptotic chi-squared distribution. In the case where the correlation is mis-specified, the test statistic is not χ^2 but a linear combination of independent $\chi^2(1)$, whose limiting distribution can be approximated by χ^2, following the corrections suggested by Rao and Scott (1981, 1984, 1987).

The GEE estimator is inefficient when the working correlation matrix is mis-specified and it also depends on the nuisance parameters $\boldsymbol{\alpha}$. To overcome these drawbacks, Qu, Lindsay and Li (2000) proposed an alternative approach to estimate coefficients in longitudinal data via quadratic inference functions (QIF). This method of estimation approach does not require any more assumptions than the GEE.

2.4 Quadratic inference functions (QIF)

We assume that $\boldsymbol{R}_i = \boldsymbol{R}$, i.e. correlation is the same for all individuals. The QIF approximates the inverse of the correlation matrix by a linear combination of several basis matrices

$$\boldsymbol{R}^{-1} \approx \sum_{i=0}^{m} a_i \boldsymbol{M}_i, \tag{2}$$

where $\boldsymbol{M}_0, ..., \boldsymbol{M}_m$ are known matrices with components only 0 or 1, $\boldsymbol{M}_0$ is the identity matrix, and $a_0, ..., a_m$ are unknown constants depending on correlation structure, $\boldsymbol{\alpha}$ and cluster size. The decomposition is exact for some common working correlation matrices. For the independent correlation matrix, $\boldsymbol{R}^{-1}$ can be written $a_0\boldsymbol{M}_0$. For the exchangeable correlation matrix, $\boldsymbol{R}^{-1}$ can be written $a_0\boldsymbol{M}_0 + a_1\boldsymbol{M}_1$ where $\boldsymbol{M}_1$ is a matrix with 0 on the diagonal and 1 off the diagonal. For the first-order autoregressive correlation matrix (AR1). $\boldsymbol{R}^{-1}$ can be written as $a_0\boldsymbol{M}_0 + a_1\boldsymbol{M}_1 + a_2\boldsymbol{M}_2$ where $\boldsymbol{M}_1$ has 1 on the two main off-diagonal and 0 elsewhere, and $\boldsymbol{M}_2$ has 1 on the diagonal corners. In cases when it is difficult to determine a suitable working correlation structure, Qu and Li (2006) suggest a hybrid working correlation that combines basis matrices from several working correlations. For the unstructured correlation structure, Qu and Lindsay (2003) proposed the conjugate gradient method to construct the basis matrices. Note that, for an independent correlation structure, the QIF method is the same as GEE method.

With $\boldsymbol{R}^{-1}$ in (1) being replaced by (2), the GEE estimating function becomes

$$\boldsymbol{S}_N(\boldsymbol{\beta}) \approx \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} (a_0\boldsymbol{M}_0 + \cdots + a_m\boldsymbol{M}_m) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})),$$

which is a linear combination of the elements of the extended score $\boldsymbol{g}_N(\boldsymbol{\beta})$ i.e.

$$\boldsymbol{g}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{\beta}) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \\ \sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \\ \vdots \\ \sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_m\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \end{pmatrix}$$

with coefficients $\boldsymbol{a} = (a_0, a_1, \cdots, a_m)^T$.

Qu, Lindsay and Li (2000) proposed a max-info GEE method to estimate regression parameters. Max-info GEE differs from the classical GEE in estimating the correlation parameters $\boldsymbol{a}$, instead of $\boldsymbol{\alpha}$. The optimal $\boldsymbol{a}$ is found by maximizing the trace of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}(\boldsymbol{a})$.

Qu, Lindsay and Li (2000) also proposed the QIF method to estimate regression parameters. Assume that the elements in the extended score $\boldsymbol{g}_N(\boldsymbol{\beta})$ are linearly independent. It is easy to show $E_{\boldsymbol{\beta}}[\boldsymbol{g}_N(\boldsymbol{\beta})] = 0$ by definition of $\boldsymbol{g}_N(\boldsymbol{\beta})$. Since the dimension of $\boldsymbol{g}_N(\boldsymbol{\beta})$ is $(m+1)p$ and greater than the dimension of unknown parameter $\boldsymbol{\beta}$, there is no way to solve the estimating equations such that $\boldsymbol{g}_N(\boldsymbol{\beta}) = 0$ to find the estimator. The generalized method of moments (GMM) can be applied here, see Hansen (1982). The idea of the GMM is to find an element in parameter space that sets the linear combination of $\boldsymbol{g}_N(\boldsymbol{\beta})$ as close to zero as possible. The GMM estimators are the minimizers of a class of weighted lengths of $\boldsymbol{g}_N(\boldsymbol{\beta})$ indexed by

weight matrices $\boldsymbol{W}$, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\Omega} \boldsymbol{g}_N^T(\boldsymbol{\beta})\boldsymbol{W}^{-1}\boldsymbol{g}_N(\boldsymbol{\beta}).$$

Remarks:

1. To achieve consistency of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{W}$ is assumed to be a non-singular non-stochastic positive-definite matrix or replaced by $\hat{\boldsymbol{W}}$ - a strongly consistent estimator of $\boldsymbol{W}$; and

2. if $\boldsymbol{W}$ is the covariance matrix of $\boldsymbol{g}_N(\boldsymbol{\beta})$, then $\hat{\boldsymbol{\beta}}$ is optimal and has minimum (in the sense of Löwner ordering, see Page 207 in Zhang (1999)) asymptotic variance matrix.

Following the GMM idea, Qu, Lindsay and Li (2000) defined quadratic inference functions

$$Q_N(\boldsymbol{\beta}) = N\boldsymbol{g}_N^T(\boldsymbol{\beta})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\boldsymbol{g}_N(\boldsymbol{\beta}) \tag{3}$$

where $C_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^N \boldsymbol{g}_i(\boldsymbol{\beta})\boldsymbol{g}_i^T(\boldsymbol{\beta})$ is the sample covariance matrix. A QIF estimator, $\hat{\boldsymbol{\beta}}_N$, of $\boldsymbol{\beta}$ is given as

$$\hat{\boldsymbol{\beta}}_N = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_N(\boldsymbol{\beta}). \tag{4}$$

The corresponding estimating equations are

$$N^{-1}\frac{\partial Q_N(\boldsymbol{\beta})}{\partial\beta_k} = 2\frac{\partial \boldsymbol{g}_N^T(\boldsymbol{\beta})}{\partial\beta_k})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\boldsymbol{g}_N(\boldsymbol{\beta}) - \boldsymbol{g}_N^T(\boldsymbol{\beta})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\frac{\partial)\boldsymbol{C}_N(\boldsymbol{\beta})}{\partial\beta_k})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\boldsymbol{g}_N(\boldsymbol{\beta}) = 0, \tag{5}$$

for $k = 1, ..., p$. Let $\dot{\boldsymbol{Q}}_N(\boldsymbol{\beta}) = (\partial Q_N(\boldsymbol{\beta})/\partial\beta_1, \cdots, \partial Q_N(\boldsymbol{\beta})/\partial\beta_p)^T$. The solution to (5) can be computed from the Newton-Raphson iterative algorithm:

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} + \ddot{\boldsymbol{Q}}_N(\boldsymbol{\beta}^{(l)})\dot{\boldsymbol{Q}}_N(\boldsymbol{\beta}^{(l)}),$$

where $\ddot{\boldsymbol{Q}}_N(\boldsymbol{\beta})$ is a $p \times p$ matrix $\ddot{\boldsymbol{Q}}_N(\boldsymbol{\beta}) = \frac{\partial^2 Q_N(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$.

As we can see from above, Q_N does not depend on $\boldsymbol{a} = (a_0, a_1, \cdots, a_m)^T$, so that there is no need to estimate the coefficient vector $\boldsymbol{a}$. That is, the QIF method does not rely on whether an appropriate estimate of the correlation parameter is available or not.

Large sample properties of the QIF method include (Hansen 1982 and Qu 1998)

1. $\hat{\boldsymbol{\beta}}_N$ is a consistent estimator of $\boldsymbol{\beta}$;

2. Let $\boldsymbol{\beta}_0$ be the true parameter value, as $N \to \infty$, $\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, (\boldsymbol{D}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}_0)^{-1})$ where $\boldsymbol{D}_0 = \lim_{N\to\infty} E[\partial\boldsymbol{g}_N(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}]$ and $\boldsymbol{\Sigma} = \lim_{N\to\infty} \boldsymbol{COV}[\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta})]$.

Under the correct specification of correlation structure, the asymptotic covariance of $\hat{\boldsymbol{\beta}}_N$ reaches the minimum, in the sense of Löwner ordering. This property ensures that QIF is more efficient than GEE when the working correlation is mis-specified and as efficient as GEE when the working correlation is correct. That is, QIF is more robust to correlation mis-specification than the GEE.

For hypothesis testing, QIF has similar property to the log likelihood function and can be used as an inference function. For example, to test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, one can use the test statistic $Q_N(\boldsymbol{\beta}_0) - Q_N(\hat{\boldsymbol{\beta}}_N)$ which asymptotically follows $\chi^2(p)$ where p

is the dimension of regression parameter vector $\boldsymbol{\beta}$. QIF also can be used to test the model mean assumption (goodness of fit): $H_0 : E[\boldsymbol{g}_N(\boldsymbol{\beta})] = 0$. The good of fit test statistic $Q_N(\hat{\boldsymbol{\beta}}_N)$ follows a $\chi^2(mp)$ distribution where mp is the difference between the dimension of the extended score and the dimension of the parameter vector, see Qu, Lindsay and Li (2000). In contrast, the mean assumption of GEE cannot be verified because it does not have an objective function, unlike the QIF.

Other advantages of the QIF method have been discussed. In the literature, Qu and Song (2004) compared the robustness of the QIF approach with the GEE method when outlying observations are present. Their simulation study shows that QIF is substantially more robust than GEE. This is because the QIF influence function is bounded while the influence function of GEE is not bounded. Hu and Song (2012) considered sample size determination and power calculations for the QIF method based on the Wald test in a marginal logistic model. Han and Song (2011) considered a linear shrinkage estimator for the QIF method in the case where the weighting matrix is not invertible. Qu and Li (2006) proposed an efficient estimation procedure for varying-coefficient models for longitudinal data based on QIF. Wang and Qu (2009) and Bai, Fung and Zhu (2009) proposed an estimating procedure for single-index models based on the combination of penalized splines and QIF. Wang, Wang and Song (2012) used QIF to propose an estimation procedure that enables the analysis of merged data and accounts for different within-subject correlations and follow-up schedules in different studies. On the other hand, Westgate and Braun (2012) studied the performance of the QIF estimate, showing that QIF could produce estimates with better precision than GEE even under a mis-specified covariance structure in small to moderately sized samples, and for unequal cluster sizes and

models with only cluster-level covariates.

Recently, QIF has been attracting more and more attention because of its advantages over GEE in practice. Odueyungbo *et al.* (2008) compared the GEE and QIF methods using data from the National Longitudinal Survey of Children and Youth (NLSCY) survey. Asgari *et al.* (2013) used the QIF method to determine the factors associated with obesity in Iran.

2.5 A simulation study for continuous responses

We conduct a simulation study to compare the QIF method with the GEE method; in the meantime, we propose another version of QIF, i.e.

$$Q_N^*(\boldsymbol{\beta}) = N\boldsymbol{g}_N^T(\boldsymbol{\beta})\boldsymbol{C}_N^{-1}(\hat{\boldsymbol{\beta}}^*)\boldsymbol{g}_N(\boldsymbol{\beta}), \tag{6}$$

where $\boldsymbol{C}_N^{-1}(\hat{\boldsymbol{\beta}}^*) = N^{-1}\sum_{i=1}^N \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}^*)\boldsymbol{g}_i^T(\hat{\boldsymbol{\beta}}^*)$ and $\hat{\boldsymbol{\beta}}^*$ is a $\sqrt{N}-$consistent estimator of $\boldsymbol{\beta}$. We take $\hat{\boldsymbol{\beta}}^*$ to be the GEE estimator under working independent correlation. Let $\hat{\boldsymbol{\beta}}_N^*$ be the corresponding QIF estimator minimizing (6). It can be shown that $\boldsymbol{C}_N(\hat{\boldsymbol{\beta}}^*)$ converges to Σ almost surely, so that $\hat{\boldsymbol{\beta}}_N^*$ is as efficient as $\hat{\boldsymbol{\beta}}_N$ asymptotically. The advantage of using (6) is to avoid the differentiation of $\boldsymbol{C}_N$.

We generate continuous responses from the following model

$$y_{it} = x_{it}^{(1)}\beta_1 + x_{it}^{(2)}\beta_2 + \epsilon_{it}, \text{ for } i = 1, 2, \cdots, N, \text{ and } t = 1, 2, \cdots, T, \tag{7}$$

where $\boldsymbol{x}_i^{(1)} = (x_{i1}^{(1)}, x_{i2}^{(1)}, \cdots, x_{iT}^{(1)})^T$ where $x_{it}^{(1)} = t/T$, $\boldsymbol{x}_i^{(2)} = (x_{i1}^{(2)}, x_{i2}^{(2)}, \cdots, x_{iT}^{(2)})^T$ are generated from a multivariate normal distribution with mean $(\frac{1}{T}, \frac{2}{T}, \cdots, 1)$ and

covariance matrix I, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (1, 1)^T$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{iT})^T$ is generated from a T-dimensional normal distribution with mean 0, marginal variance 1 and different correlation structures (true correlation). We assume that the covariates are fixed.

We consider the number of individuals $N = 10, 20, 40, 80$, exchangeable (EX) and first-order autoregressive correlation structures (AR1), and correlation parameters $\alpha = 0.3, 0.7$. We simulate 1,000 independent samples from the above model, apply both the GEE and QIF methods to each sample and compute, for all methods, the mean squared error (m.s.e.) of estimators by averaging $(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2$ over all samples.

We also propose a modified GEE by replacing $\boldsymbol{R}$ by fixed $\boldsymbol{R}^*$, and modified QIF estimators by replacing $\boldsymbol{C}_N$ by fixed $\boldsymbol{C}_N^*$, where $\boldsymbol{R}^*$ and $\boldsymbol{C}_N^*$ are obtained under working independence. That is, we obtain the estimate of $\boldsymbol{\beta}$ under working independence and then calculate the estimate of $\boldsymbol{\alpha}^*$ and ϕ^* (and $\boldsymbol{R}^*$) for the GEE and $\boldsymbol{C}_N^*$ for the QIF. In the later iteration to solve estimating equations for $\boldsymbol{\beta}$, we do not update $\boldsymbol{R}^*$ and $\boldsymbol{C}_N^*$ any more. We consider modified GEE with unstructured correlation (we call fixed) and modified QIF for both exchangeable and AR1 correlation. The performance of these estimators is evaluated by simulated relative efficiency (SRE) defined as

$$\text{SRE} = \frac{\text{m.s.e. of the GEE under true correlation}}{\text{m.s.e. of the QIF (or GEE) under working correlation}}.$$

Note that SRE ≤ 1 since the GEE under true working correlation produces regression parameter estimate with the smallest MSE. A method is more efficient if the resulting

SRE is closer to 1.

We present the simulation results in Table 1-5 for $N = 20, 40, 80$ and $T = 10, 20$. We observe the following.

1. The QIF method is more efficient than the GEE method under the same misspecified correlation, and as good as the GEE when the working correlation is true. For example, when the true correlation is EX(0.7), the QIF estimator under EX working correlation has the SRE 0.99, which is as efficient as the GEE estimator under EX, while the QIF estimator under AR1 has the SRE 0.63, which is twice more efficient than the GEE estimator under AR1 (0.29).

2. The relative efficiency varies over true correlation, working correlation, sample size and cluster size.

3. Modified QIF performs as well as QIF under the same working correlation, except one case where $N = 20$ and $T = 20$ and both the true and working correlations are EX with ρ=0.7. In this case, the modified QIF is 16% less efficient than the regular QIF.

4. The modified unstructured (Fixed) GEE estimator performs as well as the unstructured GEE estimator for large and moderate sample sizes (except $N = 40$ and $T = 20$), but better for small sample sizes.

5. For a large sample size ($N = 80$) the unstructured (either fixed or not fixed) GEE estimator performs better than QIF estimators under the misspecified correlation, and as good as them under correct working correlation. For small sample sizes, the unstructured GEE estimators perform worse than the QIF

estimators regardless of the working correlation used. For moderate sample sizes, the unstructured GEE estimators perform better only when the true correlation is EX(0.7) and working correlation is AR1.

6. In some cases such as $N = 20$ and $T = 20$, the unstructured GEE estimator performs very poorly.

Modified versions can simplify the computation. Therefore, we may consider using the modified unstructured GEE estimators for large sample sizes, and the modified QIF estimators for moderate or small sample sizes.

Table 1: SRE after 1,000 simulation, $N = 80, T = 10$. GEE Fixed: GEE with R being replaced by R^* under unstructured working correlation, GEE UN: GEE under unstructured working correlation, GEE EX: GEE under exchangeable working correlation, GEE AR1: GEE under AR1 working correlation, QIF EX: QIF under exchangeable working correlation, QIF AR1: QIF under AR1 working correlation, QIF EX*: QIF under exchangeable working correlation and with C_N being replaced by C_N^*, QIF AR1*: QIF under AR1 working correlation and with C_N being replaced by C_N^*

True		Working Correlation							
		GEE				QIF			
		Fixed	UN	EX	AR1	EX	AR1	EX*	AR1*
EX	0.7	0.86	0.87	1.00	0.29	0.99	0.63	0.99	0.62
	0.3	0.90	0.90	1.00	0.63	0.99	0.77	0.99	0.76
AR1	0.7	0.92	0.90	0.68	1.00	0.91	0.97	0.91	0.98
	0.3	0.93	0.90	0.93	1.00	0.98	0.98	0.98	0.99

Table 2: SRE after 1,000 simulation, N=40, T=10

True		Working Correlation							
		GEE				QIF			
		Fixed	UN	EX	AR1	EX	AR1	EX*	AR1*
EX	0.7	0.71	0.79	1.00	0.29	0.99	0.57	0.96	0.56
	0.3	0.80	0.73	1.00	0.64	0.97	0.77	0.97	0.77
AR1	0.7	0.88	0.80	0.68	1.00	0.92	0.95	0.92	0.96
	0.3	0.87	0.77	0.92	1.00	0.99	0.96	0.99	0.98

Table 3: SRE after 1,000 simulation, N=20, T=10

True		Working Correlation							
		GEE				QIF			
		Fixed	UN	EX	AR1	EX	AR1	EX*	AR1*
EX	0.7	0.46	0.51	1.00	0.30	0.95	0.58	0.88	0.54
	0.3	0.75	0.52	1.00	0.68	0.98	0.73	0.97	0.75
AR1	0.7	0.77	0.45	0.61	1.00	0.88	0.90	0.90	0.94
	0.3	0.85	0.52	0.91	1.00	0.94	0.90	0.95	0.94

Table 4: SRE after 1,000 simulation, N=40, T=20

True		Working Correlation							
		GEE				QIF			
		Fixed	UN	EX	AR1	EX	AR1*	EX*	AR1*
EX	0.7	0.29	0.50	1.00	0.13	0.99	0.34	0.90	0.32
	0.3	0.66	0.49	1.00	0.46	0.99	0.59	1.00	0.59
AR	0.7	0.81	0.57	0.57	1.00	0.91	0.98	0.91	0.99
	0.3	0.84	0.49	0.89	1.00	0.94	0.92	0.95	0.94

Table 5: SRE after 1,000 simulation, N=20, T=20

True		Working Correlation							
		GEE				QIF			
		Fixed	UN	EX	AR1	EX	AR1	EX*	AR1*
EX	0.7	0.12	0.01	1.00	0.13	0.93	0.31	0.78	0.29
	0.3	0.45	0.01	1.00	0.45	0.97	0.54	0.93	0.56
AR1	0.7	0.89	0.01	0.50	1.00	0.86	0.91	0.87	0.95
	0.3	0.94	0.05	0.92	1.00	0.95	0.86	0.96	0.91

Chapter 3

QIF for longitudinal survey data

Longitudinal surveys are an important source of data. For example, they include the Survey of Labour and Income Dynamics (SLID) and National Longitudinal Survey of Children and Youth (NLSCY). These surveys often involve complex sampling designs, and classical methods for non-survey data may be inappropriate. Carrillo, Chen and Wu (2010) developed a survey-weighted GEE approach for analyzing longitudinal survey data. The survey-weighted GEE produces consistent estimates for the regression parameters under a joint model-design framework. Given the superior performance of the QIF method over the GEE method , we would like to extend the QIF approach to longitudinal survey data.

In parallel with Carrillo, Chen and Wu (2010), our work also proceeds under a model-design framework. Assume that there is a sequence of finite populations, indexed by ν. For a given ν, the finite population is considered to be a random sample from a super-population model ξ with population size N_ν. Furthermore, a probability sample of size n_ν is taken from the finite population according to a probability sampling design d. As $\nu \to \infty$, both $N_\nu \to \infty$ and $n_\nu \to \infty$. For

simplicity, we suppress the index ν.

Let $U = (1, ..., N)$ be the index set of units in the finite population and S be the probability sample. Let $(y_{it}, \mathbf{x}_{it})$ be the observed response value and the covariate vector associated with the ith sampled unit at time t. Let w_i be the survey weight of the sampled unit i. Without loss of generality, we may let w_i be the sampling design weight calculated as the inverse of the probability of unit i to be included into the sample. Throughout this thesis, we assume complete responses for both the response variable and the covariates. Missing data problems will be left for future research.

3.1 Pseudo generalized estimating equations (pseudo-GEE) estimator

To estimate the regression parameters in the marginal modelling approach for longitudinal data, based on a survey sample drawn from a finite population, Carrillo, Chen and Wu (2010) proposed the sample-based weighted estimating equations (WEE)

$$\boldsymbol{S}_n(\boldsymbol{\beta}) \equiv \sum_{i \in S} w_i \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0. \tag{8}$$

Note that $\boldsymbol{S}_n(\boldsymbol{\beta})$ is the the Narain-Horvitz-Thompson (NHT, Narain 1951, Horvitz and Thompson 1952) estimator of the census estimating function (EF) $\boldsymbol{S}_N(\boldsymbol{\beta})$ in (1). Let E_ξ denote the expectation under the model ξ, E_d denote the expectation under the design d, and $E_{\xi d}$ denote the expectation under model ξ and design d jointly. Under design d, $\boldsymbol{S}_n(\boldsymbol{\beta})$ is an unbiased estimator of $\boldsymbol{S}_n(\boldsymbol{\beta})$. Therefore, it has zero mean

under model ξ and design d jointly, i.e.

$$E_{\xi d}[\boldsymbol{S}_n(\boldsymbol{\beta})] = E_{\xi}[E_d(\boldsymbol{S}_n(\boldsymbol{\beta}))] = E_{\xi}[\boldsymbol{S}_N(\boldsymbol{\beta})] = 0.$$

One can solve the WEE (8) iteratively in two steps. First, estimate $\boldsymbol{\alpha}$ and ϕ given the current $\hat{\boldsymbol{\beta}}$; and then estimate $\boldsymbol{\beta}$ based on the obtained $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$. We also need to consider the survey design when estimating $\boldsymbol{\alpha}$ and ϕ in the iteration, leading to

$$\hat{\phi} = \sum_{i \in S} w_i \sum_{t=1}^{T_i} e_{it}^2 / (\sum_{i \in S} w_i T_i - p),$$

where $\mu_{it} = g^{-1}(\boldsymbol{x}_{it}^T \boldsymbol{\beta})$ using the current value of $\boldsymbol{\beta}$, $e_{it} = (y_{it} - \mu_{it})/\sqrt{v(\mu_{it})}$ and p is the number of regression parameters, and

$$\hat{\alpha} = \frac{\sum_{i \in S} w_i \sum_{t > t'} e_{it} e_{it'}}{(0.5 \sum_{i \in S} w_i T_i (T_i - 1) - p)\hat{\phi}}$$

for EX correlation structure,

$$\hat{\alpha} = \frac{\sum_{i \in S} w_i \sum_{t \leq T_i - 1} e_{it} e_{i,t+1}}{(\sum_{i \in S} w_i (T_i - 1) - p)\hat{\phi}}$$

for AR1 structure, and

$$\hat{\alpha}_{tt'} = \frac{\sum_{i \in S} w_i e_{it} e_{it'}}{(\sum_{i \in S} w_i T_i - p)\hat{\phi}}$$

if $t \neq t'$, and 1 otherwise for an unstructured correlation structure.

3.2 Pseudo quadratic inference functions (pseudo-QIF) estimator

In this section, we propose a survey-weighted QIF approach to estimate regression parameters $\boldsymbol{\beta}$ in longitudinal survey data. First we define the survey-weighted extended scores based on the sample as

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = N^{-1} \sum_{i \in S} w_i \boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})$$

where w_i are the survey weights calculated as the inverses of the first-order inclusion probabilities. Note that $\boldsymbol{g}_n(\boldsymbol{\beta})$ is the NHT estimator of $\boldsymbol{g}_N(\boldsymbol{\beta})$, which has zero mean under model ξ and design d jointly, i.e., $E_{\xi d}[\boldsymbol{g}_n(\boldsymbol{\beta})] = 0$.

The survey-weighted QIF (or pseudo-QIF) is defined as

$$Q_n(\boldsymbol{\beta}) = n \boldsymbol{g}_n^T(\boldsymbol{\beta}) \boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) \boldsymbol{g}_n(\boldsymbol{\beta}),$$

where $\boldsymbol{A}_n(\boldsymbol{\beta}) = N^{-1} \sum_{i \in S} w_i \boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}) \boldsymbol{g}_i^T(\boldsymbol{Y}_i, \boldsymbol{\beta})$. In the case where $\boldsymbol{A}_n$ is singular, one can replace $\boldsymbol{A}_n^{-1}$ by any generalized inverse; one example is the Moore-Penrose generalized inverse $\boldsymbol{A}_n^{-}$ (Pilla 2005). However, we only consider the non-singular case in our work. In addition, as discussed in Chapter 2, $\boldsymbol{A}_n(\boldsymbol{\beta})$ in $Q_n(\boldsymbol{\beta})$ can be replaced by $\boldsymbol{A}_n^* = N^{-1} \sum_{i \in S} w_i \boldsymbol{g}_i(\boldsymbol{Y}_i, \tilde{\boldsymbol{\beta}}) \boldsymbol{g}_i^T(\boldsymbol{Y}_i, \tilde{\boldsymbol{\beta}})$ where $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ (e.g., $\tilde{\boldsymbol{\beta}}$ can be the weighted-GEE estimators), but the use of $\boldsymbol{A}_n^*$ will not be further studied.

The pseudo-QIF estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ for a survey sample is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n(\boldsymbol{\beta}).$$

Due to the factor $\partial \boldsymbol{A}_n^{-1}(\boldsymbol{\beta})/\partial\beta_k = -\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_k}\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})$ (Petersen and Pedersen 2012), minimizing $Q_n(\boldsymbol{\beta})$ leads to solving p estimating equations

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_k} = 2n\frac{\partial \boldsymbol{g}_n^T(\boldsymbol{\beta})}{\partial\beta_k}\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta}) - n\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_k}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}) = 0, \qquad (9)$$

for $k = 1, ..., p$. In matrix notation, we denote $\dot{g}_n(\boldsymbol{\beta}) = \partial \boldsymbol{g}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}) = \partial Q_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}) = \frac{\partial^2 Q_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$. Usually, a closed form solution is not available to (9), and a Newton-Raphson iterative algorithm can be used to find an approximation of the solution,

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \ddot{\boldsymbol{Q}}_n^{-1}(\boldsymbol{\beta}^{(k)})\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}^{(k)}), \qquad (10)$$

The starting value can be $\boldsymbol{\beta}^{(0)} = \tilde{\boldsymbol{\beta}}$. The second derivative is

$$\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}) = 2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}) + \boldsymbol{R}_n(\boldsymbol{\beta})$$

where $\boldsymbol{R}_n(\boldsymbol{\beta})$ is a $p \times p$ matrix with the $(k,l)^{th}$ element being

$$2n\frac{\partial^2 \boldsymbol{g}_n^T(\boldsymbol{\beta})}{\partial\beta_k\partial\beta_l}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}) - 2n\frac{\partial \boldsymbol{g}_n^T(\boldsymbol{\beta})}{\beta_k}\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_l}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}) + 2n\frac{\partial \boldsymbol{g}_n^T(\boldsymbol{\beta})}{\partial\beta_l}\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_k}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta})$$
$$+2n\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_l}\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_k}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}) - n\boldsymbol{g}_n^T\boldsymbol{A}_n^{-1}\frac{\partial^2 \boldsymbol{A}_n(\boldsymbol{\beta})}{\partial\beta_k\partial\beta_l}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}).$$

The first term in (9) at $\boldsymbol{\beta}$ is $O_p(n^{1/2})$ and the second term at $\boldsymbol{\beta}$ is $O_p(1)$; thus, solving (9) is asymptotically equivalent to solving

$$\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}) \approx 2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta}) = 0.$$

Similarly, $\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta})$ can approximated by $2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\dot{\boldsymbol{g}}_n(\boldsymbol{\beta})$ as $\boldsymbol{R}_n(\boldsymbol{\beta}) = o_p(n)$. Although the approximation can simplify the calculation, this solution may not provide the minimum value of $Q_n(\boldsymbol{\beta})$.

Under mild conditions in Lemma 1, we may show that the weight matrix $\|\boldsymbol{A}_n(\boldsymbol{\beta}) - N^{-1}\sum_{i=1}^N \boldsymbol{V}_i\| \xrightarrow{p} 0$ as $n \to \infty$ with respect to ξ and d, where $\boldsymbol{V}_i = \boldsymbol{COV}_\xi(\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}))$. Note that $\boldsymbol{A}_0 \neq \lim_{n\to\infty} \boldsymbol{COV}_{\xi d}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}))$ under most sampling designs with the exception of simple random sampling (SRS). As a result, $\hat{\boldsymbol{\beta}}_n$ may not be optimal under the model and design joint randomization. This is different from the QIF within a non-survey context.

3.3 Example

We conducted a simulation study to compare the survey-weighted QIF to the unweighted QIF for survey data under an informative sampling design.

We use the model-design approach to simulate 1,000 samples. Finite populations are generated from the following superpopulation model:

$$Y_{it} = \beta x_{it} + \gamma_i + \epsilon_{it} \text{ for } i = 1, ..., N \text{ and } t = 1, ..., T,$$

where $N = 5,000$, $T = 10$, $\beta = 1$ and $\boldsymbol{x}_i = (0.1, 0.2, ..., 1.0)^T$; In addition, γ_i is

generated independently from $N(0, 0.7)$, and ϵ_{it} are generated independently from $N(0, 0.3)$. Thus, the vector of within-subject observation $\boldsymbol{Y}_i$ has an exchangeable correlation structure with correlation parameter 0.7.

Define a size variable Z such that $Z_i = [1+\exp(2.5-1.5\gamma_i)]^{-1}$. From each population we draw a sample of n subjects, via Rao-Sampford's method (Rao 1965, Sampford 1967), with selection probabilities proportional to size without replacement. Using the sampling weights (the inverse of the first order inclusion probabilities), we define the weighted QIF and compute the weighted QIF estimate.

We produce the unweighted GLM estimate, weighted/unweighted GEE estimates, weighted/unweighted QIF estimates, and weighted QIF estimates. For the GEE , we consider unstructured, AR1, and EX correlations. For the QIF, we consider AR1 and EX correlations. We use the relative bias (RB) and the relative efficiency (RE) criteria to evaluate the performance of different estimators under different sample sizes. The RB is defined as

$$RB(\hat{\beta}) = \frac{E[\hat{\beta}] - \beta}{\beta} \times 100\%,$$

and the RE is defined as

$$RE(\hat{\beta}) = \frac{m.s.e.(\hat{\beta})}{m.s.e.(\hat{\beta}_{WEE,true})}$$

where $\hat{\beta}_{WEE,true}$ is the weighted GEE estimator under correctly specified correlation (exchangeable in this case). In this simulation, we use $B = 1000$ replicates and so $E[\hat{\beta}] \approx \frac{1}{1000}\sum_{b=1}^{1000}\hat{\beta}^{(b)}$ and $m.s.e.(\hat{\beta}) \approx \frac{1}{1000}\sum_{b=1}^{1000}(\hat{\beta}^{(b)} - \beta)^2$ are the Monte Carlo (MC) estimates.

Unweighted GLM estimators always give the worst result since they ignore both the within-subject correlations and the sampling design. Both unweighted GEE and QIF are biased with RB around 84% for independent working correlation, 36% for AR correlation, and 21% for the exchangeable correlation (true). The RBs do not decrease as the sample size increases. Among weighted estimators, the GEE under exchangeable correlation has the smallest mean squared error (MSE). Under an exchangeable structure, the performance of QIF is slightly less efficient than the GEE when the sample size is 20 or 40, while the QIF is as good as the GEE when the sample size is 80. One possible reason for this is that the C_n matrix in QIF cannot converge to the true covariance of $\boldsymbol{g}_n$ when the sample size is small. When the working correlation is AR, the MSE of weighted GEE estimator is around 1.85 times larger than weighted QIF estimator for sample size n =20, 2.25 times as large as for sample size n =40, and 2.36 times as large as for sample size n =80. The weighted GEE and QIF may also be subject to a small bias due to the approximation of the estimation procedure and sampling process. As the sample size increases, the bias diminishes. In this simulation, the weighted unstructured GEE seems to work well for moderate or large n. However, we did not compare it to weighted QIF with unstructured correlation. We conclude that weighting should be used to remove the estimation bias due to informative sampling, and weighted QIF is better than weighted GEE when working correlation is far from the true structure, and as good as weighted GEE when it is close.

Table 6: Relative bias and relative efficiency of weighted and unweighted GEE and QIF estimators for regression paramater β, true correlation is exchangeable

Sample Size	Method	RE	RB
20	Unweighted GLM	2317%	84%
	weighted GEE unstructured	234%	14%
	Unweighted GEE exchangeable	119%	15%
	weighted GEE exchangeable	100%	4%
	Unweighted GEE AR1	764%	46%
	weighted GEE AR1	383%	10%
	Unweighted QIF AR1	534%	37%
	weighted QIF AR1	207%	10%
	Unweighted QIF exchangeable	216%	21%
	weighted QIF exchangeable	110%	4%
40	Unweighted GLM	4108%	84%
	weighted GEE unstructured	147%	8%
	Unweighted GEE exchangeable	166%	15%
	weighted GEE exchangeable	100%	3%
	Unweighted GEE AR1	1262%	45%
	weighted GEE AR1	409%	7%
	Unweighted QIF AR1	835%	36%
	weighted QIF AR1	182%	6%
	Unweighted QIF exchangeable	320%	21%
	weighted QIF exchangeable	106%	3%
80	Unweighted GLM	7383%	84%
	weighted GEE unstructured	120%	5%
	Unweighted GEE exchangeable	247%	14%
	weighted GEE exchangeable	100%	1%
	Unweighted GEE AR1	2136%	44%
	weighted GEE AR1	436%	3%
	Unweighted QIF AR1	1388%	36%
	weighted QIF AR1	184%	3%
	Unweighted QIF exchangeable	498%	21%
	weighted QIF exchangeable	100%	1%

3.4 Point Estimation

In this section, we show that the pseudo-QIF estimator is a consistent estimator of the true regression parameter $\boldsymbol{\beta}_0$. First we state the following conditions for the extended scores $\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})$ and the sampling design. Assume that

C1 the parameter space Ω is a compact subset of $\mathbb{R}^p$;

C2 $\sup_i E_\xi[h_i^4(\boldsymbol{Y}_i)] < \infty$ and $\sup_i E_\xi \|\boldsymbol{Y}_i\| < \infty$ where $h_i(\boldsymbol{Y}_i) = \sup_{\boldsymbol{\beta}\in\Omega} \|\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})\|$, $i = 1, 2, ...$ and $\|.\|$ is the Euclidean norm, i.e. $\|\boldsymbol{A}_{p,q}\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}^2} = \sqrt{Tr(\boldsymbol{A}\boldsymbol{A}^T)}$ where a_{ij} is the (i, j) element of matrix $\boldsymbol{A}_{p,q}$ with p rows and q columns;

C3 Equicontinuity of $\phi_i(\boldsymbol{\beta}) = \boldsymbol{g}_i(\boldsymbol{y}_i, \boldsymbol{\beta})$: for any $c > 0$ and sequence $\{\boldsymbol{y}_i\}$ satisfying $\|\boldsymbol{y}_i\| \leq c$, we have: there is $\delta_\epsilon > 0$ such that, if $\|t - s\| \leq \delta_\epsilon$, then $\sup_i \|\phi_i(t) - \phi_i(s)\| \leq \epsilon$ for $t, s \in \mathcal{O}$ where $\mathcal{O}$ is any open subset of Ω;

C4 For every $N \geq 1$ and $\boldsymbol{\beta} \in \Omega$, the function $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[\boldsymbol{g}_n(\boldsymbol{\beta})]$ has the property that for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|>\epsilon} \|\Delta_N(\boldsymbol{\beta})\| > \delta_\epsilon$;

C5 $\boldsymbol{A}(\boldsymbol{\beta}) = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} E[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})\boldsymbol{g}_i^T(\boldsymbol{Y}_i, \boldsymbol{\beta})]$ is positive-definite for all $\boldsymbol{\beta} \in \Omega$. $\inf_{\boldsymbol{\beta}\in\Omega} \lambda_{\min}[\boldsymbol{A}(\boldsymbol{\beta})] > 0$ and $\sup_{\boldsymbol{\beta}\in\Omega} \lambda_{\max}[\boldsymbol{A}(\boldsymbol{\beta})] < \infty$ where $\lambda_{\min}[\boldsymbol{A}]$ and $\lambda_{\max}[\boldsymbol{A}]$ is the smallest and largest eigenvalues of the matrix A, respectively. Also, $\boldsymbol{A}_n(\boldsymbol{\beta}) = N^{-1}\sum_{i\in S} w_i \boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})\boldsymbol{g}_i^T(\boldsymbol{Y}_i, \boldsymbol{\beta})$, such that $\inf_{\boldsymbol{\beta}} \lambda_{\min}[\boldsymbol{A}_n(\boldsymbol{\beta})] > 0, \forall n$ with probability one and $\sup_n \sup_{\boldsymbol{\beta}\in\Omega} \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\| < \infty$.

C6 the survey weights w_i satisfy $N^{-1}\sum_{i\in S} w_i Z_i - N^{-1}\sum_{i=1}^{N} Z_i = O_p(1/\sqrt{n})$ for any constant Z_i such that $N^{-1}\sum_{i=1}^{N} Z_i^2 = O(1)$, where "$p$" is in probability with respect to the sampling design.

These conditions are not unusual. Carrillo, Chen and Wu (2010) used similar conditions to show the model-design consistency of the survey-weighted GEE estimator. Condition C1 ensures the existence of the parameter estimator; In condition C2, fourth order boundedness is required for the convergence of $\boldsymbol{A}_n$; Equicontinuity in condition C3 is described in page 364 of Shao (2003) and used to show the convergence of a sequence of functions of random variables; under condition C4, $\boldsymbol{\beta}$ is uniquely identified; condition 5 is required for the consistency of $\boldsymbol{A}_n^{-1}$; also, $\inf_{\boldsymbol{\beta}} \lambda_{\min}[\boldsymbol{A}(\boldsymbol{\beta})] > 0$ guarantees the invertibility of $\boldsymbol{A}(\boldsymbol{\beta})$; and finally condition C6 assures the design consistency of the NHT estimator.

The following lemma plays a fundamental role for the proof of the consistency of the pseudo-QIF estimator. It provides the asymptotic unbiasedness of estimating functions under the model and the sampling design, which ensures the model-design consistency of the parameter estimator under the theory of estimating functions (Godambe and Thompson 2009).

Lemma 1. *Suppose that conditions (C1), (C2), and (C6) specified above hold. Then, as* $n \to \infty$ *and* $N \to \infty$*,*

$$\sup_{\boldsymbol{\beta} \in \Omega} \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\| \xrightarrow{p} 0 \text{ under } \xi \text{ and } d,$$

where $\boldsymbol{g}_n(\boldsymbol{\beta}) = N^{-1} \sum_{i \in S} w_i g_i(\boldsymbol{Y}_i, \boldsymbol{\beta})$, $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[\boldsymbol{g}_n(\boldsymbol{\beta})] = N^{-1} \sum_{i=1}^{N} E_{\xi}[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})]$ *and "*p*" denotes in probability with respect to both the model* ξ *and the sampling design* d*.*

The proof can be found in Carrillo, Chen and Wu (2010).

Lemma 2. *Let* $\boldsymbol{A}_n(\boldsymbol{\beta}) = N^{-1} \sum_{i \in S} w_i \boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}) \boldsymbol{g}_i^T(\boldsymbol{Y}_i, \boldsymbol{\beta})$ *and*

$\boldsymbol{A}(\boldsymbol{\beta}) = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} E_{\xi}[\boldsymbol{g}_i(\boldsymbol{Y}_i,\boldsymbol{\beta})\boldsymbol{g}_i^T(\boldsymbol{Y}_i,\boldsymbol{\beta})]$. *If conditions (C1), (C2), (C5), and (C6) hold, then as* $n\to\infty, N\to\infty$

$$\sup_{\boldsymbol{\beta}\in\Omega} \left\|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\right\| \xrightarrow{p} 0,$$

where "p" denotes in probability with respect to both the model ξ *and the sampling design* d.

Proof. First, given the fourth-order boundedness in condition (C2), we can extend the result of Lemma 1 in a straightforward manner and obtain

$$\sup_{\boldsymbol{\beta}\in\Omega} \|\boldsymbol{A}_n(\boldsymbol{\beta}) - \boldsymbol{A}(\boldsymbol{\beta})\| \xrightarrow{p} 0 \text{ under } \xi \text{ and } d.$$

as $n, N\to\infty$.

Since $\boldsymbol{A}$ is symmetric and positive-definite in (C5), $\|\boldsymbol{A}^{-1}(\boldsymbol{\beta})\| \le Tr(\boldsymbol{A}^{-1}(\boldsymbol{\beta})) \le k\max_i \lambda_i[\boldsymbol{A}^{-1}(\boldsymbol{\beta})] = k\frac{1}{\lambda_{\min}[\boldsymbol{A}(\boldsymbol{\beta})]}$ where k is the dimension of $\boldsymbol{A}$. For condition (C5),

$$\sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}^{-1}(\boldsymbol{\beta})\right\| = \frac{1}{\inf_{\boldsymbol{\beta}} \lambda_{\min}[\boldsymbol{A}(\boldsymbol{\beta})]} < \infty.$$

Also,

$$\sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\right\| = \frac{1}{\inf_{\boldsymbol{\beta}} \lambda_{\min}[\boldsymbol{A}_n(\boldsymbol{\beta})]} < \infty, \forall n, \text{ a.s.}$$

To show the convergence of inverse of matrix sequences, we write

$$\begin{aligned}
\sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\right\| &= \sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}^{-1}(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\boldsymbol{A}_n(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\right\| \\
&= \sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}^{-1}(\boldsymbol{\beta})[\boldsymbol{A}(\boldsymbol{\beta}) - \boldsymbol{A}_n(\boldsymbol{\beta})]\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\right\|
\end{aligned}$$

$$\leq \sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}^{-1}(\boldsymbol{\beta})\right\| \sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\right\| \sup_{\boldsymbol{\beta}} \left\|\boldsymbol{A}(\boldsymbol{\beta}) - \boldsymbol{A}_n(\boldsymbol{\beta})\right\|,$$

we have

$$\sup_{\boldsymbol{\beta}\in\Omega} \left\|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\right\| \to^p 0 \text{ under } \xi \text{ and } d.$$

as $n, N \to \infty$. □

Lemma 3. *Let* $Q_N(\boldsymbol{\beta}) = N\Delta_N^T(\boldsymbol{\beta})\boldsymbol{A}^{-1}(\boldsymbol{\beta})\Delta_N(\boldsymbol{\beta})$ *and* $Q_n(\boldsymbol{\beta}) = n\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta})$. *If conditions C1, C2, C4, C5 and C6 hold, then*

$$\sup_{\boldsymbol{\beta}\in\Omega} \left|n^{-1}Q_n(\boldsymbol{\beta}) - N^{-1}Q_N(\boldsymbol{\beta})\right| \xrightarrow{p} 0 \text{ under } \xi \text{ and } d.$$

Proof. First, condition (C2) $\sup_i E_\xi[h_i^4(\boldsymbol{Y}_i)] < \infty$ implies $\sup_i \|E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})]\| < M$ for some $M > 0$. Then, $\forall N \geq 1$, we have

$$\begin{aligned}
\|\Delta_N(\boldsymbol{\beta})\| &\leq \frac{1}{N}\sum_{i=1}^{N} \|E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})]\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \sup_i \|E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})]\| \\
&\leq M,
\end{aligned}$$

hence, $\sup_{\boldsymbol{\beta}} \|\Delta_N(\boldsymbol{\beta})\| < M$, $\forall N \geq 1$.

Decompose the left hand side as follows

$$\begin{aligned}
\left|n^{-1}Q_n(\boldsymbol{\beta}) - N^{-1}Q_N(\boldsymbol{\beta})\right| &= \left|\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N^T(\boldsymbol{\beta})\boldsymbol{A}^{-1}(\boldsymbol{\beta})\Delta_N(\boldsymbol{\beta})\right| \\
&\leq \left|[\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})]^T\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})[\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \\
&+ 2\left|\Delta_N^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})[\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})]\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \quad \left|\Delta_N^T(\boldsymbol{\beta})[\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})]\Delta_N(\boldsymbol{\beta})\right| \\
&\leq \quad \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\|^2 \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\| \\
&+ \quad 2\|\Delta_N^T(\boldsymbol{\beta})\| \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\| \cdot \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\| \\
&+ \quad \|\Delta_N(\boldsymbol{\beta})\|^2 \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\| \\
&\leq \quad \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\|^2 \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\| \\
&+ \quad 2M \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\| \cdot \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\| \\
&+ \quad M^2 \cdot \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\|
\end{aligned}
$$

From Lemma 1, $\sup_{\boldsymbol{\beta}\in\Omega} \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\| \to^p 0$ and from Lemma 2, $\sup_{\boldsymbol{\beta}\in\Omega} \|\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) - \boldsymbol{A}^{-1}(\boldsymbol{\beta})\| \to^p 0$. Then, as $n, N \to \infty$,

$$
\sup_{\boldsymbol{\beta}\in\Omega} \left|n^{-1}Q_n(\boldsymbol{\beta}) - N^{-1}Q_N(\boldsymbol{\beta})\right| \to^p 0 \text{ under } \xi \text{ and } d.
$$

□

Theorem 1. *Let* $Q_n(\boldsymbol{\beta}) = n\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta})$ *where* $\boldsymbol{g}_n(\boldsymbol{\beta}) = N^{-1}\sum_{i\in S} w_i\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})$ *with* $\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}) : \mathbb{R}^{T_i} \times \Omega \to \mathbb{R}^r (r > p)$ *being a sequence of vector functions. Let* $Q_N(\boldsymbol{\beta}) = N\Delta_N^T(\boldsymbol{\beta})\boldsymbol{A}^{-1}(\boldsymbol{\beta})\Delta_N(\boldsymbol{\beta})$ *where* $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[\boldsymbol{g}_n(\boldsymbol{\beta})]$. *Let* $\boldsymbol{\beta}_0 \in \Omega$ *be such that* $\Delta_N(\boldsymbol{\beta}_0) = 0$. *If conditions C1 - C6 hold, then,* $\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n(\boldsymbol{\beta})$ *exists, and as* $n \to \infty$ *(and* $N \to \infty$*),*

$$
\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0 \text{ under } \xi \text{ and } d. \tag{11}
$$

Proof. Since Ω is compact, $\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n(\boldsymbol{\beta})$ exists. Let $C = \sup_{\boldsymbol{\beta}\in\Omega} \lambda_{\max}[\boldsymbol{A}(\boldsymbol{\beta})] > 0$ as in condition C5. According to Rayleigh's principle (Horn

and Johnson 1985 Page 234),

$$\begin{aligned}
\left\|\Delta_N(\hat{\boldsymbol{\beta}}_n)\right\|^2 &= \Delta_N(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}^{-1/2}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}^{-1/2}(\hat{\boldsymbol{\beta}}_n)\Delta_N(\hat{\boldsymbol{\beta}}_n) \\
&\leq \lambda_{\max}[\boldsymbol{A}(\hat{\boldsymbol{\beta}}_n)]\left\|\boldsymbol{A}^{-1/2}(\hat{\boldsymbol{\beta}}_n)\Delta_N(\hat{\boldsymbol{\beta}}_n)\right\|^2 \\
&\leq C\left\|\boldsymbol{A}^{-1/2}(\hat{\boldsymbol{\beta}}_n)\Delta_N(\hat{\boldsymbol{\beta}}_n)\right\|^2 \\
&= C\left\|\boldsymbol{A}^{-1/2}(\hat{\boldsymbol{\beta}}_n)\Delta_N(\hat{\boldsymbol{\beta}}_n)\right\|^2 - C\left\|\boldsymbol{A}_n^{-1/2}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}_n)\right\|^2 \\
&+ C\left\|\boldsymbol{A}_n^{-1/2}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}_n)\right\|^2 \\
&\leq C\left|N^{-1}Q_N(\hat{\boldsymbol{\beta}}_n) - n^{-1}Q_n(\hat{\boldsymbol{\beta}}_n)\right| + C\frac{1}{n}Q_n(\hat{\boldsymbol{\beta}}_n) \\
&\leq C\sup_{\boldsymbol{\beta}\in\Omega}\left|N^{-1}Q_N(\boldsymbol{\beta}) - n^{-1}Q_n(\boldsymbol{\beta})\right| + \frac{C}{n}Q_n(\hat{\boldsymbol{\beta}}_n) \\
&\leq o_p(1) + \frac{C}{n}Q_n(\hat{\boldsymbol{\beta}}_n) \\
&\leq \frac{C}{n}Q_n(\boldsymbol{\beta}) + o_p(1) \\
&\leq C\left(n^{-1}Q_n(\boldsymbol{\beta}_0) - N^{-1}Q_N(\boldsymbol{\beta}_0)\right) + C\frac{1}{N}Q_N(\boldsymbol{\beta}_0) + o_p(1) \\
&= \frac{1}{N}Q_N(\boldsymbol{\beta}) + o_p(1) \text{ (as } Q_N(\boldsymbol{\beta}_0) = 0 \text{)} \\
&= o_p(1) \text{ under } \xi \text{ and } d.
\end{aligned}$$

Therefore, $\|\Delta_N(\hat{\boldsymbol{\beta}}_n)\|^2 = o_p(1)$ under ξ and d.

According to condition C4, for any $\epsilon > 0$, if $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \epsilon$, then $\|\Delta_N(\hat{\boldsymbol{\beta}}_n)\| \geq \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|>\epsilon}\|\Delta_N(\boldsymbol{\beta})\| > \delta_\epsilon$; therefore, as $N \to \infty$ and $n \to \infty$,

$$P_{\xi d}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \epsilon) \leq P_{\xi d}(\|\Delta_N(\hat{\boldsymbol{\beta}}_n)\| > \delta_\epsilon) \to 0,$$

that is, $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$ under ξ and d. □

3.5 Variance estimation

Now we consider variance estimation for the pseudo-QIF estimator $\hat{\boldsymbol{\beta}}_n$ under the joint model-design random mechanism. Linearization and replication resampling methods (bootstrap or jackknife) are commonly used for variance estimation.

We first consider the variance of extended scores $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta})$. Let $\boldsymbol{COV}_\xi$, $\boldsymbol{COV}_d$ and $\boldsymbol{COV}_{\xi d}$ denote the covariance with respect to the model, the variance w.r.t. the design, and the total variance (under both model and design), respectively. Under the framework of joint model and design random mechanisms, the total variance of $\boldsymbol{g}_n(\boldsymbol{\beta}_0)$ can be decomposed into two terms

$$\boldsymbol{COV}_{\xi d}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)) = \boldsymbol{COV}_\xi(E_d[\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)]) + E_\xi[\boldsymbol{COV}_d(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0))]. \tag{12}$$

The first term represents the variance due to the model and the second term represents the variance due to sampling design. Since $\boldsymbol{COV}_\xi(E_d[\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)]) = O(n/N)$ and $E_\xi[\boldsymbol{COV}_d(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0))] = O(1)$ (Carrillo, Chen and Wu 2010), when the sampling fraction $f = n/N$ is small, i.e. $(n << N)$, we may neglect the first term and approximate the total variance by

$$\boldsymbol{COV}_{\xi d}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)) \approx E_\xi[\boldsymbol{COV}_d(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0))]. \tag{13}$$

Then, a consistent estimator of the total variance can be taken to be

$$\hat{\boldsymbol{COV}}_{\xi d}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)) = \hat{\boldsymbol{COV}}_d(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)),$$

where $\hat{\boldsymbol{COV}}_d$ is a design unbiased variance estimator of $\boldsymbol{COV}_d$. The variance estimation requires the calculation of the second order inclusion probability. Using results 2.8.9 and 2.8.11 in Sarndal, Swensson and Wretman (1992), the Horvitz-Thompson variance estimator for any size design is given by

$$\vartheta_{HT}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)) = nN^{-2}\sum_{i\in S}\sum_{j\in S}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\boldsymbol{g}_i(\boldsymbol{\beta}_0)}{\pi_i}\frac{\boldsymbol{g}_j^T(\boldsymbol{\beta}_0)}{\pi_j},$$

where $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$. When the sample size is fixed, the Sen-Yates-Grundy variance estimator can be used to obtain

$$\vartheta_{SYG}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)) = -0.5nN^{-2}\sum_{i\in S}\sum_{i\neq j\in S}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{\boldsymbol{g}_i(\boldsymbol{\beta}_0)}{\pi_i} - \frac{\boldsymbol{g}_j(\boldsymbol{\beta}_0)}{\pi_j}\right)^2,$$

As $\boldsymbol{\beta}_0$ is unknown, we may replace it by $\hat{\boldsymbol{\beta}}_n$.

To estimate the variance of $\hat{\boldsymbol{\beta}}_n$, we expand $\dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)$ around $\boldsymbol{\beta}_0$,

$$n^{-1}\dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n) = n^{-1}\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) + n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_{\xi d}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|).$$

As $\hat{\boldsymbol{\beta}}_n$ solves the equations such that $\dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n) = 0$, and $\hat{\boldsymbol{\beta}}_n \to^p \boldsymbol{\beta}_0$ as $n \to \infty$,

$$\begin{aligned}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &\approx -\{\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\}^{-1}\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) \\ &\approx -\{2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + 2n\ddot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}_n(\boldsymbol{\beta}_0)\}^{-1}2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}_n(\boldsymbol{\beta}_0).\end{aligned}$$

Under the assumptions in Theorem 3, $\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = \boldsymbol{D}_0 + o_p(1)$ where $\boldsymbol{D}_0 = \lim_{n\to\infty} E_{\xi d}[\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)]$, $\boldsymbol{A}_n(\boldsymbol{\beta}_0) = \boldsymbol{A}_0 + o_p(1)$ where $\boldsymbol{A}_0 = \lim_{n\to\infty} E_{\xi d}[\boldsymbol{A}_n(\boldsymbol{\beta}_0)]$, and $\boldsymbol{g}_n(\boldsymbol{\beta}_0) = o_p(1)$,

$$\boldsymbol{COV}_{\xi d}(\hat{\boldsymbol{\beta}}_n) \approx [\boldsymbol{D}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{D}_0]^{-1} \boldsymbol{D}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{COV}_{\xi d}(\boldsymbol{g}_n(\boldsymbol{\beta}_0)) \boldsymbol{A}_0^{-1} \boldsymbol{D}_0 [\boldsymbol{D}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{D}_0]^{-1}. \quad (14)$$

Thus, a variance estimator of $\hat{\boldsymbol{\beta}}_n$ can be obtained by replacing $\boldsymbol{D}_0$ with $\hat{\boldsymbol{D}}_0 = \dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)$, $\boldsymbol{A}_0$ by $\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n)$, and $\boldsymbol{COV}_{\xi d}(\boldsymbol{g}_n(\boldsymbol{\beta}_0))$ by $\hat{\boldsymbol{COV}}_{\xi d}(\boldsymbol{g}_n(\boldsymbol{\beta}_0))$.

The linearization method requires knowledge of the second order inclusion probabilities π_{ij}. They are often difficult to compute under a complex designs such as probability proportional to size (PPS). Several methods have been proposed to approximate π_{ij}'s. In our simulation study, we consider their approximation for the Rao-Sampford PPS design.

In complex surveys, we may also resort to replication methods such as the the Jackknife and the bootstrap. Using replication methods, the variance of many statistics can be estimated. Krewski and Rao (1981) showed that the jackknife provides a consistent estimator of the variance when the parameter of interest is a smooth function of the population. Efron (1979, 1982) proposed the bootstrap method for non-survey data. Rao and Wu (1988) discussed a rescaling bootstrap variance estimation for stratified multi-stage sampling. Following them, Rao, Wu and Yue (1992) introduced the concept of bootstrap weights. The procedures which are used to generate bootstrap weights are described below.

1. Draw a random sample of size m_h, with replacement, independently from the n_h sample units in stratum h.

2. Let r_{hi} be the number of times that sample unit i in stratum h is selected in

the resample. Calculate bootstrap weights

$$w_i^{(b)} = [1 - (\frac{m_h}{n_h - 1})^{1/2} + (\frac{m_h}{n_h - 1})^{1/2} \frac{n_h}{m_h} r_{hi}] w_i.$$

3. Repeat the above steps B times, i.e. $b = 1, ..., B$. We set B = 500. For the b^{th} set of bootstrap weights, calculate $\hat{\theta}^{(b)}$, using the weights $w^{(b)}$ in the formula of $\hat{\theta}$.

4. The bootstrap variance estimator is given by

$$\hat{V}_B = \frac{1}{B} \sum_{i=1}^{B} [\hat{\theta}^{(b)} - \hat{\theta}]^2.$$

Kovar, Rao and Wu (1988) showed that when $m_h = n_h - 1$, the bootstrap variance estimator performs well for a smooth function. In our simulation, we use the bootstrap procedure described above.

Let $w^{(b)}$ denote the b^{th} set of bootstrap weights. For each set of weights, we compute $\boldsymbol{g}_n^{(b)}(\boldsymbol{\beta}) = \sum_{i \in S} w_i^{(b)} \boldsymbol{g}_i(y_i, \boldsymbol{\beta})$ and $Q_n^{(b)}(\boldsymbol{\beta}) = n \boldsymbol{g}_n^{(b)T}(\boldsymbol{\beta}) \boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) \boldsymbol{g}_n^{(b)}(\boldsymbol{\beta})$. The corresponding estimating equation is

$$\dot{\boldsymbol{Q}}_n^{(b)}(\boldsymbol{\beta}) \approx 2n \dot{\boldsymbol{g}}_n^{(b)T}(\boldsymbol{\beta}) \boldsymbol{A}_n^{-1}(\boldsymbol{\beta}) \boldsymbol{g}_n^{(b)}(\boldsymbol{\beta}) = 0.$$

Lipsitz, Dear and Zhao (1994) proposed an one-step bootstrap using $\hat{\boldsymbol{\beta}}_n$ as the starting point; for example, setting

$$\hat{\boldsymbol{\beta}}_n^{(1)} = \hat{\boldsymbol{\beta}}_n + \left[\ddot{\boldsymbol{Q}}_n^{(b)}(\hat{\boldsymbol{\beta}}_n) \right]^{-1} \dot{\boldsymbol{Q}}_n^{(b)}(\hat{\boldsymbol{\beta}}_n),$$

where $\ddot{\boldsymbol{Q}}_n^{(b)}(\boldsymbol{\beta}) \approx 2n\dot{\boldsymbol{g}}_n^{(b)T}(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\dot{\boldsymbol{g}}_n^{(b)}(\boldsymbol{\beta})$.

Rao and Tausi (2004) considered the estimating function (EF) bootstrap variance estimator of $\hat{\boldsymbol{\beta}}_n$ is given by

$$\hat{\boldsymbol{COV}}_B(\hat{\boldsymbol{\beta}}_n) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_n^{(b)} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n^{(b)} - \hat{\boldsymbol{\beta}}_n)^T. \tag{15}$$

where

$$\hat{\boldsymbol{\beta}}_n^{(b)} = \hat{\boldsymbol{\beta}}_n + 2n\left[\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)\right]^{-1}\dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n^{(b)}(\hat{\boldsymbol{\beta}}_n). \tag{16}$$

The EF bootstrap avoids the inversion of a large dimensional matrix and produces a more stable variance estimate (Roberts et al. 2003). We may justify (15) as follows:

$$\begin{aligned}
\hat{\boldsymbol{COV}}_B(\hat{\boldsymbol{\beta}}_n) &= \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_n^{(b)} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n^{(b)} - \hat{\boldsymbol{\beta}}_n)^T \\
&= \frac{1}{B}\sum_{b=1}^{B} 4n^2\left[\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)\right]^{-1}\dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{g}_n^{(b)}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{g}_n^{(b)T}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)\left[\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)\right]^{-1} \\
&= 4n^2\left[\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)\right]^{-1}\dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\left\{\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{g}_n^{(b)}\boldsymbol{g}_n^{(b)T}\right\}\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)\left[\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)\right]^{-1} \\
&\approx \boldsymbol{J}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{COV}}_B(\boldsymbol{g}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)\boldsymbol{J}_n^{-1}(\hat{\boldsymbol{\beta}}_n),
\end{aligned}$$

where $\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_n) = \dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)$. This will approximate the right hand side of $\boldsymbol{COV}_{\xi d}(\hat{\boldsymbol{\beta}}_n)$.

Chapter 4

Asymptotic normality and chi-squared tests

Qu, Lindsay and Li (2010) studied large sample inference properties of QIF in the non-survey context. They showed that QIF has similar properties as twice log likelihood for testing hypotheses about a regression parameter. Moreover, due to over-identification, QIF can be directly used to test the mean model assumption, while log likelihood does not have this property. In this chapter, we study the asymptotic properties of the survey-weighted QIF when making inference using data from a longitudinal survey sample. First, we establish asymptotic normality for the survey weighted QIF estimator $\hat{\boldsymbol{\beta}}_n$. Second, we construct several tests for hypotheses about the regression parameter $\boldsymbol{\beta}$, namely, a Wald test, a generalized score test, and likelihood-ratio-type test. Third, we construct a goodness-of-fit test for testing the mean model assumptions.

4.1 Asymptotic normality

The asymptotic normality of the survey-weighted QIF estimator $\hat{\boldsymbol{\beta}}_n$ is established under the joint model-design framework. Following the usual procedure under the theory of estimating functions, we first show the asymptotic normality of sample estimating functions under the joint model and design set-up. Let ξ denote the model and d denote the sampling design.

Lemma 4. *(Yuan and Jennrich, 1998, Theorem 6) Suppose that* $\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})$ *is such that*

(1) $E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}_0)] = 0$, *for all* i. *If* $\boldsymbol{V}_i = \boldsymbol{COV}_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}_0)]$, *then* $\boldsymbol{\Sigma}_\xi = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{V}_i$ *exists and is positive-definite;*

(2) there exists some $C < \infty$, *such that,* $E_\xi \|\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}_0)\|^3 < C$ *for all* i.

Then, as $N \to \infty$,

$$\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_\xi), \text{ under } \xi, \tag{17}$$

where $\boldsymbol{g}_N(\boldsymbol{\beta}_0) = 1/N \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}_0)$.

We consider the sampling survey case and further assume that survey weights w_i satisfy

$$\sqrt{n}(N^{-1} \sum_{i\in S} w_i z_i - N^{-1} \sum_{i=1}^{N} z_i) \xrightarrow{d} N(0, \sigma^2) \text{ as } N, n \to \infty \tag{18}$$

for constants z_i such that $N^{-1} \sum_{i=1}^{N} z_i^2 = O(1)$ and some non-random σ^2. This assumption is sufficient for obtaining the Central Limit Theorem (CLT) in the context of survey sampling (Binder 1983).

Conditions for the survey version of CLT have been discussed for common sampling designs; for simple random sampling by Hajek (1960), for rejective sampling by Hajek (1964), for probability proportional to the size without replacement by Rosen (1972a, 1992b, 1997), Fuller (1975), and for stratified multi-stage probability proportional to size with replacement by Krewki and Rao (1981).

Example 1 (Krewski and Rao (1981) Theorem 3.1)**.** *Consider stratified two-stage with replacement sampling design, such that* $n_h \geq 2$ *PSUs are selected from the* N_h *PSUs in the* h^{th} *stratum with probabilities* $p_{hi} > 0$ *where* $i = 1, ..., N_h$ *and* $h = 1, ..., L$. *Let* M_h *denote the total number of units in stratum* h.

Let $\boldsymbol{y}$ *be a p-dimension vector of variables of interest. Write* $\boldsymbol{Y} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \boldsymbol{Y}_{hi}$ *where* $\boldsymbol{Y}_{hi}$ *is the total of* $\boldsymbol{y}$ *in the* i^{th} *PSU of the* h^{th} *stratum; then* $\bar{\boldsymbol{Y}} = \boldsymbol{Y}/M$ *and* $M = \sum_{h=1}^{L} M_h$.

Let $\hat{\boldsymbol{Y}} = \sum_{h=1}^{L} \hat{\boldsymbol{Y}}_h$ *and* $\hat{\boldsymbol{Y}}_h = \sum_{i=1}^{n_h} \hat{\boldsymbol{Y}}_{hi}/(n_h p_{hi})$ *where* $\hat{\boldsymbol{Y}}_{hi}$ *is an unbiased estimator of* $\boldsymbol{Y}_{hi}$. *Also let* $\bar{\boldsymbol{y}} = \hat{\boldsymbol{Y}}/M$. *Assume that*

(C1) $\sum_{h=1}^{L} W_h E \left| y_{hik} - \bar{Y}_{hk} \right|^{2+\delta} = O(1)$ *for some* $\delta > 0 (k = 1, ..., p)$ *where* $W_h = M_h/M$ *denote the weight of the* h^{th} *stratum,*

(C2) $max_{1 \leq h \leq L} n_h = O(1)$,

(C3) $max_{1 \leq h \leq L} W_h = O(L^{-1})$;

(C4) $n \sum W_h^2 \boldsymbol{\Gamma}_h / n_h \rightarrow \boldsymbol{\Gamma}$ *where* $\boldsymbol{\Gamma}_h = E[(\boldsymbol{y}_{hi} - \bar{\boldsymbol{Y}}_h)(\boldsymbol{y}_{hi} - \bar{\boldsymbol{Y}}_h)^T]$ *and* $\boldsymbol{\Gamma}$ *is positive definite;*

Then, $n^{1/2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{Y}}) \rightarrow^d N(0, \boldsymbol{\Gamma})$ *as* $L \rightarrow \infty$.

Following Schenker and Welsh (1988), Chen and Rao (2007), Fuller (2009), we establish the asymptotic normality of $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)$ under the model and design joint framework.

Theorem 2. *Let* $\boldsymbol{g}_N(\boldsymbol{\beta}_0) = N^{-1}\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{Y}_i,\boldsymbol{\beta}_0)$ *and* $\boldsymbol{g}_n(\boldsymbol{\beta}_0) = N^{-1}\sum_{i\in S} w_i\boldsymbol{g}_i(\boldsymbol{Y}_i,\boldsymbol{\beta}_0)$. *Let* $\mathcal{F}$ *denote the finite population. Assume that*

1. $n/N \to f \geq 0$;
2. $\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta}_0) \xrightarrow{d} N(0,\boldsymbol{\Sigma}_\xi)$ *under* ξ, *where* $\boldsymbol{\Sigma}_\xi$ *is positive-definite;*
3. *Given* $\mathcal{F}$, $\sqrt{n}\{\boldsymbol{g}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}_N(\boldsymbol{\beta}_0)\} \xrightarrow{d} N(0,\boldsymbol{\Sigma}_\pi)$ *under sampling design* d, *where* $\boldsymbol{\Sigma}_\pi$ *is a non-stochastic value under the model.*

Then, as $n, N \to \infty$,

$$\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) \xrightarrow{d} N(0,\boldsymbol{\Sigma}_0) \text{ under } \xi \text{ and } d, \tag{19}$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_\pi + f\boldsymbol{\Sigma}_\xi$.

Proof. First, we decompose $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)$ into a sum of two random terms such that

$$\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) = \sqrt{\frac{n}{N}}\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta}_0) + \sqrt{n}(\boldsymbol{g}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}_N(\boldsymbol{\beta}_0)), \tag{20}$$

where the first term depends only on the model parameters, and the second term on the sample design and the realized population.

Let c be any non-zero vector in $\mathcal{R}^p$ of length one, i.e. $\|c\| = 1$. Then

$$c^T\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) = \sqrt{n/N}c^T\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta}_0) + c^T\sqrt{n}(\boldsymbol{g}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}_N(\boldsymbol{\beta}_0)) \tag{21}$$

$$= \quad V_n + U_n, \tag{22}$$

where, by assumption 2, $V_n = \sqrt{n/N}c^T\sqrt{N}\boldsymbol{g}_N(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \delta_v^2)$ under ξ with $\delta_v^2 = fc^T\boldsymbol{\Sigma}_\xi c$ and $\delta_v > 0$, and also, by assumption 3, given $\mathcal{F}$, $U_n = c^T\sqrt{n}(\boldsymbol{g}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}_N(\boldsymbol{\beta}_0)) \xrightarrow{d} N(0, \delta_u^2)$ with $\delta_u^2 = c^T\boldsymbol{\Sigma}_\pi c$ and $\delta_u > 0$.

Let $\Phi(t)$ denote the cumulative distribution function (CDF)of standardized normal distribution, then $\Phi(t)$ is a continuous and bounded function, by Lemma 3.2 in Rao (1962), $U_n \xrightarrow{d} N(0, \delta_u^2)$ leads to

$$\lim_{n\to\infty} \sup_{t\in R} |\Pr\{U_n/\delta_u \le t\} - \Phi(t)| = 0.$$

Let $S = \delta_u^{-1}(\delta t - {}_n)$ where $\delta = \sqrt{\delta_v^2 + \delta_u^2}$. Then, we have

$$\begin{aligned}
|\Pr\{\frac{U_n + V_n}{\delta} \le t\} - \Phi(t)| &= |\Pr\{U_n/\delta_u \le \delta_u^{-1}(\delta t - V_n)\} - \Phi(t)| \\
&= |E[\Pr\{U_n/\delta_u \le S|\mathcal{F}\}] - E[\Phi(S)] + E[\Phi(S)] - \Phi(t)| \\
&\le |E[\Pr\{U_n/\delta_u \le S|\mathcal{F}\}] - E[\Phi(S)]| + |E[\Phi(S)] - \Phi(t)| \\
&\le E[|\Pr\{U_n/\delta_u \le S|\mathcal{F}\} - \Pr(Z \le S|\mathcal{F})|] + |E[\Phi(S)] - \Phi(t)|,
\end{aligned}$$

where Z_u represents a standard normal random variable such that $Z_u \sim N(0, 1)$.

As $n \to \infty$, the first term in the right hand side satisfies

$$|\Pr\{U_n/\delta_u \le S|\mathcal{F}\} - \Pr(Z_u \le S|\mathcal{F})| \le \sup_{t\in R} |\Pr\{U_n/\delta_u \le t\} - \Phi(t)| \to 0.$$

For the second term in the right hand side, since $V_n \to^d V$ where $V \sim N(0, \delta_v^2)$

and $\Phi(\cdot)$ is continuous and bounded, by the continuous mapping theorem we obtain

$$E[\Phi(S)] = E[\Phi(\delta_u^{-1}(\delta t - V_n))] \to E[\Phi(\delta_u^{-1}(\delta t - V))] \text{ as } n \to \infty.$$

Let Z be another standard normal random variable independent of V, then

$$\begin{aligned} E[\Phi(\delta_u^{-1}(\delta t - V))] &= E[\Pr\{Z \leq \delta_u^{-1}(\delta t - V)\}] \\ &= \Pr\{\frac{\delta_u Z + V}{\delta} \leq t\} \\ &= \Phi(t), \end{aligned}$$

where we used the fact that $\frac{\delta_u Z + V}{\delta} \sim N(0,1)$. This completes the proof. □

To establish the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$, we make further assumptions.

(A) Let $\boldsymbol{\beta}_0$ be the true parameter and $N_{\boldsymbol{\beta}_0}$ be a neighbourhood of $\boldsymbol{\beta}_0$ in parameter space Ω. Assume that for all $\boldsymbol{\beta} \in N_{\boldsymbol{\beta}_0}$, $Q_n(\boldsymbol{\beta})$ is thrice differentiable.

(B) The second derivative of the QIF satisfies

$$E\{n^{-1}\frac{\partial^2 Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\}^2 < \infty \text{ for } j,k = 1,...,p.$$

The eigenvalues of $\Omega_n(\boldsymbol{\beta}) = E\{n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta})\}$ are uniformly bounded by positive constants C_1 and C_2 for all n.

(C) For all y and $\boldsymbol{\beta} \in N_{\boldsymbol{\beta}_0}$,

$$|n^{-1}\frac{\partial^3 Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k \partial \beta_l}| \leq M_{jkl}(y)$$

where $E[M_{jkl}^2(\boldsymbol{Y})] < \infty$ for all j, k, l.

These conditions are common under the theory of estimating equations, as in Theorem 5.14 in Shao (2003).

Theorem 3. *Let $\hat{\boldsymbol{\beta}}_n$ be the survey-weighted QIF estimator such that $\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n(\boldsymbol{\beta})$. Let $\boldsymbol{\beta}_0$ be the true parameter and uniquely satisfy $E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta}_0)] = 0$ for all i. Assume that $\boldsymbol{D}_0 = \lim_{n\to\infty} E_{\xi d}[\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)]$ exists and $\boldsymbol{D}_0 = O(1)$ and $\boldsymbol{A}_0 = \lim_{n\to\infty} E_{\xi d}[\boldsymbol{A}_n(\boldsymbol{\beta}_0)]$ exists and is a positive-definite matrix. Suppose (A) - (C) hold. If*

1. *$\hat{\boldsymbol{\beta}}_n$ solves the equation $\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}) = 0$, and $\hat{\boldsymbol{\beta}}_n \to^p \boldsymbol{\beta}_0$ under ξ and d as $n \to \infty$,*
2. *$\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) \to^p \boldsymbol{D}_0$ as $n \to \infty$ under ξ and d,*
3. *$\boldsymbol{A}_n(\boldsymbol{\beta}_0) \to^p \boldsymbol{A}_0$ as $n \to \infty$ under ξ and d,*
4. *$\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_0)$ as $n \to \infty$, under ξ and d.*

Then, as $n, N \to \infty$,

$$n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{J}^{-1}\left[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{\Sigma}_0 A_0^{-1}\boldsymbol{D}_0\right]\boldsymbol{J}^{-1}), \text{ under } \xi \text{ and } d, \tag{23}$$

where $\boldsymbol{J} = \boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0$.

Proof. Expanding each element in $\dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)$ around $\boldsymbol{\beta}$ leads to

$$\begin{aligned} n^{-1}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_n)}{\partial\beta_k} &= n^{-1}\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial\beta_k} + n^{-1}\partial\frac{\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial\beta_k}}{\partial\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ &+ n^{-1}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T\partial^2\frac{\frac{\partial Q_n(\boldsymbol{\beta}^*)}{\partial\beta_k}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \text{ for } k = 1, ..., p \end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_n$.

Due to the fact that $\dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n) = 0$, under ξ and d,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = -\sqrt{n}\{n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\}^{-1}\{n^{-1}\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\} + o_p(1), \tag{24}$$

since $n^{-1}\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = O_p(n^{-1/2})$, $n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = O_p(1)$ and $n^{-1}\partial^2\frac{\frac{\partial Q_n(\boldsymbol{\beta}^*)}{\partial\beta_k}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = O_p(1)$ (condition C).

Also, $\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = \boldsymbol{D}_0 + o_p(1)$ and $\boldsymbol{A}_n(\boldsymbol{\beta}_0) = \boldsymbol{A}_0 + o_p(1)$ gives

$$\begin{aligned} n^{-1}\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial\beta_k} &= 2\frac{\partial \boldsymbol{g}_n^T(\boldsymbol{\beta}_0)}{\partial\beta_k}\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}_n(\boldsymbol{\beta}_0) + \boldsymbol{g}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}\frac{\partial \boldsymbol{A}_n(\boldsymbol{\beta}_0)}{\partial\beta_k}\boldsymbol{A}_n^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}_0) \\ &= 2\frac{\partial \boldsymbol{g}_n^T(\boldsymbol{\beta}_0)}{\partial\beta_k}\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(n^{-1/2}) \text{ for } k = 1, ..., p. \end{aligned}$$

That is, $n^{-1}\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = 2\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(n^{-1/2})$. Similarly, $n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = 2\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_p(1)$. Replacing them into (24) gives

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= -[\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)]^{-1}\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(1) \\ &= -[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0 + o_p(1)]^{-1}[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1} + o_p(1)]\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(1) \\ &= -[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(1). \end{aligned}$$

By Slutsky's theorem and conditions (2)-(4), as $n, N \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{J}^{-1}\left[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{\Sigma}_0 A_0^{-1}\boldsymbol{D}_0\right]\boldsymbol{J}^{-1}).$$

under ξ and d. □

4.2 Hypothesis tests

Let $\hat{\boldsymbol{J}} = \dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)$, which is a consistent estimator of $\boldsymbol{J} = \boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0$. Let $\hat{\boldsymbol{\Sigma}}_0$ be a consistent estimator of $\boldsymbol{\Sigma}_0 = \boldsymbol{COV}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0))$. We define

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \hat{\boldsymbol{J}}^{-1}[\dot{\boldsymbol{g}}_n^T(\hat{\boldsymbol{\beta}}_n)\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{\Sigma}}_0\boldsymbol{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)]\hat{\boldsymbol{J}}^{-1}, \tag{25}$$

then $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{COV}(\sqrt{n}\hat{\boldsymbol{\beta}}_n)$.

Partition the regression parameter $\boldsymbol{\beta}$ into $\boldsymbol{\beta}^T = (\boldsymbol{\phi}^T, \boldsymbol{\psi}^T)$ where $\boldsymbol{\phi}$ is a vector of the first q component of $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ is a vector of the first $p-q$ components. We write $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ correspondingly, i.e.

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{V}}_{\boldsymbol{\phi\phi}} & \hat{\boldsymbol{V}}_{\boldsymbol{\phi\psi}} \\ \hat{\boldsymbol{V}}_{\boldsymbol{\psi\phi}} & \hat{\boldsymbol{V}}_{\boldsymbol{\psi\psi}} \end{bmatrix}. \tag{26}$$

Also, we write

$$\hat{\boldsymbol{J}} = \begin{bmatrix} \hat{\boldsymbol{J}}_{\boldsymbol{\phi\phi}} & \hat{\boldsymbol{J}}_{\boldsymbol{\phi\psi}} \\ \hat{\boldsymbol{J}}_{\boldsymbol{\psi\phi}} & \hat{\boldsymbol{J}}_{\boldsymbol{\psi\psi}} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{J}}^{-1} = \begin{bmatrix} \hat{\boldsymbol{J}}^{\boldsymbol{\phi\phi}} & \hat{\boldsymbol{J}}^{\boldsymbol{\phi\psi}} \\ \hat{\boldsymbol{J}}^{\boldsymbol{\psi\phi}} & \hat{\boldsymbol{J}}^{\boldsymbol{\psi\psi}} \end{bmatrix},$$

where $\hat{\boldsymbol{J}}^{\boldsymbol{\phi\phi}} = (\hat{\boldsymbol{J}}_{\boldsymbol{\phi\phi}} - \hat{\boldsymbol{J}}_{\boldsymbol{\phi}\psi}\hat{\boldsymbol{J}}_{\boldsymbol{\psi\psi}}^{-1}\hat{\boldsymbol{J}}_{\boldsymbol{\psi\phi}})^{-1}$, $\hat{\boldsymbol{J}}^{\boldsymbol{\phi\psi}} = -\hat{\boldsymbol{J}}_{\boldsymbol{\phi\phi}}^{-1}\hat{\boldsymbol{J}}_{\boldsymbol{\phi\psi}}(\hat{\boldsymbol{J}}_{\boldsymbol{\psi\psi}} - \hat{\boldsymbol{J}}_{\boldsymbol{\psi\phi}}\hat{\boldsymbol{J}}_{\boldsymbol{\phi\phi}}^{-1}\hat{\boldsymbol{J}}_{\boldsymbol{\phi\psi}})^{-1}$, and $\hat{\boldsymbol{J}}^{\boldsymbol{\psi\psi}} = (\hat{\boldsymbol{J}}_{\boldsymbol{\psi\psi}} - \hat{\boldsymbol{J}}_{\boldsymbol{\psi\phi}}\hat{\boldsymbol{J}}_{\boldsymbol{\phi\phi}}^{-1}\hat{\boldsymbol{J}}_{\boldsymbol{\phi\psi}})^{-1}$.

Suppose that $\boldsymbol{\phi}$ is parameter vector of interest and $\boldsymbol{\psi}$ is the nuisance parameter vector. We test $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$ vs. $H_a : \boldsymbol{\phi} \neq \boldsymbol{\phi}_0$.

4.2.1 Wald test

Using the asymptotic normality of $\hat{\boldsymbol{\beta}}_n^T = (\hat{\boldsymbol{\phi}}_n^T, \hat{\boldsymbol{\psi}}_n^T)$, we may construct the generalized Wald test with test statistic

$$T_W = n(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)^T \hat{\boldsymbol{V}}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1} (\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0). \tag{27}$$

It can be shown that

$$T_W \xrightarrow{d} \chi^2(q) \text{ as } n \to \infty, \text{ under } \xi \text{ and } d.$$

However, the inverse of $\hat{\boldsymbol{V}}_{\boldsymbol{\phi}\boldsymbol{\phi}}$ involves the estimation of $\boldsymbol{\Sigma}_0$ and is often unstable. Following Rotnitzky and Jewell (1990), we construct an alternative Wald test using

$$T_W^* = n(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)^T \hat{W}_{\boldsymbol{\phi}}^{-1} (\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0), \tag{28}$$

where $\hat{W}_{\boldsymbol{\phi}} = \hat{\boldsymbol{J}}^{\boldsymbol{\phi}\boldsymbol{\phi}}$.

Proposition 1 (Rotnitzky and Jewell, 1990)**.** *Under mild regularity conditions and provided the model is correctly specified,* $T_W^* \cong \sum_{j=1}^q c_j \chi_j^2$*, where "*$\cong$*" represents equivalence in distribution and* $\chi_j^2, j = 1, ..., q$ *are independent* $\chi^2(1)$ *random variables,* $c_1 \geq c_2 \geq \cdots \geq c_q$ *are the eigenvalues of* $\boldsymbol{P} = \boldsymbol{P}_0^{-1}\boldsymbol{P}_1$*, with*

$$\boldsymbol{P}_0 = \tilde{\boldsymbol{D}}_0^T \boldsymbol{A}_0^{-1} \tilde{\boldsymbol{D}}_0, \boldsymbol{P}_1 = \tilde{\boldsymbol{D}}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{A}_0^{-1} \tilde{\boldsymbol{D}}_0, \text{ and} \tag{29}$$

$$\tilde{\boldsymbol{D}}_0 = \begin{cases} \boldsymbol{D}_0^{(1)} - \boldsymbol{D}_0^{(2)}(\boldsymbol{D}_0^{(2)T}\boldsymbol{A}_0^{-1}\boldsymbol{D}_0^{(2)})^{-1}(\boldsymbol{D}_0^{(2)T}\boldsymbol{A}_0^{-1}\boldsymbol{D}_0^{(1)}) & (p > q) \\ \boldsymbol{D}_0 & (p = q) \end{cases} \tag{30}$$

where $\boldsymbol{D}_0^{(1)}$ *and* $\boldsymbol{D}_0^{(2)}$ *are the* $r \times q$ *and* $(r - p + q) \times q$ *matrices of the first* q *and last* $p - q$ *rows of* $\boldsymbol{D}_0$ *and* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_\pi + f\boldsymbol{\Sigma}_\xi$.

Comments:

1. If $\boldsymbol{A}_0 = \boldsymbol{\Sigma}_0$, then $c_j = 1$ for all j's, which implies $\boldsymbol{V}_{\boldsymbol{\beta}} = W_{\boldsymbol{\beta}}$ and $T_W^* = T_W \xrightarrow{d} \chi^2(q)$ under the joint ξ and d. In practice, the c_j's are usually unknown. We replace $\boldsymbol{D}_0$ by $\hat{\boldsymbol{D}}_0 = \dot{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}_n)$, $\boldsymbol{A}_0$ by $\hat{\boldsymbol{A}}_0 = \boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n)$ and $\boldsymbol{\Sigma}_0$ by $\hat{\boldsymbol{\Sigma}}_0$ to find $\hat{\boldsymbol{P}} = \hat{\boldsymbol{P}}_0^{-1}\hat{\boldsymbol{P}}_1$. In addition, c_j are replaced by $\hat{c}_j$, the eigenvalues of $\hat{\boldsymbol{P}}$.

2. We may approximate the distribution of T_W^* by a χ^2 variable. Following Rao and Scott (1981), $T_W^{*(1)} = T_W^*/\bar{c} \sim \chi_q^2$ where $\bar{c_1} = q^{-1}\sum_{j=1}^q c_j$; following Satterwaite (1946), $T_W^{*(2)} = T_W^{(1)}/(q^{-1}\sum_{j=1}^q c_j^2/\bar{c}^2) \sim \chi_r^2$ where $r = q/(q^{-1}\sum_{j=1}^q c_j^2/\bar{c}^2)$. In fact, we do not need to calculate the individual c_j since $\bar{c} = q^{-1}trace(\boldsymbol{P})$ and $\sum_{j=1}^q c_j^2 = trace(\boldsymbol{P}^2)$ (Rao and Scott, 1984).

4.2.2 Likelihood-ratio-type test

Corresponding to the partition of $\boldsymbol{\beta}$, we rewrite the QIF as $Q_n(\boldsymbol{\beta}) = Q_n(\boldsymbol{\phi}, \boldsymbol{\psi})$. To test $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$, we consider the likelihood-ratio-type test based on the survey-weighted QIF with the test statistic

$$T_{QIF}^* = Q_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) - Q_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n), \tag{31}$$

where $\breve{\boldsymbol{\psi}}_n = \arg\min_{\boldsymbol{\psi}} Q_n(\boldsymbol{\phi}_0, \boldsymbol{\psi})$ and $(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) = \arg\min_{(\boldsymbol{\phi},\boldsymbol{\psi})} Q_n(\boldsymbol{\phi}, \boldsymbol{\psi})$.

Theorem 4. *Under null hypothesis $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$, if conditions A - C in 4.1 hold,*

$$T^*_{QIF} = T^*_W + o_p(1) \text{ under } \xi \text{ and } d.$$

Proof. We use the following notation for the partial derivative and second derivative:

$$\dot{\boldsymbol{Q}}_n(\boldsymbol{\phi}, \boldsymbol{\psi}) = \begin{pmatrix} \frac{\partial Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\phi}} \\ \\ \frac{\partial Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \end{pmatrix} = \begin{pmatrix} \dot{\boldsymbol{Q}}_{n\boldsymbol{\phi}} \\ \\ \dot{\boldsymbol{Q}}_{n\boldsymbol{\psi}} \end{pmatrix}_{(\boldsymbol{\phi},\boldsymbol{\psi})}$$

and

$$\ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}, \boldsymbol{\psi}) = \begin{pmatrix} \frac{\partial^2 Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} & \frac{\partial^2 Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\psi}^T} \\ \\ \frac{\partial^2 Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\phi}^T} & \frac{\partial^2 Q_n(\boldsymbol{\phi},\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T} \end{pmatrix} = \begin{pmatrix} \ddot{\boldsymbol{Q}}_{n\boldsymbol{\phi}\boldsymbol{\phi}} & \ddot{\boldsymbol{Q}}_{n\boldsymbol{\phi}\boldsymbol{\psi}} \\ \\ \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}} & \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}} \end{pmatrix}_{(\boldsymbol{\phi},\boldsymbol{\psi})}.$$

From Theorem 1, we know $n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}, \boldsymbol{\psi}) \xrightarrow{p} 2\boldsymbol{D}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{D}_0$ and $\hat{\boldsymbol{\beta}}_n^T = (\hat{\boldsymbol{\phi}}_n^T, \hat{\boldsymbol{\psi}}_n^T)$ and $\breve{\boldsymbol{\psi}}_n$ are $\sqrt{n}$-consistent. Since $\ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}, \boldsymbol{\psi})$ is continuous in the neighbourhood of $\boldsymbol{\beta}$, we obtain

$$n^{-1}\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) = n^{-1} \begin{pmatrix} \ddot{\boldsymbol{Q}}_{n\boldsymbol{\phi}\boldsymbol{\phi}} & \ddot{\boldsymbol{Q}}_{n\boldsymbol{\phi}\boldsymbol{\psi}} \\ \\ \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}} & \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}} \end{pmatrix}_{(\hat{\boldsymbol{\phi}}_n,\hat{\boldsymbol{\psi}}_n)} \xrightarrow{p} 2\boldsymbol{D}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{D}_0 = 2 \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$$

under ξ and d. Similarly, under the null hypothesis,

$$n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) \xrightarrow{p} 2 \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}.$$

Expand the proposed test statistic as

$$\begin{aligned} Q_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) - Q_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) &= [Q_n(\boldsymbol{\phi}_0, \boldsymbol{\psi}) - Q_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n)] - [Q_n(\boldsymbol{\phi}_0, \boldsymbol{\psi}) - Q_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n)] \\ &= \frac{1}{2} \begin{pmatrix} \hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix}^T \ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) \begin{pmatrix} \hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix} \qquad (32) \\ &- \frac{1}{2} \begin{pmatrix} 0 \\ \breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix}^T \ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) \begin{pmatrix} 0 \\ \breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix} + o_p(1). \end{aligned}$$

The minimization procedures lead to two equations

$$\dot{\boldsymbol{Q}}_{n\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) = 0 \text{ and } \dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) = 0.$$

From the first equation, we have

$$0 = \dot{\boldsymbol{Q}}_{n\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) = \dot{\boldsymbol{Q}}_{n\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi}) + \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) + o_p(\sqrt{n}); \qquad (33)$$

From the second equation, we have

$$0 = \dot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) = \dot{\boldsymbol{Q}}_n(\boldsymbol{\phi}_0, \boldsymbol{\psi}) + \ddot{\boldsymbol{Q}}_n(\boldsymbol{\phi}_0, \boldsymbol{\psi}) \begin{pmatrix} \hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix} + o_p(\sqrt{n}), \tag{34}$$

so that,

$$0 = \dot{\boldsymbol{Q}}_{n\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi}) + \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) + o_p(\sqrt{n}). \tag{35}$$

Combining equations (33) and (35), we obtain

$$\ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) = \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + \ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) + o_p(\sqrt{n}).$$

Assuming $\ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})$ is non-singular,

$$(\breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) = \ddot{\boldsymbol{Q}}^{-1}_{n\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})\ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}}(\boldsymbol{\phi}_0, \boldsymbol{\psi})(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + (\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}) + o_p(1/\sqrt{n}),$$

we have

$$\begin{pmatrix} 0 \\ \breve{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \ddot{\boldsymbol{Q}}^{-1}_{n\boldsymbol{\psi}\boldsymbol{\psi}}\ddot{\boldsymbol{Q}}_{n\boldsymbol{\psi}\boldsymbol{\phi}} & \mathbf{I} \end{pmatrix}_{(\boldsymbol{\phi}_0, \boldsymbol{\psi})} \begin{pmatrix} \hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi} \end{pmatrix}.$$

Replacing this term in (32) gives

$$\begin{aligned} Q_n(\boldsymbol{\phi}_0, \breve{\boldsymbol{\psi}}_n) - Q_n(\hat{\boldsymbol{\phi}}_n, \hat{\boldsymbol{\psi}}_n) &= (\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)^T(\boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21})(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + o_p(1) \\ &= n(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)^T W_{\boldsymbol{\phi}}^{-1}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + o_p(1), \end{aligned}$$

Therefore, as $n \to \infty$ and $N \to \infty$,

$$T^*_{QIF} = T^*_W + o_p(1) \text{ under } \xi \text{ and } d.$$

□

4.3 Test for model goodness of fit H_0 : $E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})] = 0$, for all i

For non survey data, since $\hat{\boldsymbol{\beta}}_N$ is obtained by solving p linear combination of the r components of $\boldsymbol{g}_N$ to zero, there remain $r-p$ linear combinations of $\boldsymbol{g}_N$ that should be close to zero under the above model assumption. On these grounds, Qu, Lindsay and Li (2000) used $Q_N(\hat{\boldsymbol{\beta}}_N)$ to construct the goodness of fit test of marginal model mean assumption $H_0 : E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})] = 0$, for all i, that is, $Q_N(\hat{\boldsymbol{\beta}}_N) \sim \chi^2_{r-p}$ asymptotically. Hansen (1982) called this test an "over-identifying restriction" test.

Similarly, we propose a test statistic T^*_G to test the over-identifying conditions for survey sample data, using the pseudo-QIF. It is given by

$$T^*_G = Q_n(\hat{\boldsymbol{\beta}}_n). \tag{36}$$

Proposition 2. *Under null hypothesis $H_0 : E_\xi[\boldsymbol{g}_i(\boldsymbol{Y}_i, \boldsymbol{\beta})] = 0$, for all i, if conditions A- C in 4.1 hold,*

$$T^*_G \cong \sum_{j=1}^{l} c_j \chi^2_j \text{ under } \xi \text{ and } d,$$

where " $\cong$ " means asymptotic equivalence in distribution, χ^2_j, $j = 1, ..., l = r - p$ are

independent $\chi^2(1)$ *random variables, and* $c_1 \geq c_2 \geq \cdots \geq c_l > 0$ *are the eigenvalues of* $\boldsymbol{\Sigma}_0 \left\{ \boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1} \right\}$.

Proof. From assumption (3) and $\hat{\boldsymbol{\beta}}_n \to^p \boldsymbol{\beta}_0$,

$$\|n^{-1}\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n) - n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\| \xrightarrow{p} 0.$$

Also, $n^{-1}\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = 2\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0 + o_p(1)$. Thus, $n^{-1}\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n) = 2\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0 + o_p(1)$.

Expand $Q_n(\boldsymbol{\beta}_0)$ around $\hat{\boldsymbol{\beta}}_n$,

$$Q_n(\boldsymbol{\beta}_0) = Q_n(\hat{\boldsymbol{\beta}}_n) + \frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n)^T\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n) + o_p(1) \text{ under } \xi \text{ and } d. \tag{37}$$

That is,

$$\begin{aligned} Q_n(\hat{\boldsymbol{\beta}}_n) &= Q_n(\boldsymbol{\beta}_0) - \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T\ddot{\boldsymbol{Q}}_n(\hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(1) \\ &= n\boldsymbol{g}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_0^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}_0) - n\boldsymbol{g}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(1) \\ &= n\boldsymbol{g}_n^T(\boldsymbol{\beta}_0)\left\{\boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\right\}\boldsymbol{g}_n(\boldsymbol{\beta}_0) + o_p(1) \end{aligned}$$

Since $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_0)$ and $\left\{\boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\right\}$ is symmetric, by 20.28 in Seber (2007),

$$T_G^* \cong \sum_{j=1}^{l} c_j\chi_j^2,$$

where χ_j^2, $j = 1, ..., l$ are independent $\chi^2(1)$ random variables, and $c_1 \geq c_2 \geq \cdots \geq c_l$ are the eigenvalues of $\boldsymbol{\Sigma}_0 \left\{\boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\right\}$. □

Note that, when $\boldsymbol{A}_0 = \boldsymbol{\Sigma}_0$,

$$\boldsymbol{\Sigma}_0 \left\{ \boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1}\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{A}_0^{-1} \right\} = \mathbf{I}_r - \boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}$$

is idempotent. By 4.11 (c) in Seber(2007), we have

$$\begin{aligned}
\text{rank}(\mathbf{I}_r - \boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}) &= \text{trace } (\mathbf{I}_r - \boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}) \\
&= \text{trace } (\mathbf{I}_r) - \text{trace } (\boldsymbol{D}_0[\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}) \\
&= r - \text{trace } ([\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0]^{-1}\boldsymbol{D}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{D}_0) \\
&= r - p
\end{aligned}$$

as the dimension of $\boldsymbol{D}_0$ is $r \times p$. Therefore, by 20.29 in Seber (2007), $T_G^* \cong \chi^2_{r-p}$.

Chapter 5

Simulation studies

Two simulation studies are carried out: one for a continuous outcome and the other for a binary outcome. We investigate the performance of the proposed method on point estimation and inference.

5.1 Continuous outcome

We generated a finite population from the following model for a continuous outcome Y,

$$Y_{it} = \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \beta_3 x_{it}^{(3)} + \beta_4 x_{it}^{(4)} + \gamma_i + \epsilon_{it} \tag{38}$$

for $i = 1, ..., N$ and $t = 1, ..., T$, where $N = 50,000$, $T = 5$, $\boldsymbol{\beta} = (1, 0, 0, 0)^T$, $\boldsymbol{x}_i^{(1)} = (0.2, 0.4, ..., 1.0)^T$, $x_{it}^{(2)} \sim \exp(1)$, $x_{it}^{(3)} \sim N(1, 1)$, $x_{it}^{(4)} \sim N(0, 1)$, γ_i is generated independently from $N(0, 0.7)$, and ϵ_{it} is generated independently from $N(0, 0.3)$. Thus, the response vector of the within-subject observation, Y_i, has an exchangeable correlation structure with correlation parameter 0.7. Define a size variable

Z such that $Z_i = [1 + \exp(2.5 - 0.5\gamma_i)]^{-1}$. From this population, we use Rao-Sampford's method to draw B=1,000 independently sample repeatedly of size n with selection probabilities proportional to size without replacement. The selection is implemented in PROC SURVEYSELECT in SAS 9.3. We consider three sample sizes $n = 250, 500, 750$.

For each sample, sampling weights (inverse of selection probabilities) are used to construct survey-weighted QIF. We consider three types of working correlation structures: independent (IN), exchangeable (EX), and first-order autoregressive (AR1). Linearisation method (14) and bootstrap method (15) were used for variance estimation. We approximate the joint selection probabilities using the terms to $O(N^{-3})$ in Equation 5.5 in Connor (1966). Using the estimate and the variance estimate, we construct 95% confidence intervals (C.I.) $(\hat{\beta}_j - 1.96se(\hat{\beta}_j), \hat{\beta}_j + 1.96se(\hat{\beta}_j))$. We consider testing the hypothesis $H_0 : \phi = (\beta_2, \beta_3, \beta_4) = 0$ vs. H_a : at least one of them is not zero. We compute the Wald test statistic T_W, alternative Wald statistic T_W^*, and the likelihood-ratio-type test statistic T_{QIF}^*. We also consider testing the zero mean assumption $H_0 : E[g_i(\beta)] = 0$, using the goodness-of-fit statistic T_G^*.

In Tables 7 - 9, we report the relative bias (RB) of the survey-weighted QIF estimator and associated variance estimator, as well as the coverage rates of the 95% confidence intervals. We calculate the relative bias of parameter estimators and associated variance estimator as $RB(\hat{\beta}) = (E[\hat{\beta}] - \beta)/\beta$ and $RB(\hat{V}) = (E[\hat{V}] - V)/V$. We approximate $E[\hat{\beta}] \approx B^{-1}\sum_{b=1}^{B}\hat{\beta}^{(b)}$ and $E[\hat{V}] \approx B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)}$ where $\hat{\beta}^{(b)}$ and $\hat{V}^{(b)}$ are the estimate of β and associated variance estimate from the b^{th} simulated sample. The RBs for the QIF estimator are always less than 1% and decrease as n increases. Also, the RBs for the variance estimator are less than

10%. The 95% C.I. coverage is close to the nominal rate, which is consistent with the asymptotic normality of the QIF estimator. One-step EF bootstrap variance estimates are very close to the corresponding linearization variance estimates. The weighted QIF estimator and associated variance estimator perform the best when the working correlation structure is correctly specified.

In Table 10, we report the rejection rates for all the tests at significant levels 0.01, 0.05 and 0.10. For those tests whose null distribution is asymptotically equivalent to a linear combination of χ^2 variables, we consider a chi-squared distribution under the Rao-Scott first- and second-order adjustment. For the tests on regression parameters, likelihood-type test performs better than the Wald test and alternative Wald test, especially when the working correlation is AR1. When the effective sample size is small, the Wald and alternative Wald tests seem to over-reject H_0 because the small sample size may result in instability of inverse covariance matrix when the dimension of regression parameters is large. When the sample size is large, the three tests have the similar performance. For the goodness of fit test, the rejection rates of T_G^* are always close to the significance levels. Rao-Scott first-order and second-order adjustments have similar performance.

Table 7: Relative bias of the weighted QIF estimator and associated variance estimate ($\times 10^{-3}$) and coverage rates of 95% C.I.

Size	**IN**				**EX**				**AR**			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
250												
$E[\hat{\beta}]$	1.004	0.002	-0.005	0.001	1.005	0.002	-0.001	0.000	1.009	0.002	0.000	0.001
$RB(\hat{\beta})$	0.4%	-	-	-	0.5%	-	-	-	0.9%	-	-	-
$V(\hat{\beta})$	6.19	0.87	0.83	0.98	2.61	0.32	0.30	0.35	4.41	0.43	0.44	0.47
$E[\hat{V}(\hat{\beta})]$	6.34	0.84	0.85	0.99	2.56	0.30	0.31	0.32	4.29	0.40	0.40	0.43
$RB(\hat{V})$	2.4%	-3.7%	2.8%	1.3%	-1.9%	-6.2%	0.8%	-6.1%	-2.7%	-8.0%	-7.4%	-6.7%
95% C.I.	0.94	0.93	0.95	0.95	0.95	0.95	0.95	0.94	0.93	0.94	0.94	0.95
$E[\hat{V}_B(\hat{\beta})]$	6.25	0.82	0.83	0.97	2.57	0.30	0.31	0.33	4.31	0.40	0.40	0.44
$RB(\hat{V}_B)$	1.1%	-6.3%	0.6%	-0.9%	-1.5%	-6.2%	1.3%	-5.6%	-2.2%	-8.0%	-7.3%	-6.4%

Table 8: Relative bias of the weighted QIF estimator and associated variance estimate ($\times 10^{-3}$)) and coverage rates of 95% C.I.

Size	**IN**				**EX**				**AR**			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
500												
$E[\hat{\beta}]$	1.001	0.003	-0.005	0.002	1.001	0.002	-0.001	0.001	1.003	0.003	-0.001	0.001
$RB(\hat{\beta})$	0.1%	-	-	-	0.1%	-	-	-	0.3%	-	-	-
$V(\hat{\beta})$	3.12	0.41	0.43	0.54	1.31	0.16	0.17	0.18	2.22	0.20	0.21	0.24
$E[\hat{V}(\hat{\beta})]$	3.23	0.44	0.44	0.50	1.31	0.16	0.15	0.16	2.22	0.21	0.21	0.22
$RB(\hat{V})$	3.5%	8.0%	1.0%	-6.5%	0.4%	-3.7%	-7.0%	-5.9%	0.2%	3.0%	-3.6%	-6.3%
95% C.I.	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.94
$E[\hat{V}_B(\hat{\beta})]$	3.24	0.44	0.44	0.50	1.32	0.16	0.16	0.17	2.23	0.21	0.21	0.22
$RB(\hat{V}_B)$	3.9%	8.4%	1.4%	-5.9%	1.1%	-2.9%	-5.8%	-5.5%	1.4%	4.0%	-2.9%	-5.4%

Table 9: Relative bias of the weighted QIF estimator and associated variance estimate ($\times 10^{-3}$), and coverage rates of 95% C.I.

Size	**IN**				**EX**				**AR**			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
750												
$E[\hat{\beta}]$	1.002	0.002	-0.004	0.000	1.001	0.001	-0.001	0.000	1.002	0.002	-0.001	0.000
$RB(\hat{\beta})$	0.2%	-	-	-	0.1%	-	-	-	0.2%	-	-	-
$V(\hat{\beta})$	2.33	0.28	0.28	0.33	0.91	0.11	0.10	0.12	1.64	0.14	0.13	0.16
$E[\hat{V}(\hat{\beta})]$	2.16	0.29	0.29	0.34	0.88	0.10	0.10	0.11	1.50	0.14	0.14	0.15
$RB(\hat{V})$	-7.5%	1.8%	4.6%	1.1%	-3.4%	-0.9%	0.4%	-7.8%	-8.4%	-1.9%	3.2%	-5.2%
95% C.I.	0.94	0.94	0.95	0.95	0.94	0.94	0.95	0.93	0.94	0.94	0.95	0.93
$E[\hat{V}_B(\hat{\beta})]$	2.19	0.29	0.29	0.34	0.89	0.11	0.10	0.11	1.52	0.14	0.14	0.15
$RB(\hat{V}_B)$	-5.9%	3.1%	5.9%	2.2%	-2.2%	0.3%	1.5%	-7.1%	-7.3%	-1.9%	4.3%	-4.1%

Table 10: Rejection rates of Wald test, alternative Wald test, Likelihood-ratio-type test for $H_0 : \boldsymbol{\phi} = (\beta_2, \beta_3, \beta_4) = 0$ and Goodness of fit test of $H_0 : E_\xi[\boldsymbol{g}_i(Y_i, \boldsymbol{\beta})] = 0$ for all i, at level 0.01, 0.05 and 0.1

Size	**Statistics**		**IN**			**EX**			**AR**		
			0.01	**0.05**	**0.10**	**0.01**	**0.05**	**0.10**	**0.01**	**0.05**	**0.10**
250											
	Wald		0.02	0.06	0.11	0.01	0.06	0.12	0.02	0.07	0.14
	Alternative Wald	**RS1**	0.01	0.06	0.10	0.01	0.05	0.12	0.02	0.06	0.13
		RS2	0.01	0.06	0.10	0.01	0.05	0.12	0.02	0.06	0.13
	Likelihood-Type	**RS1**	0.01	0.05	0.09	0.01	0.04	0.09	0.01	0.04	0.10
		RS2	0.01	0.05	0.09	0.01	0.04	0.09	0.01	0.04	0.10
	Goodness of Fit	**RS1**	-	-	-	0.02	0.06	0.10	0.01	0.06	0.11
		RS2	-	-	-	0.01	0.06	0.10	0.01	0.05	0.11
500											
	Wald		0.01	0.06	0.13	0.02	0.07	0.12	0.01	0.07	0.13
	Alternative Wald	**RS1**	0.01	0.06	0.12	0.01	0.08	0.12	0.01	0.07	0.13
		RS2	0.01	0.06	0.12	0.01	0.08	0.12	0.01	0.07	0.13
	Likelihood-Type	**RS1**	0.01	0.05	0.11	0.01	0.06	0.10	0.01	0.06	0.11
		RS2	0.01	0.05	0.11	0.01	0.06	0.10	0.01	0.06	0.11
	Goodness of Fit	**RS1**	-	-	-	0.01	0.05	0.10	0.01	0.04	0.10
		RS2	-	-	-	0.01	0.05	0.10	0.01	0.05	0.10
750											
	Wald		0.01	0.05	0.10	0.01	0.06	0.12	0.02	0.06	0.12
	Alternative Wald	**RS1**	0.01	0.05	0.10	0.01	0.06	0.12	0.01	0.06	0.12
		RS2	0.01	0.05	0.09	0.01	0.06	0.12	0.01	0.06	0.12
	Likelihood-Type	**RS1**	0.01	0.04	0.09	0.01	0.05	0.11	0.01	0.05	0.11
		RS2	0.01	0.04	0.09	0.01	0.05	0.11	0.01	0.05	0.11
	Goodness of Fit	**RS1**	-	-	-	0.01	0.06	0.10	0.01	0.05	0.09
		RS2	-	-	-	0.01	0.06	0.10	0.01	0.05	0.09

5.2 Binary outcome

In this section, we consider a binary response and stratified sampling design. Finite populations with binary response were generated from the following model

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_{it}^{(2)} + \beta_3 x_i^{(1)} x_{it}^{(2)} + \beta_4 x_{it}^{(3)} + \beta_5 x_{it}^{(4)}, \tag{39}$$

for $i = 1, ..., N$ and $t = 1, ..., T$, where $N = 5,000$ and $T = 5$, $\pi_{it} = \Pr(Y_{it} = 1)$, $\boldsymbol{\beta} = (0.25, 0.25, 0.25, 0, 0, 0)^T$, $x_i^{(1)} = 0$ or 1 , $\boldsymbol{x}_i^{(2)} = (0, 1, \cdots, 4)^T$, $x_{it}^{(3)}$ and $x_{it}^{(4)}$ are generated from exp(1) independently. The values for the covariates are generated only once. Half of the population have $x_i^{(1)} = 0$ or $x_i^{(1)} = 1$. Pairwise odds ratios (ORs) are used to describe the within-subject association among repeated measurements of the same subject, and assumed to be the same for all subjects. The ORs are set as follows: $OR_{t,t+1} = 3.8$ for $t = 1, , 4$, $OR_{1,3} = OR_{1,4} = OR_{2,4} = 2.214$, $OR_{1,5} = OR_{2,5} = OR_{3,5} = 2.185$. Since the pairwise OR matrix is symmetric, we only need to specify the upper triangle elements. With the above set-up, we generate the observations on the response variable, using the function *rmvbin* in R-package *bindata* (Leisch, Weingessel, and Hornik, 2015).

The finite population is divided into two strata. Stratum A is such that $x_i^{(1)} = 1$. For Stratum B, $x_i^{(1)} = 0$. The stratum sizes are $N_A = N_B = 2,500$. Samples are drawn from two strata, A and B, independently under a simple random sampling design (SRS) with different sample sizes n_A and n_B. We consider three cases when the total sample sizes are 150, 300, and 450. The sample size from Stratum A is always twice as large as that drawn from Stratum B. For each total sample size, we simulate $B = 500$ samples.

For each sample, we compute the weighted QIF estimate $\hat{\beta}_n$ and associated linearization variance estimate under an AR1 working correlation. The weighted QIF estimate and associated variance estimate are used to construct 95% confidence intervals. We also calculate a bootstrap variance estimate using the Rao-Wu rescaling bootstrap weights, and a one-step EF method. In addition, we compute the Wald statistic T_W, the generalized Wald statistic T_W^* and the likelihood type test statistic T_{QIF} for $H_0 : \boldsymbol{\phi} = (\beta_3, \beta_4, \beta_5) = 0$, as well as the goodness of fit test statistic T_G^*.

Table 11 presents the relative bias (RB) of the weighted estimator and the associated linearization variance estimator. The RBs of the parameter estimator are mostly less than 5% and decrease as the sample size increases. The RBs of the variance estimator are between 10% and 20% for a sample size of 150; however, they drop dramatically when the sample size increases. Specifically, they are less than 10% for sample sizes 300 and 450. The bootstrap performs as well as the linearization method.

Table 12 gives the rejection rates of Wald test, alternative Wald test, Likelihood-ratio-type test for $H_0 : \boldsymbol{\phi} = (\beta_3, \beta_4, \beta_5) = 0$ and Goodness of fit test of $H_0 : E_\xi[g_i(Y_i, \beta)] = 0$ for all i, at level 0.05 and 0.1. For the tests on regression parameters, the Wald test T_W and alternative Wald test T_W^* have similar performance. Likelihood-type test T_{QIF}^* performs better than them when n= 150 and 300, while the three tests perform very similarly when n = 450. For the goodness of fit test, T_g^* has the good performance for all sample sizes and the rejection rates are close to the significance level. Rao-Scott first-order and second-order adjustments perform very similarly.

Figure 1 presents the observed p-values against the expected p-values for different

test statistics. The observed p-values of $Q_n(\hat{\beta}_n)$ and T_{QIF} are calculated by using the chi-squared distribution under Rao-Scott first-order adjustment. The p-values for T_W are calculated under the $\chi^2(2)$ distribution. Since the null hypothesis is true, we expect them to be very close. For a sample size of 150, the observed p-values for goodness of fit statistic and chi-squared test statistic are very close to the expected p-values and slightly different at the upper tail; the observed p-values for Wald statistic are slightly smaller than the expected p-values at the lower tail, which coincides with the underestimation of variance. When the sample size is more than 300, the observed p-values for all three statistics are very close to the expected values.

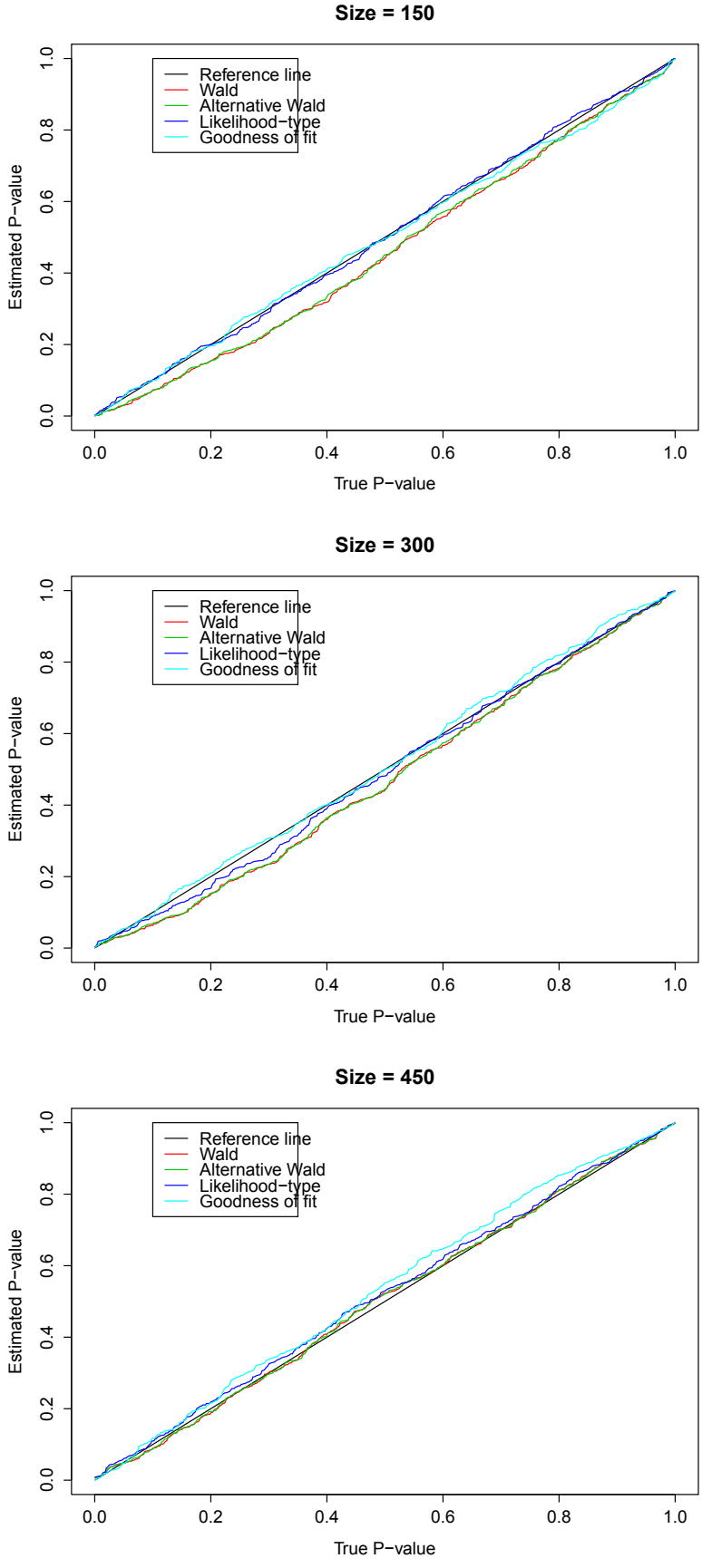
Table 11: Relative bias of the weighted QIF estimator and associated variance estimator and coverage rates of 95% C.I. (AR1 working correlation)

		β_0	β_1	β_2	β_3	β_4	β_5
	TRUE	**0.25**	**0.25**	**0.25**	**0**	**0**	**0**
				AR			
150	$E[\hat{\beta}]$	0.269	0.251	0.262	-0.004	-0.002	-0.003
	$RB(\hat{\beta})$	7.5%	0.5%	4.9%	-	-	-
	$V(\hat{\beta})$	0.076	0.113	0.010	0.016	0.020	0.022
	$E[\hat{V}(\hat{\beta})]$	0.063	0.096	0.009	0.014	0.019	0.019
	$RB(\hat{V}(\hat{\beta}))$	-18.1%	-14.3%	-9.4%	-10.1%	-5.6%	-11.5%
	95% CI	92.2%	92.8%	93.0%	94.2%	94.0%	93.6%
	$E[\hat{V}_B(\hat{\beta})]$	0.062	0.096	0.009	0.014	0.019	0.019
	$RB(\hat{V}_B(\hat{\beta}))$	-18.2%	-14.5%	-9.0%	-10.3%	-5.2%	-11.4%
300	$E[\hat{\beta}]$	0.252	0.252	0.258	-0.003	0.006	0.004
	$RB(\hat{\beta})$	0.9%	0.7%	3.1%	-	-	-
	$V(\hat{\beta})$	0.032	0.048	0.005	0.007	0.011	0.011
	$E[\hat{V}(\hat{\beta})]$	0.031	0.048	0.004	0.007	0.010	0.010
	$RB(\hat{V}(\hat{\beta}))$	-2.0%	0.1%	-3.7%	-1.7%	-14.6%	-7.2%
	95% CI	94.8%	95.0%	94.2%	93.0%	93.8%	94.8%
	$E[\hat{V}_B(\hat{\beta})]$	0.032	0.048	0.004	0.007	0.010	0.010
	$RB(\hat{V}_B(\hat{\beta}))$	-1.7%	-0.2%	-3.5%	-1.3%	-14.6%	-6.5%
450	$E[\hat{\beta}]$	0.257	0.244	0.255	-0.002	0.002	-0.007
	$RB(\hat{\beta})$	2.9%	-2.5%	1.8%	-	-	-
	$V(\hat{\beta})$	0.021	0.031	0.003	0.004	0.007	0.006
	$E[\hat{V}(\hat{\beta})]$	0.021	0.032	0.003	0.005	0.006	0.006
	$RB(\hat{V}(\hat{\beta}))$	0.7%	4.8%	2.2%	3.6%	-4.4%	5.4%
	95% CI	95.8%	95.8%	94.4%	94.4%	94.2%	95.4%
	$E[\hat{V}_B(\hat{\beta})]$	0.021	0.032	0.003	0.005	0.006	0.006
	$RB(\hat{V}_B(\hat{\beta}))$	0.7%	4.9%	2.2%	4.0%	-4.1%	-5.4%

Table 12: Rejection rates of Wald test, alternative Wald test, Likelihood-ratio-type test for $H_0 : \boldsymbol{\phi} = (\beta_3, \beta_4, \beta_5) = 0$ and Goodness of fit test of $H_0 : E_\xi[\boldsymbol{g}_i(Y_i, \boldsymbol{\beta})] = 0$ for all i, at level 0.05 and 0.1

Size			**0.05**	**0.1**
150				
	Wald		0.08	0.14
	Alternative Wald	**RS1**	0.07	0.14
		RS2	0.07	0.14
	Likelihood-Type	**RS1**	0.04	0.10
		RS2	0.04	0.10
	Goodness of fit	**RS1**	0.05	0.10
		RS2	0.05	0.09
300				
	Wald		0.08	0.16
	Alternative Wald	**RS1**	0.08	0.16
		RS2	0.08	0.16
	Likelihood-Type	**RS1**	0.06	0.12
		RS2	0.06	0.12
	Goodness of fit	**RS1**	0.04	0.10
		RS2	0.04	0.10
450				
	Wald		0.06	0.11
	Alternative Wald	**RS1**	0.05	0.11
		RS2	0.05	0.11
	Likelihood-Type	**RS1**	0.04	0.10
		RS2	0.04	0.10
	Goodness of fit	**RS1**	0.05	0.09
		RS2	0.05	0.08

Figure 1: Q-Q plots for p-values under H_0



Chapter 6

Penalized QIF for variable selection

High-dimensional data may arise from large-scale surveys or from surveys after being linked to administrative data files. These data are used to explore relationships between outcome variables and covariates or factors. Variable selection is an important topic in regression analysis. Missing important factors (underfitting) can lead to estimation bias and poor prediction performance. Including too many factors (overfitting) can make a model unnecessarily complex and difficult to interpret, and may produce unstable estimates.

Classical variable selection criteria include Mallows's C_p (Mallows 1973), the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (BIC; Schwarz 1978). All these criteria lead to a model fitting criterion penalized by the number of variables in the model (L_0-penalty). However, best subset selection procedures are subject to several limitations. First, due to the discontinuity of L_0 penalty, they cannot be carried out automatically. All possible subsets of variables need to be considered. If a large number of variables are available, these selection

procedures become computationally difficult. Second, these criteria are unstable especially when co-linearity exists and a small change in data might cause a different subset of variables to be selected (Breiman, 1996). Third, under these criteria, variable selection and estimation are carried out separately. The uncertainty in variable selection is ignored at the estimation stage. On the other hand, ridge regression (Hoerl and Kennard 1970) with a L_2-penalty is used to deal with co-linearity problem and can produce stable parameter estimates; however, it does not have the ability to do variable selection. More sophisticated procedures based on L_1-penalty have been proposed to avoid the instability of classical selection procedures. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) for multiple linear regression. Fan and Li (2001) suggested that a good model selection procedure should possess the so-called "ORACLE" properties: 1) identifying the right subset of variables to be included in the model and 2) providing the optimal estimates. They showed that the LASSO tends to overfit the model and proposed the smoothly clipped absolute deviation (SCAD) approach. Zou (2006) proposed another version of LASSO, the adaptive LASSO (ALASSO). Using these methods one can select variables and estimate unknown regression parameters simultaneously; therefore, the variability due to variable selection is taken into account in the parameter estimation.

For the marginal models approach in longitudinal data analysis, several variable selection methods have been developed. Pan (2001) proposed the quasi-likelihood information criterion (QIC) as a modification of the AIC to apply to models fitted by GEEs, assuming independent working correlation. Cantoni, Flemming, and Ronchetti (2005) proposed a generalized version of Mallow's C_p which minimizes the

prediction error. Dziak and Li (2006) proposed the generalized BIC. Dziak (2006) proposed the penalized GEE and penalized QIF with sophisticated penalties such as LASSO and SCAD. The penalized GEE method was also studied by Johnson, Lin and Zeng (2008). Wang and Qu (2009) developed a BIC procedure based on the QIF to incorporate correlation information. The penalized QIF improves the penalized GEE in estimation efficiency under mis-specified correlation. In addition, it does not require the estimation of correlation parameters.

Variable selection for complex survey data has been challenging, as one needs to consider the sampling design and variable selection jointly. Only a few papers can be found in the literature. Xu, Chen and Mantel (2013) proposed a pseudo-likelihood-based BIC criterion for variable selection. Lumley and Scott (2015) modified the AIC and BIC to handle complex survey sample data based on pseudo likelihood. For longitudinal survey data, Wang, Wang and Wang (2014) developed the penalized survey weighted GEE for marginal models. In this chapter, we consider variable selection based on survey weighted QIF with a family of L_1 penalties.

This chapter is organized as follows. In Section 2, we first define the penalized survey weighted QIF. In Section 3, we establish some asymptotic properties. In Section 4, we propose its implementation via the local quadratic approximation and tuning parameter selection. We also investigate the asymptotic behavior of the proposed BIC procedure. We use simulation studies in Section 5 to assess the performance of our approach, and provide concluding remarks in Section 6.

6.1 Penalized survey weighted QIF

Let $(\boldsymbol{y}_i, \mathbf{x}_i)$ be the measurements on unit i where $\boldsymbol{y}_i$ is the response vector and $\mathbf{x}_i$ is a matrix of d explanatory variables, given by $\mathbf{x}_{it} = (x_{it1}, ..., x_{itd})^T$. We assume that the marginal mean full model is correctly specified as

$$g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta}, \tag{40}$$

where $\mu_{it} = E[Y_{it}]$ and $g(\cdot)$ is a link function. We also assume that some coefficients in the full model are zero while others are not. Without loss of generality, we may write $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^T, \boldsymbol{\beta}_{\mathcal{N}}^T)^T$ where $\boldsymbol{\beta}_{\mathcal{A}}$ is a vector of length d_a for the active super-population coefficients, and $\boldsymbol{\beta}_{\mathcal{N}} = 0$ is a vector of length $d - d_a$ for inactive coefficients. Our goal is to identify the inactive coefficients $\boldsymbol{\beta}_{\mathcal{N}}$ and estimate the active coefficients $\boldsymbol{\beta}_{\mathcal{A}}$.

Let U denote a finite population generated from the above model. A random sample S under sampling design π is drawn from the finite population. Let w_i be the sampling weight attached to the i^{th} individual. We assume that $(\boldsymbol{y}_i, \mathbf{x}_i)$, $i \in S$ are observed for all individuals in the sample at all time points $t = 1, ..., T_i$. We assume $T_i = T, \forall i$. We are interested in constructing a variable selection procedure for the marginal models approach based on the sample survey.

Johnson, Lin, and Zeng (2008) proposed a consistent variable selection method for longitudinal data based on penalized estimating functions,

$$\boldsymbol{S}_N^{(P)}(\beta) = \boldsymbol{S}_N(\boldsymbol{\beta}) - N\boldsymbol{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\text{sgn}(\boldsymbol{\beta}) = 0 \tag{41}$$

where $\boldsymbol{S}_N$ is the finite population GEE function defined in (1), $\boldsymbol{\lambda} = (\lambda_1, .., \lambda_d)^T$,

and $\boldsymbol{q}_\lambda(\boldsymbol{\beta}) = \text{diag}\{\dot{p}_{\lambda_1}(|\beta_1|), ..., \dot{p}_{\lambda_d}(|\beta_d|)\}$ for some penalty function p_{λ_j} with a tuning parameter λ_j. Without loss of generality, we may let λ_j be the same for all β_j, that is, $\lambda_j = \lambda$.

Commonly used penalties $p_\lambda(\theta)$ for $\theta > 0$ include

1. Hard threshold penalty: $p_\lambda(\theta) = \lambda^2 - (\theta - \lambda)^2 I_{\{\theta<\lambda\}}$ and $\dot{p}_\lambda(\theta) = 2(\lambda - \theta) I_{\{\theta<\lambda\}}$.

2. LASSO: $p_\lambda(\theta) = \lambda\theta$ and $\dot{p}_\lambda(\theta) = \lambda$.

3. ALASSO: $p_\lambda(\theta) = \lambda\hat{w}\theta$ and $\dot{p}_\lambda(\theta) = \lambda\hat{w}$ where $\hat{w} = |\tilde{\theta}|^{-\gamma}$ for some $\gamma > 0$ and $\tilde{\theta}$ is the minimiser of unpenalized survey weighted QIF .

4. SCAD:

$$p_\lambda(\theta) = \lambda\theta I_{\{\theta<\lambda\}} + \frac{(a^2-1)\lambda^2 I_{\theta\geq\lambda} - (\theta - a\lambda)^2 I_{\{\lambda\leq\theta<a\lambda\}}}{2(a-1)}$$

and

$$\dot{p}_\lambda(\theta) = \lambda\left\{ I_{\{\theta<\lambda\}} + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I_{\{\theta\geq\lambda\}} \right\}$$

with $a = 3.7$ as suggested by Fan and Li (2001).

Wang, Wang, and Wang (2014) extended the penalized GEE to longitudinal survey data and proposed the penalized survey-weighted GEE

$$\boldsymbol{S}_n^{(P)}(\boldsymbol{\beta}) = \boldsymbol{S}_n(\boldsymbol{\beta}) - N\boldsymbol{q}_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\text{sgn}(\boldsymbol{\beta}) = 0 \tag{42}$$

where $\boldsymbol{S}_n$ is the unpenalized survey weighted GEE as defined in (8).

The Bayesian information criterion is used to select the tuning parameters. For

a continuous outcome,

$$BIC(\lambda) = \log\{\frac{1}{\hat{N}_T}\sum_{i\in S}\sum_{t=1}^{T} w_i e_{it}(\hat{\boldsymbol{\beta}}_\lambda)\} + \frac{\log(n)}{n} df(\hat{\boldsymbol{\beta}}_\lambda)$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the resulting estimator of $\boldsymbol{\beta}$ for tuning parameter λ, $\hat{N}_T = T\sum_{i\in S} w_i$, $e_{it}(\boldsymbol{\beta}) = y_{it} - \mu_{it}(\boldsymbol{\beta})$ and $df(\boldsymbol{\beta})$ is the number of non-zero elements of $\boldsymbol{\beta}$ (also is the number of predictors entering the model). For a binary outcome, we consider

$$BIC(\lambda) = -2QL(\hat{\boldsymbol{\beta}}_\lambda) + \frac{\log(n)}{n} df(\hat{\boldsymbol{\beta}}_\lambda)$$

where $QL(\beta)$ is the quasi-likelihood for binary data with logit link under working independence (Dziak 2006). Here the tuning parameter λ is selected such that $\lambda = \text{argmin}_\lambda BIC(\lambda)$. For survey data, Lumley and Scott (2015) propose $dBIC(\lambda)$ that replaces $\log(n)$ by $\log(n^*)$, where n^* is the effective sample size, in $BIC(\lambda)$ to account for the sampling design. Note that $dBIC(\lambda)$ and $BIC(\lambda)$ are asymptotically similar as the design effects are $O(1)$. For small or moderate sample sizes, Lumley and Scott (2015) show that $dBIC(\lambda)$ performs better than $BIC(\lambda)$ when the design effect is not close to one.

We propose here estimation methods based on penalized survey weighted QIF for longitudinal survey data. The penalized survey weighted QIF is defined as

$$Q_n^P(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta}) + k_n \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \tag{43}$$

where $Q_n(\boldsymbol{\beta}) = n\boldsymbol{g}_n^T(\boldsymbol{\beta})\boldsymbol{A}_n^{-1}(\boldsymbol{\beta})\boldsymbol{g}_n(\boldsymbol{\beta})$ is the survey weighted QIF as defined in Chapter

3, $k_n = O(n)$ is a positive scaling parameter, the penalties $p_{\lambda_n}(\cdot)$ are non-negative functions, symmetric at zero and have a continuous second derivative except possibly at zero, and λ_n is the tuning parameter. We consider only the SCAD penalty. Dziak (2006) showed that the best choice of k_n is $k_n = 2n{\cdot}T/\hat{\sigma}^2$ for linear models, $k_n = 2n{\cdot}T$ for the GLM with $\phi = 1$, or $k_n = 2n \cdot T/\hat{\phi}$ when over-dispersion is present. The penalized QIF estimator of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n^P(\boldsymbol{\beta}). \tag{44}$$

6.2 Local quadratic approximation (LQA)

The minimization of the function defined in (43) is challenging because the L_1-family penalties are singular at $\beta_j = 0$ and their second derivatives, $\ddot{p}_\lambda(|\beta_j|)$, are not continuous. Fan and Li (2001) proposed a unified algorithm to solve the minimization problem via a local quadratic approximation (LQA). The LQA can be used for a more general penalized form

$$PL(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + k_n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{45}$$

where $l(\boldsymbol{\beta})$ is a loss function for $\boldsymbol{\beta}$; for example, a negative log-likelihood, a sum of squared errors or the QIF. Let $\boldsymbol{\beta}_{opt}$ be the minimizer of $PL(\boldsymbol{\beta})$ and $\boldsymbol{\beta}^{(0)}$ be an initial value close to $\boldsymbol{\beta}_{opt}$. In practice $\boldsymbol{\beta}^{(0)}$ can be found by minimizing $l(\boldsymbol{\beta})$. If $\beta_j^{(0)}$ is close

to zero, we set $\beta_j = 0$. As $\beta_j \approx \beta_{0j}$, we can locally approximate $p_\lambda(|\beta_j|)$ by

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{\dot{p}_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}). \tag{46}$$

This reduces the minimization problem to minimizing a quadratic form with respect to $\boldsymbol{\beta}$

$$l(\boldsymbol{\beta}^{(0)}) + \dot{l}(\boldsymbol{\beta}^{(0)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^T\ddot{l}(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{P}_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta} \tag{47}$$

where $\dot{l}(\boldsymbol{\beta}^{(0)}) = \frac{\partial l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}}$, $\ddot{l}(\boldsymbol{\beta}^{(0)}) = \frac{\partial^2 l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}$, and

$$P_\lambda(\boldsymbol{\beta}^{(0)}) = k_n \text{diag}\{\dot{p}_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \cdots, \dot{p}_\lambda(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}. \tag{48}$$

We apply the Newton-Raphson algorithm

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} - \{\ddot{l}(\boldsymbol{\beta}^{(0)}) + \boldsymbol{P}_\lambda(\boldsymbol{\beta}^{(0)})\}^{-1}\{\dot{l}(\boldsymbol{\beta}^{(0)}) + \boldsymbol{U}_\lambda(\boldsymbol{\beta}^{(0)})\} \tag{49}$$

where $\boldsymbol{U}_\lambda(\boldsymbol{\beta}^{(0)}) = \boldsymbol{P}_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(0)}$.

An iterative algorithm is suggested where

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} - \{\ddot{l}(\boldsymbol{\beta}^{(k-1)}) + \boldsymbol{P}_\lambda(\boldsymbol{\beta}^{(k-1)})\}^{-1}\{\dot{l}(\boldsymbol{\beta}^{(k-1)}) + \boldsymbol{U}_\lambda(\boldsymbol{\beta}^{(k-1)})\}. \tag{50}$$

At each iteration, if some coefficient $\beta_j^{(k-1)}$ has an absolute value that is less than a pre-defined cut-off, say 0.001, then this coefficient is set to zero at the next iteration.

6.3 Asymptotic properties

Regularity conditions on the quadratic inference function are required to establish asymptotic properties of the penalized estimator. Similarly, we assume

(A) Let $\boldsymbol{\beta}_0$ be the true parameter and $N_{\boldsymbol{\beta}_0}$ be a neighbourhood of $\boldsymbol{\beta}_0$ in parameter space Ω. Assume that for all $\boldsymbol{\beta} \in N_{\boldsymbol{\beta}_0}$, $Q_n(\boldsymbol{\beta})$ is thrice differentiable.

(B) The second derivative of the QIF satisfies

$$E\{n^{-1}\frac{\partial^2 Q_n(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\}^2 < \infty \text{ for } j,k = 1,...,d.$$

The eigenvalues of $\Omega_n(\boldsymbol{\beta}) = E\{n^{-1}\ddot{Q}_n(\boldsymbol{\beta})\}$ are uniformly bounded by positive constants C_1 and C_2 for all n.

C. For all $\boldsymbol{y}$ and $\boldsymbol{\beta} \in N_{\boldsymbol{\beta}_0}$,

$$|n^{-1}\frac{\partial^3 Q_n(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k\partial\beta_l}| \leq M_{jkl}(\boldsymbol{y}), \text{ almost surely}$$

where $E[M_{jkl}^2(\boldsymbol{y})] < \infty$ for all j,k,l.

We also assume that the penalty $p_{\lambda_n}(|\beta_j|)$ satisfies the following conditions (Wang, Wang, and Wang 2014),

(1) for nonzero fixed $\beta_j \neq 0$, $\lim_{n\to\infty} \max_j n^{1/2}\dot{p}_{\lambda_n}(|\beta_j|) = 0$, and $\lim_{n\to\infty} \max_j n^{1/2}\ddot{p}_{\lambda_n}(|\beta_j|) = 0$,

(2) for any constant $M > 0$, $\lim_{n\to\infty} n^{1/2} \inf_{|\beta_j|<Mn^{-1/2}} \dot{p}_{\lambda_n}(|\beta_j|) = \infty$.

Both conditions are important for establishing the "ORACLE" properties. the first one ensures the $\sqrt{n}$-consistency and the second one leads to sparsity. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, the SCAD and hard threshold penalties satisfy conditions (1) and (2); if $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$, the ALASSO penalty satisfies both conditions. However, in the case of the LASSO penalty, conditions (1) and (2) cannot hold at the same time.

Theorem 5. *($\sqrt{n}$- consistency) Let $\boldsymbol{\beta}_0$ be the true parameter. Under the assumptions (A) - (C) and conditions (1) and (2) on penalty functions, there exists a local minimiser of $Q_n^P(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_n$, such that*

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2}),$$

where "p" is under ξ and π.

Proof. Following Fan and Li (2001), it suffices to prove that $\forall \epsilon > 0$, with probability at least $1-\epsilon$, there exist some constant C_ϵ such that a local minimiser $\hat{\boldsymbol{\beta}}_n$ exists in the interior of the ball $\left\{\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u} : \|\boldsymbol{u}\| \le C_\epsilon\right\}$.

$$\begin{aligned}
\text{Diff}(n, \boldsymbol{u}) &= Q_n^P(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}) - Q_n^P(\boldsymbol{\beta}_0) \\
&= Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}) - Q_n(\boldsymbol{\beta}_0) + k_n \sum_{j=1}^{d} \left[p_{\lambda_n}(\left|\beta_{0j} + n^{-1/2}u_j\right|) - p_{\lambda_n}(|\beta_{0j}|)\right] \\
&= n^{-1/2}\boldsymbol{u}^T \dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) + \frac{1}{2}n^{-1}\boldsymbol{u}^T \left[\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\right] \boldsymbol{u} + \frac{1}{6}n^{-3/2} \sum_{j,k,l=1}^{d} \frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_j \partial\beta_k \partial\beta_l} u_j u_k u_l \\
&+ k_n \sum_{j=1}^{d} \dot{p}_{\lambda_n}(|\beta_{0j}|) n^{-1/2} u_j + \frac{k_n}{2} \sum_{j=1}^{d} \ddot{p}_{\lambda_n}(|\beta_{0j}^{**}|) n^{-1} u_j^2.
\end{aligned}$$

In the above right hand side, for the first term,

$$n^{-\frac{1}{2}}\boldsymbol{u}^T\dot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0) = n^{-\frac{1}{2}}\boldsymbol{u}^T[2n\boldsymbol{g}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_p(n^{1/2})] = O_p(1)\left\|\boldsymbol{u}\right\|,$$

as $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) = O_p(1)$, $\boldsymbol{A}_n(\boldsymbol{\beta}_0) = O_p(1)$, and $\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = O_p(1)$; for the second term,

$$\frac{1}{2}n^{-1}\boldsymbol{u}^T\left[\ddot{\boldsymbol{Q}}_n(\boldsymbol{\beta}_0)\right]\boldsymbol{u} = \frac{1}{2}n^{-1}\boldsymbol{u}^T\left[2n\dot{\boldsymbol{g}}_n^T(\boldsymbol{\beta}_0)\boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_0)\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0 + o_p(n)\right]\boldsymbol{u} = O_p(1)\left\|\boldsymbol{u}\right\|^2;$$

for the third term, $\frac{1}{6}n^{-3/2}\sum_{j,k,l=1}^{d}\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_k\partial\beta_l}u_ju_ku_l = O_p(n^{-1/2})\left\|\boldsymbol{u}\right\|^3$ as $n^{-1}\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_k\partial\beta_l} = O_p(1)$ according to condition (C); for the fourth term, $k_n\sum_{j=1}^{d}\dot{p}_{\lambda_n}(|\beta_{0j}|)n^{-1/2}u_j = o(1)\left\|\boldsymbol{u}\right\|$ by condition (1); for the fifth term, $\frac{k_n}{2}\sum_{j=1}^{d}\ddot{p}_{\lambda_n}(|\beta_{0j}^{**}|)n^{-1}u_j^2 = o(1)\left\|\boldsymbol{u}\right\|^2$ by condition (1).

By choosing a sufficiently large C_ϵ and n, the second term dominates the others; then,

$$\text{Diff}(n, \boldsymbol{u}) > 0.$$

Therefore, there exists at least one local minimiser of $Q_n^P(\boldsymbol{\beta})$ inside the ball. □

Lemma 5 (Sensitivity)**.** *The active coefficients are included in the model with probability approaching one, i.e.,* $\Pr(\exists j \in \mathcal{A} : \hat{\beta}_{nj} = 0) = o(1)$.

Proof. Let $\epsilon > 0$ be arbitrary fixed. Since $\beta_{0j} \neq 0$ for $j \in \mathcal{A}$, if for some n, $\exists j \in \mathcal{A}$ such that $\hat{\beta}_{nj} = 0$, then $|\hat{\beta}_{nj} - \beta_{0j}| > \epsilon$. Hence,

$$\begin{aligned}\Pr(\exists j \in \mathcal{A} : \hat{\beta}_{nj} = 0) &< \Pr(\exists j \in \mathcal{A} : |\hat{\beta}_{nj} - \beta_{0j}| > \epsilon) \\ &\leq \Pr(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \epsilon)\end{aligned}$$

$$= o(1), \text{ using Theorem 5.}$$

□

Lemma 6 (Sparsity)**.** *There exists a local minimiser of $Q_n^P(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_n$, such that $\hat{\boldsymbol{\beta}}_{n\mathcal{A}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_{0\mathcal{A}}$, and that with probability converging to one, $\hat{\boldsymbol{\beta}}_{n\mathcal{N}} = \boldsymbol{\beta}_{0\mathcal{N}} = 0$.*

Proof. It is sufficient to show that with probability one, for any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, and for some constant $C > 0$ and all $j \in \mathcal{N}$, we have

$$\begin{aligned} &\frac{\partial Q_n^P(\boldsymbol{\beta})}{\partial \beta_j} > 0, && \text{for } 0 < \beta_j < Cn^{-1/2}; \\ &\frac{\partial Q_n^P(\boldsymbol{\beta})}{\partial \beta_j} < 0, && \text{for } -Cn^{-1/2} < \beta_j < 0. \end{aligned} \tag{51}$$

To show (51), by Taylor's expansion, we have

$$\begin{aligned} \frac{\partial Q_n^P(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} + k_n \dot{p}_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= \frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{0l}) \\ &\quad + \frac{1}{2}\sum_{k,l=1}^{d} \frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \partial \beta_k}(\beta_l - \beta_{0l})(\beta_k - \beta_{0k}) + k_n \dot{p}_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= I + II + III + IV, \end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. In the right hand side of the above equation, we

first have $I = \frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_p(n^{1/2})$; second,

$$II = \sum_{l=1}^{d} \frac{\partial Q_n^2(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) = \sum_{l=1}^{d} O_p(n)(\hat{\beta}_l - \beta_{0l}) = O_p(n^{1/2})$$

since $\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} = O_p(n)$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$; third, as $\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \partial \beta_k} = O_p(n)$

$$III = \frac{1}{2} \sum_{k,l=1}^{d} \frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \partial \beta_k}(\beta_l - \beta_{0l})(\beta_k - \beta_{0k}) = \sum_{k,l=1}^{d} O_p(n)(\beta_l - \beta_{0l})(\beta_k - \beta_{0k}) = O_p(1);$$

fourth, $k_n n^{-1/2} = O(n^{1/2})$ as $k_n = O(n)$ is positive and $n^{1/2}\dot{p}_{\lambda_n}(|\beta_j|) \to \infty$ as $n \to \infty$ for $|\beta_j| < Cn^{-1/2}$ by condition (2). Then, for $j \in \mathcal{N}$,

$$\begin{aligned} \frac{\partial Q_n^P(\boldsymbol{\beta})}{\partial \beta_j} &= k_n \dot{p}_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p(n^{1/2}) \\ &= k_n n^{-1/2}\left[n^{1/2}\dot{p}_{\lambda_n}(|\beta_j|)\right]\mathrm{sgn}(\beta_j) + O_p(n^{1/2}) \\ &= k_n n^{-1/2}\left[n^{1/2}\dot{p}_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p(1)\right], \end{aligned} \tag{52}$$

thus, as $n \to \infty$, the sign of $\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j}$ is determined by the sign of β_j, which completes our proof. □

Let $\boldsymbol{\Sigma}_0 = \lim_{n\to\infty} \boldsymbol{COV}(\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0))$ and $\boldsymbol{A}_0 = \lim_{n\to\infty} E[\boldsymbol{A}_n(\boldsymbol{\beta}_0)]$. Also, let

$$\boldsymbol{D}_0 = \lim_{n\to\infty} E[\dot{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)] = (\boldsymbol{D}_{0\mathcal{A}}^T, \boldsymbol{D}_{0\mathcal{N}}^T)^T.$$

Theorem 6 (Normality)**.** *Under conditions (A) - (C) and (1)- (2),*

$$\sqrt{n}(\boldsymbol{\Omega}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n \mathcal{A}})\left[\hat{\boldsymbol{\beta}}_{n\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} + (\boldsymbol{\Omega}_{\mathcal{A}} + \boldsymbol{\Sigma}_{\lambda_n \mathcal{A}})^{-1}\boldsymbol{b}_n\right] \to^d N(0, V_{\mathcal{A}})$$

where

$$\boldsymbol{\Omega}_{\mathcal{A},n} = n^{-1}\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\beta_{\mathcal{A}}\partial\beta_{\mathcal{A}}^T},$$

$$\boldsymbol{\Sigma}_{\lambda_n\mathcal{A}} = n^{-1}k_n \cdot diag\{\ddot{p}_{\lambda_n}(|\beta_{01}|), \cdots, \ddot{p}_{\lambda_n}(|\beta_{0d_a}|)\},$$

$$\boldsymbol{b}_n = n^{-1}k_n(\dot{p}_{\lambda_n}(|\beta_{0j}|)sgn(\beta_{01}), \cdots, \dot{p}_{\lambda_n}(|\beta_{0j}|)sgn(\beta_{0d_a})),$$

$$and\ \boldsymbol{V}_{\mathcal{A}} = 4\boldsymbol{D}_{0\mathcal{A}}^T\boldsymbol{A}_0^{-1}\boldsymbol{\Sigma}_0 A_0^{-1}\boldsymbol{D}_{0\mathcal{A}}.$$

Proof. Using Lemma 6, let $\hat{\boldsymbol{\beta}}_{n\mathcal{A}}$ be a $\sqrt{n}-$consistent estimator of $\boldsymbol{\beta}_{0\mathcal{A}}$, then $\hat{\boldsymbol{\beta}}_{n\mathcal{A}}$ is a local minimiser of the function $Q_n^P(\hat{\boldsymbol{\beta}}_n)$, and so

$$\frac{\partial Q_n^P(\hat{\boldsymbol{\beta}}_n)}{\partial\beta_j} = 0, \text{ with } j \in \mathcal{A}.$$

By a Taylor series expansion around β_{0j} for $j \in \mathcal{A}$, we have

$$\begin{aligned}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_n)}{\partial\beta_j} + k_n\dot{p}_{\lambda_n}(|\hat{\beta}_{nj}|)\text{sgn}(\hat{\beta}_{nj}) = \frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial\beta_j} &+ \sum_{l=1}^{d_a}\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l}(\hat{\beta}_{nl} - \beta_{0l}) \\ &+ \frac{1}{2}\sum_{l,l'=1}^{d_a}\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_{l'}}(\hat{\beta}_{nl} - \beta_{0l})(\hat{\beta}_{nl'} - \beta_{0l'}) \\ &+ k_n\dot{p}_{\lambda_n}(|\beta_{0j}|)\text{sgn}(\beta_{0j}) + k_n[\ddot{p}_{\lambda_n}(|\beta_{0j}|) + o_p(1)](\hat{\beta}_{nj} - \beta_{0j}),\end{aligned}$$

where β_j^* lies between β_{0j} and $\hat{\beta}_{nj}$. That is,

$$-n^{-1}\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_{\mathcal{A}}} = \left[n^{-1}\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_{\mathcal{A}}\partial\boldsymbol{\beta}_{\mathcal{A}}^T} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}}\right](\hat{\boldsymbol{\beta}}_{n\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) + \boldsymbol{b}_n + o_p(n^{-1/2}),$$

as $n^{-1}k_n\dot{p}_{\lambda_n}(|\beta_{0j}|)\text{sgn}(\beta_{0j}) = o_p(1)$ by condition (1) and $n^{-1}\sum_{l,l'=1}^{d_a}\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_{l'}}(\hat{\beta}_{nl} - \beta_{0l})(\hat{\beta}_{nl'} - \beta_{0l'}) = o_p(1)$ by condition (C) and the $\sqrt{n}-$consistency of $\hat{\boldsymbol{\beta}}_{n\mathcal{A}}$.

Then, we have

$$-2\boldsymbol{D}_{0\mathcal{A}}^T\boldsymbol{A}_0^{-1}\boldsymbol{g}_n(\boldsymbol{\beta}_0) = \left[n^{-1}\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_{\mathcal{A}}\partial\boldsymbol{\beta}_{\mathcal{A}}^T} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}}\right](\hat{\boldsymbol{\beta}}_{n\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) + \boldsymbol{b}_n + o_p(n^{-1/2}).$$

As a result,

$$\sqrt{n}(\boldsymbol{\Omega}_{\mathcal{A},n}+\boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})\left[\hat{\boldsymbol{\beta}}_{n\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} + (\boldsymbol{\Omega}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})^{-1}\boldsymbol{b}_n\right] = -2\boldsymbol{D}_{0\mathcal{A}}^T(\boldsymbol{\beta}_0)\boldsymbol{A}_0^{-1}\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0)+o_p(1).$$

We have shown that $\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\beta}_0) \to^d N(0, \boldsymbol{\Sigma}_0)$. Therefore, as $n \to \infty$,

$$\sqrt{n}(\boldsymbol{\Omega}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})\left[\hat{\boldsymbol{\beta}}_{n\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} + (\boldsymbol{\Omega}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})^{-1}\boldsymbol{b}_n\right] \to^d N(0, \boldsymbol{V}_{\mathcal{A}}).$$

□

An asymptotic variance estimator of $\hat{\boldsymbol{\beta}}_{n\mathcal{A}}$ is given as

$$n^{-1}(\hat{\boldsymbol{\Omega}}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})^{-1}\hat{\boldsymbol{V}}_{\mathcal{A}}(\hat{\boldsymbol{\Omega}}_{\mathcal{A},n} + \boldsymbol{\Sigma}_{\lambda_n\mathcal{A}})^{-1} \tag{53}$$

where $\hat{\boldsymbol{\Omega}}_{\mathcal{A},n} = n^{-1}\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_n)}{\partial\boldsymbol{\beta}_{\mathcal{A}}\partial\boldsymbol{\beta}_{\mathcal{A}}^T}$ and $\hat{\boldsymbol{V}}_{\mathcal{A}} = 4\hat{\boldsymbol{D}}_{0\mathcal{A}}^T\hat{\boldsymbol{A}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{A}}_0^{-1}\hat{\boldsymbol{D}}_{0\mathcal{A}}$.

Bootstrap methods can also be used for variance estimation. We generate bootstrap weights using the Rao-Wu rescaling method and consider a one-step bootstrap method. Following (50), we calculate the estimate for each set of bootstrap weights,

$$\hat{\boldsymbol{\beta}}^{(b)} = \hat{\boldsymbol{\beta}}_n - \{\ddot{\boldsymbol{Q}}^{(b)}(\hat{\boldsymbol{\beta}}_n) + \boldsymbol{P}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n)\}^{-1}\{\dot{\boldsymbol{Q}}^{(b)}(\hat{\boldsymbol{\beta}}_n) + \boldsymbol{U}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n)\}, \tag{54}$$

where $\boldsymbol{P}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n)$ replaces $\hat{\boldsymbol{\beta}}_n$ by $\boldsymbol{\beta}^{(0)}$ in (49), $\boldsymbol{U}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{P}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{\beta}}_n$, and $\boldsymbol{Q}^{(b)}(\hat{\boldsymbol{\beta}}_n)$ is the bootstrap weighted QIF calculated by using the b^{th} set of bootstrap weights. The diagonal elements of $\boldsymbol{P}_{\lambda_n}(\hat{\boldsymbol{\beta}}_n)$ and $\boldsymbol{U}_{\lambda}(\hat{\boldsymbol{\beta}}_n)$ corresponding to $\boldsymbol{\beta}_{\mathcal{N}}$ are set to zero.

6.4 Selection of tuning parameter

The performance of penalized survey weighted QIF relies on the choice of tuning parameter λ_n. Fan and Li (2001) proposed generalized cross-validation (GCV) to choose the tuning parameter. However, Wang, Li and Tsai (2007a) showed that the GCV approach is similar to the AIC (Wang, Li and Tsai 2007b) and tends to overfit the model and select null variables as non-zero components. In contrast, the BIC is able to identify the true model consistently. Dziak (2006) carefully studied the performance of different tuning parameter selectors for penalized QIF through simulation and the BIC worked best for the penalized QIF. For longitudinal survey data, Wang, Wang, and Wang (2014) selected the tuning parameter for the penalized GEE that minimized the survey weighted BIC (WBIC). Wang and Qu (2009) proposed the BIQIF selection criterion, which is a version of the BIC using the QIF as an objective function. Cho and Qu (2013) used the BIQIF as a tuning parameter selector for longitudinal data with a diverging number of parameters. In our work, we propose the survey weighted BIQIF for selecting the tuning parameter.

Let $\Theta = (0, \lambda_{max})$ be the range within which we consider the tuning parameter λ_n. The upper limit λ_{max} satisfies $\lambda_{max} \to 0$ as $n \to \infty$. For an arbitrary tuning

parameter $\lambda \in \Theta$ and each n, we define a function WBIC(λ) as

$$\text{WBIC}(\lambda) = Q_n(\hat{\boldsymbol{\beta}}_\lambda) + \log(n) df(\hat{\boldsymbol{\beta}}_\lambda)$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the penalized coefficient obtained by minimizing (43) and $df(\hat{\boldsymbol{\beta}}_\lambda)$ is the number of non-zero components of $\hat{\boldsymbol{\beta}}_\lambda$. Therefore, $\lambda_n = \arg\min_{\lambda \in \Theta} \text{WBIC}(\lambda)$ and $\hat{\boldsymbol{\beta}}_{\lambda_n}$ is the penalized QIF estimate of $\boldsymbol{\beta}$.

Let c_0 denote a set of indices of covariates whose corresponding coefficients are non-zero in the true model. Let $C_0 = \{c_0\}$, $C_+ = \{c : c_0 \subset c, c \neq c_0\}$, and $C_- = \{c : c_0 \not\subset c\}$, where $\not\subset$ denotes that at least one element in c_0 is not in c. Then, C_+ gives the set of indices for all over-fitted models and C_- gives the set of indices for all under-fitted models.

For a given tuning parameter, λ, let c_λ denote the set of indices of selected covariates with non-zero coefficients. Partition Θ into three parts based on selected models

$$\Theta_+ = \{\lambda : c_\lambda \in C_+\}, \Theta_- = \{\lambda : c_\lambda \in C_-\}, \Theta_0 = \{\lambda : c_\lambda = c_0\}.$$

Let $\hat{\boldsymbol{\beta}}_T$ be the unpenalized QIF estimate under C_0 and define

$$\text{WBIC}_0 = Q_n(\hat{\boldsymbol{\beta}}_T) + \log(n) df(\hat{\boldsymbol{\beta}}_T).$$

Lemma 7. *If conditions (A) - (C) and (1) - (2) hold,*

$$\Pr\{WBIC_{\lambda_n} = WBIC_0\} \to 1,$$

where $WBIC_{\lambda_n} = WBIC(\lambda_n)$.

Proof. Let $\hat{\boldsymbol{\beta}}_{\lambda_n}$ be the penalized QIF estimator of $\boldsymbol{\beta}$ corresponding to λ_n. By Theorem 5 and Lemma 5 and 6, $\hat{\boldsymbol{\beta}}_{\lambda_n}$ satisfies the oracle properties under conditions (A) - (C) and (1) - (2). That is, with probability one, $\hat{\beta}_{\lambda_n j} = 0$ for $j \in \mathcal{N}$ and $\hat{\beta}_{\lambda_n j}$ solve

$$n^{-1/2}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n})}{\partial \beta_j} + n^{-1/2}k_n\dot{p}_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|)\text{sgn}(\hat{\beta}_{\lambda_n j}) = 0, \text{ for } j \in \mathcal{A}. \tag{55}$$

and $df(\hat{\boldsymbol{\beta}}_{\lambda_n}) = d_a$. We can show that

$$n^{-1/2}k_n\dot{p}_{\lambda_n}(\hat{\beta}_{\lambda_n j})\text{sgn}(\hat{\beta}_{\lambda_n j}) = k_n n^{-1}n^{1/2}\dot{p}_{\lambda_n}(\hat{\beta}_{\lambda_n j})\text{sgn}(\hat{\beta}_{\lambda_n j}) = o_p(1)$$

since $n^{1/2}\dot{p}_{\lambda_n}(\hat{\beta}_{\lambda_n j}) = o_p(1)$ for $j \in \mathcal{A}$ with probability tending one by condition 1 and $k_n = O(n)$. Also,

$$n^{-1/2}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n})}{\partial \beta_j} = n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \beta_j} + o_p(1) = O_p(1).$$

Therefore, with the probability tending to one, $\hat{\beta}_{\lambda_n j}$ for $j \in \mathcal{A}$ solves estimating equations under the unpenalized QIF,

$$n^{-1/2}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n})}{\partial \beta_j} = 0, \text{ for } j \in \mathcal{A}. \tag{56}$$

Thus, $\Pr\{\hat{\beta}_{\lambda_n j} = \hat{\beta}_{T j}, j \in \mathcal{A}\} \to 1$ as $n \to \infty$. Since Q_n is continuous, with probability tending to one,

$$Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) = Q_n(\hat{\boldsymbol{\beta}}_T), \tag{57}$$

as $n \to \infty$.

Therefore, with probability tending to one,

$$\begin{aligned} \text{WBIC}_{\lambda_n} &= Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) + \log(n) d_a \\ &= Q_n(\hat{\boldsymbol{\beta}}_T) + \log(n) d_a \\ &= \text{WBIC}_0, \end{aligned}$$

as $n \to \infty$. □

Lemma 8. *If conditions (A) - (D) and (1) - (2) hold,*

$$P(\inf_{\lambda \in \Theta_- \cup \Theta_+} WBIC_\lambda > WBIC_{\lambda_n}) \to 1.$$

Proof. (Case 1: **Underfitted model**) First, with probability one, we have

$$\begin{aligned} \inf_{\lambda \in \Theta_-} \text{WBIC}(\lambda) - \text{WBIC}_{\lambda_n} &= \inf_{\lambda \in \Theta_-} \text{WBIC}_\lambda - \text{WBIC}_0 + \text{WBIC}_0 - \text{WBIC}_{\lambda_n} \\ &> \inf_{\lambda \in \Theta_-} Q_n(\hat{\boldsymbol{\beta}}_\lambda) - Q_n(\hat{\boldsymbol{\beta}}_T) - \log(n) df(\hat{\boldsymbol{\beta}}_T) + o_p(n) \\ &> \min_{c \in C_-} Q_n(\boldsymbol{\beta}_c) - Q_n(\hat{\boldsymbol{\beta}}_T) - \log(n) df(\hat{\boldsymbol{\beta}}_T) + o_p(n), \end{aligned}$$

where $\boldsymbol{\beta}_c$ is any coefficient vector for $c \in C_-$.

First, we have $Q_n(\hat{\boldsymbol{\beta}}_T) = O_p(1)$ as $Q_n(\hat{\boldsymbol{\beta}}_T)$ is the goodness of it test statistic following a distribution of a linear combination of χ^2_1 (see Section 4.3). Second, $\log(n) df(\hat{\boldsymbol{\beta}}_T) = o_p(n)$. Third, for any underfitted model $c \in C_-$, there exists some $\epsilon >$ such that $\|\boldsymbol{\beta}_c - \boldsymbol{\beta}_0\| > \epsilon$. By assumption C4 in section 3.4, $\boldsymbol{g}_n(\boldsymbol{\beta}_c) = E[\boldsymbol{g}_n(\boldsymbol{\beta}_c)] + o_p(1) = O_p(1)$, which leads to $Q_n(\boldsymbol{\beta}_c) = n \boldsymbol{g}_n^T(\boldsymbol{\beta}_c) \boldsymbol{A}_n^{-1}(\boldsymbol{\beta}_c) \boldsymbol{g}_n(\boldsymbol{\beta}_c) = O_p(n)$. Then, we

have

$$\Pr\{\inf_{\lambda\in\Theta_-} \text{WBIC}_\lambda - \text{WBIC}_{\lambda_n} > 0\} \to 1.$$

(Case 2: **Overfitted model**)

$$\begin{aligned}
\inf_{\lambda\in\Theta_+} \text{WBIC}_\lambda - \text{WBIC}_{\lambda_n} &\geq \inf_{\lambda\in\Theta_+} Q_n(\hat{\boldsymbol{\beta}}_\lambda) - Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) + [\min_{\lambda\in\Theta_+} df(\hat{\boldsymbol{\beta}}_\lambda) - df(\hat{\boldsymbol{\beta}}_{\lambda_n})]\log(n) \\
&\geq \min_{c\in C_+} Q_n(\hat{\boldsymbol{\beta}}_c) - Q_n(\hat{\boldsymbol{\beta}}_T) + Q_n(\hat{\boldsymbol{\beta}}_T) - Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) + \log(n) \\
&= \min_{c\in C_+}[Q_n(\hat{\boldsymbol{\beta}}_c) - Q_n(\hat{\boldsymbol{\beta}}_T)] + Q_n(\hat{\boldsymbol{\beta}}_T) - Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) + \log(n),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_c$ is the unpenalized QIF estimate for an overfitted model such that $c \in C_+$.

As we have shown in Section 4.2.2, for an overfitted model c, $Q_n(\hat{\boldsymbol{\beta}}_T) - Q_n(\hat{\boldsymbol{\beta}}_c)$ follows the same distribution as a linear combination of χ^2_1. Then, $Q_n(\hat{\boldsymbol{\beta}}_T) - Q_n(\hat{\boldsymbol{\beta}}_c) = O_p(1)$ and $\min_{c\in C_+}[Q_n(\hat{\boldsymbol{\beta}}_c) - Q_n(\hat{\boldsymbol{\beta}}_T)] = O_p(1)$. Also, $Q_n(\hat{\boldsymbol{\beta}}_T) - Q_n(\hat{\boldsymbol{\beta}}_{\lambda_n}) = O_p(1)$. Therefore, as $n \to \infty$, $\log(n)$ dominates other two terms and

$$\Pr(\inf_{\lambda\in\Theta_+} \text{WBIC}_\lambda > \text{WBIC}_{\lambda_n}) \to 1.$$

□

Lemma (7) and (8) indicate that, with probability tending to 1, the WBIC(λ) procedure selects λ_n that identifies the model converging to the true model as $n \to \infty$.

6.5 Simulation

In this section, we conduct simulation studies to show the numerical performance of the proposed variable selection method based on weighted QIF.

6.5.1 Continuous outcome

First, we generate a population of N = 10000 individuals from the following model

$$y_{it} = (\beta_1 + v_i)x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \cdots + \beta_7 x_{it}^{(7)} + \epsilon_{it}, i = 1, 2, ..., 10000, t = 1, 2, ..., 5$$

where $\boldsymbol{\beta} = (3.5, 1.5, 0, 0, 2, 0, 0)^T$. The covariates are generated independently from a multivariate normal distribution $N(0, \boldsymbol{\Sigma}_x)$ where $\boldsymbol{\Sigma}_x = (1 - \rho_x)\mathbf{I} + \rho_x \mathbf{1}\mathbf{1}^T$ with $\rho_x = 0.5$, v_i is the random slope independently generated from $N(0, 1)$, and ϵ_{ij} are random errors independently generated from $N(0, 1)$. The observations from the same individual are correlated because they have the common v_i.

Second, we select $K = 1000$ Monte Carlo (MC) samples of size n under a given sampling design. We consider two designs, SRS and Rao-Sampford PPS, and two sample sizes, n =100 and 200 for each design. For the PPS design, the size variable z_i is defined as $z_i = \exp(0.5v_i)$; therefore, the sampling design is informative.

Third, we select the variables using the procedures based on both weighted and unweighted QIF for each MC sample. Three working correlations; independence (IN), first-order autoregressive (AR1) and exchangeable (EX) are considered. We produce regression coefficient estimates and associated variance estimates. For the SRS design, we consider only the "sandwich" estimates; while for the PPS design, we also consider the bootstrap variance estimates.

The performance of the selection procedures are assessed as follows. First, we present the percentage of exact selection among K simulations where only three covariates with non-zero coefficients are selected (C), the percentage of over-selection where at least four covariates including three covariates with non-zero coefficients are

selected (O), and the percentage of under-selection where at least one of the three covariates with non-zero coefficients is not selected(U). We use mean squared errors (MSE) to assess the performance of estimation after variable selection, replaced by its Monte Carlo estimate

$$MSE(\hat{\boldsymbol{\beta}}) \approx K^{-1} \sum_{k=1}^{K} (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}),$$

where k is the index of MC samples. Second, we report the relative bias for the estimate of non-zero coefficients that is calculated as $RB(\hat{\beta}_j) = (E[\hat{\beta}_j] - \beta_j)/\beta_j \times 100\%$ where $E[\hat{\beta}_j]$ is evaluated as $E[\hat{\beta}_j] \approx K^{-1} \sum_{k=1}^{K} \hat{\beta}_j^{(k)}$. Third, we assess the performance of two variance estimation methods, linearization and bootstrap. Following Fan and Li (2001), we evaluate standard error using median absolute deviation divided by 0.6745 (SD) of the coefficient estimates from K simulations, the median of the K "sandwich" estimated SD's (SD_m^L or SD_m^B), and median absolute deviation error of the K estimated standard errors divided by 0.6745 (SD_{mad}^L or SD_{mad}^B). The one-step bootstrap variance estimator is considered for the PPS design.

Tables (13) -(15) present the variable selection results for the continuous outcome under SRS design. Here "PQIF" represents penalized QIF, "ORACLE" represents the true model that always identifies the zero coefficients and the nonzero coefficients, and "PGEE" represents the penalized GEE. As expected, the unweighted methods are as good as weighted ones since the SRS is non-informative. All methods work well. They select the non-zero predictor and exclude the zero predictors. The penalized QIF performs slightly better than the penalized GEE in terms of the number of models correctly specified and model error and is as good as the oracle methods. For

the most part, the performance of model selection improves as sample size increases.

Table (14) shows that all methods produce consistent estimates under SRS design. We see that all the RBs of PQIF and ORACLE are less than 1%; they are usually smaller than the RBs of PGEE. Table (15) gives standard deviation estimates for PQIF estimates using linearization and bootstrap. The two methods perform similarly and produce reasonable standard deviation estimates.

Table (16) - (18) presents numerical results under PPS sampling. The unweighted methods perform slightly better than weighted methods in terms of the percentage of times the true model is exactly selected, however they have much larger MSEs. The penalized QIF performs much better than PGEE, especially when the sample size is small. The unweighted methods always yield coefficient estimates for β_1 with relative bias of 8%, which does not decrease as sample size increases. On the other hand, the relative bias of β_1 for weighted methods are less than 2%, and decrease as sample size increases. This confirms that when selecting variables based on survey data under informative sampling, one needs to take account of sampling design. Again, linearization and bootstrap produce similar, and reasonable standard deviation estimates for the weighted penalized QIF.

Table 13: Variable selection for continuous outcome with $p = 7$ under SRS design

SIZE	METHOD	IN				EX				AR			
		C	**O**	**U**	**MSE**	**C**	**O**	**U**	**MSE**	**C**	**O**	**U**	**MSE**
						Unweighted/Weighted							
100	**PQIF**	100.0	0.0	0.0	0.021	100.0	0.0	0.0	0.010	100.0	0.0	0.0	0.012
	ORACLE	100.0	0.0	0.0	0.021	100.0	0.0	0.0	0.010	100.0	0.0	0.0	0.012
	PGEE	94.7	5.3	0.0	0.027	98.5	1.5	0.0	0.017	97.3	2.7	0.0	0.021
200	**PQIF**	100.0	0.0	0.0	0.010	100.0	0.0	0.0	0.047	100.0	0.0	0.0	0.005
	ORACLE	100.0	0.0	0.0	0.010	100.0	0.0	0.0	0.005	100.0	0.0	0.0	0.005
	PGEE	97.5	2.5	0.0	0.011	99.8	0.2	0.0	0.008	99.0	1.0	0.0	0.009

Table 14: Relative bias of regression coefficient estimates for continuous outcome under SRS

SIZE	METHOD	IN			EX			AR		
		β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5
					Unweighted/Weighted					
100	**PQIF**	0.10%	-0.09%	-0.37%	0.10%	0.65%	0.13%	0.21%	0.31%	0.04%
	ORACLE	0.10%	-0.08%	-0.37%	0.10%	0.64%	0.12%	0.19%	0.31%	0.04%
	PGEE	-0.02%	-2.70%	-1.05%	0.04%	-1.22%	-0.38%	0.06%	-1.85%	-0.66%
200	**PQIF**	0.00%	-0.07%	-0.18%	-0.11%	0.49%	0.18%	0.05%	0.20%	0.07%
	ORACLE	0.00%	-0.06%	-0.18%	-0.11%	0.49%	0.18%	0.05%	0.20%	0.07%
	PGEE	-0.03%	-0.90%	-0.30%	-0.06%	-0.47%	0.10%	-0.02%	-0.55%	-0.18%

Table 15: Standard deviations of PQIF estimators for continuous outcome under SRS design

SIZE	CORR	β_1			β_2			β_5		
		SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$
100	**IN**	0.097	0.086(0.010)	0.093(0.012)	0.079	0.070(0.009)	0.076(0.010)	0.074	0.069(0.008)	0.073(0.010)
	EX	0.076	0.063(0.006)	0.063 (0.007)	0.046	0.040(0.004)	0.040(0.005)	0.046	0.039(0.004)	0.040(0.004)
	AR	0.083	0.066(0.007)	0.066(0.009)	0.055	0.045(0.004)	0.045(0.005)	0.052	0.044(0.004)	0.044(0.005)
200	**IN**	0.072	0.063(0.006)	0.065(0.007)	0.056	0.051(0.004)	0.053(0.005)	0.053	0.050(0.004)	0.052(0.005)
	EX	0.052	0.046(0.003)	0.046(0.004)	0.033	0.029(0.002)	0.029(0.002)	0.031	0.029(0.002)	0.029(0.002)
	AR	0.055	0.049(0.004)	0.049(0.004)	0.036	0.033(0.002)	0.033(0.003)	0.035	0.032(0.002)	0.032(0.003)

Table 16: Variable selection for continuous outcome with $p = 7$ under PPS design

SIZE	METHOD	IN				EX				AR			
		C	O	U	MSE	C	O	U	MSE	C	O	U	MSE
Unweighted													
100	**PQIF**	100.0	0.0	0.0	0.078	100.0	0.0	0.0	0.078	100.0	0.0	0.0	0.080
	ORACLE	100.0	0.0	0.0	0.077	100.0	0.0	0.0	0.078	100.0	0.0	0.0	0.080
	PGEE	87.9	12.1	0.0	0.078	93.0	7.0	0.0	0.078	91.0	9.0	0.0	0.079
200	**PQIF**	100.0	0.0	0.0	0.069	100.0	0.0	0.0	0.070	100.0	0.0	0.0	0.070
	ORACLE	100.0	0.0	0.0	0.069	100.0	0.0	0.0	0.070	100.0	0.0	0.0	0.070
	PGEE	97.8	2.2	0.0	0.069	99.3	0.7	0.0	0.070	98.8	1.2	0.0	0.070
Weighted													
100	**PQIF**	99.0	1.0	0.0	0.017	99.4	0.6	0.0	0.018	99.5	0.5	0.0	0.019
	ORACLE	100.0	0.0	0.0	0.016	100.0	0.0	0.0	0.018	100.0	0.0	0.0	0.018
	PGEE	78.2	21.8	0.0	0.018	85.7	14.3	0.0	0.017	81.3	18.7	0.0	0.017
200	**PQIF**	98.2	1.8	0.0	0.009	98.8	1.2	0.0	0.009	98.6	1.4	0.0	0.010
	ORACLE	100.0	0.0	0.0	0.009	100.0	0.0	0.0	0.009	100.0	0.0	0.0	0.009
	PGEE	92.6	7.4	0.0	0.009	96.2	3.8	0.0	0.009	93.9	6.1	0.0	0.009

Table 17: Relative bias of regression coefficient estimates for continuous outcome under PPS design

SIZE	METHOD	IN			EX			AR		
		β_1	β_2	β_5	β_1	β_2	β_5	β_1	β_2	β_5
Unweighted										
100	**PQIF**	8.5%	0.0%	0.1%	8.5%	0.2%	0.1%	8.6%	0.0%	0.2%
	ORACLE	8.5%	0.1%	0.1%	8.5%	0.2%	0.1%	8.6%	0.0%	0.2%
	PGEE	8.5%	-1.5%	-0.4%	8.5%	-1.4%	0.2%	8.6%	-1.4%	-0.2%
200	**SCAD**	8.4%	0.1%	0.1%	8.4%	0.2%	0.2%	8.4%	0.1%	0.1%
	ORACLE	8.4%	0.1%	0.1%	8.4%	0.2%	0.2%	8.4%	0.1%	0.1%
	PGEE	8.4%	-0.9%	0.1%	8.4%	-0.7%	-0.2%	8.4%	-0.7%	0.1%
Weighted										
100	**PQIF**	0.8%	0.3%	-0.1%	1.3%	0.4%	0.0%	1.4%	0.3%	0.1%
	ORACLE	0.3%	0.4%	-0.1%	1.3%	0.4%	0.0%	1.4%	0.3%	0.1%
	PGEE	0.3%	-1.4%	-0.1%	0.3%	-1.1%	-0.0%	0.4%	-1.2%	0.0%
200	**SCAD**	0.5%	0.4%	0.0%	0.8%	0.5%	0.1%	0.9%	0.4%	0.0%
	ORACLE	0.1%	0.4%	0.0%	0.8%	0.5%	0.1%	0.9%	0.4%	0.0%
	PGEE	0.1%	-0.6%	-0.0%	0.1%	-0.4%	0.0%	0.1%	-0.5%	0.0%

Table 18: Standard deviations of PQIF estimators for continuous outcome under PPS design

SIZE	CORR	β_1			β_2			β_5		
		SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^L(SD_{mad}^L)$	$SD_m^B(SD_{mad}^B)$
100	**IN**	0.090	0.088(0.018)	0.079(0.013)	0.066	0.062(0.009)	0.056(0.008)	0.063	0.060 (0.009)	0.055(0.008)
	EX	0.094	0.072(0.012)	0.071(0.010)	0.064	0.053(0.007)	0.051(0.007)	0.060	0.051(0.007)	0.050(0.007)
	AR	0.087	0.073(0.012)	0.071(0.010)	0.069	0.053(0.007)	0.052(0.007)	0.061	0.051(0.007)	0.051(0.007)
200	**IN**	0.069	0.063(0.011)	0.059(0.008)	0.048	0.043(0.005)	0.041(0.005)	0.045	0.043(0.005)	0.041(0.004)
	EX	0.067	0.055(0.007)	0.055(0.006)	0.047	0.039(0.004)	0.039(0.004)	0.046	0.039(0.004)	0.039(0.004)
	AR	0.067	0.056(0.008)	0.056(0.007)	0.046	0.039(0.004)	0.039(0.004)	0.046	0.038(0.004)	0.039(0.004)

6.5.2 Binary outcome

We generated a finite population with correlated binary response y from the following marginal logit model

$$\text{logit}(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta} \text{ for } i = 1, ..., 30000, \text{ for } j = 1, ..., 5$$

where $\mu_{it} = \Pr(y_{it} = 1)$, $\boldsymbol{x}_{it} = (x_{it}^{(1)}, ..., x_{it}^{(10)})^T$ and $\boldsymbol{\beta} = (0.8, -0.7, -0.6, 0, 0, 0, 0, 0, 0, 0)^T$. Each covariate $x_{it}^{(k)}$ was generated independently from UN(0,0.8) for $k = 1, ..., 10$. Wicklin (2013) developed a SAS program *RandMVBinary* that enables one to simulate correlated multivariate binary data according to the algorithm of Emrich and Piedmonte (1991). We used this program to generate the correlated binary responses with an exchangeable correlation structure (EX) with correlation coefficient $\alpha = 0.4$.

Define a size variable $Z_i = \sum_t y_{it} + 1$. We used PROC SURVEYSELECT in SAS 9.3 to select finite samples under PPS with replacement. We considered two sample sizes $n = 300$ and 500. From the finite population, we generated K=500 Monte Carlo (MC) samples for each sample size.

A bootstrap method was considered for variance estimation. First, we replicate those units selected more than once and draw a bootstrap sample of size $n - 1$ with replacement and equal probabilities. Bootstrap weights were calculated using the rescaling formula given as

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

where $t_i^{(b)}$ is the number of hits of unit i in the b^{th} bootstrap sample.

Simulation results are presented in Table 19 - 21. Here, UNWGT represents unweighted QIF with SCAD penalty , PQIF represents the weighted QIF with SCAD penalty, and ORACLE represents the weighted QIF under the true model that identifies three nonzero coefficients and seven zero coefficients. Table (19) shows that the unweighted PQIF is not capable of variable selection; less than 10% of the time is the true model selected exactly. The PQIF does perform better than PGEE under AR and EX, but worse under IN.

We calculated absolute relative bias due to negative values of β, namely $(E[\hat{\beta}_i] - \beta_i)/|\beta_i|$ for $i = 1, 2, 3$. As shown in Table (20), the unweighted PQIF yields biased results. Generally, the RBs of PQIF estimates are less that those of PGEE. The RBs decrease for weighted methods but not for unweighted methods. From Table (21), we see that the one-step bootstrap method yields reasonable standard deviation estimate for binary data.

Table 19: Variable selection for binary outcome with $p = 10$

SIZE	METHOD	IN				EX				AR			
		C	O	U	MSE	C	O	U	MSE	C	O	U	MSE
300	**UNWGT**	1.8	19.0	79.2	0.648	5.6	34.2	60.2	0.538	2.4	16.0	71.6	0.610
	PQIF	66.6	9.6	23.8	0.245	80.8	11.2	8.0	0.129	75.0	11.4	13.6	0.163
	ORACLE	100.0	0.0	0.0	0.098	100.0	0.0	0.0	0.072	100.0	0.0	0.0	0.080
	PGEE	69.2	10.8	20.0	0.225	73.0	10.6	16.4	0.172	72.2	10.4	17.4	0.181
500	**UNWGT**	2.4	48.4	49.2	0.486	7.0	66.6	26.4	0.384	5.4	59.6	35.0	0.435
	PQIF	83.8	12.8	3.4	0.093	91.2	8.6	0.2	0.049	87.8	11.0	1.2	0.065
	ORACLE	100.0	0.0	0.0	0.059	100.0	0.0	0.0	0.039	100.0	0.0	0.0	0.046
	PGEE	84.0	13.6	2.4	0.087	88.2	10.0	1.8	0.054	87.0	10.2	1.8	0.062

Table 20: Absolute relative bias (ARB) for regression coefficients for binary outcome after variable selection

SIZE	METHOD	IN			EX			AR		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
300	UNWGT	23.8%	40.0%	65.0%	28.8%	30.0%	55.0%	28.8%	37.1%	61.7%
	PQIF	6.3%	1.4%	13.3%	1.3%	1.4%	6.7%	2.5%	0.0%	8.3%
	ORACLE	1.3%	1.4%	5.0%	1.3%	1.4%	3.3%	0.0%	0.0%	3.3%
	PGEE	6.3%	0.0%	11.7%	7.5%	1.4%	13.3%	6.3%	1.4%	13.3%
500	UNWGT	25.8%	29.7%	52.9%	28.4%	22.9%	41.4%	28.6%	26.6%	46.5%
	PQIF	1.7%	2.9%	5.5%	1.7%	2.3%	4.0%	1.4%	2.0%	3.9%
	ORACLE	1.5%	3.2%	4.2%	1.8%	2.4%	3.9%	1.5%	2.0%	3.2%
	PGEE	2.6%	2.6%	5.2%	3.9%	1.2%	5.8%	3.5%	0.9%	4.5%

Table 21: Standard deviations of PQIF estimators for binary outcome after model selection

CORR	SIZE	β_1		β_2		β_3	
		SD	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^B(SD_{mad}^B)$	SD	$SD_m^B(SD_{mad}^B)$
AR	300	0.181	0.172(0.018)	0.170	0.170(0.016)	0.162	0.165(0.016)
	500	0.124	0.134(0.011)	0.137	0.133(0.011)	0.125	0.131(0.011)
EX	300	0.172	0.163(0.016)	0.152	0.161(0.015)	0.142	0.156(0.015)
	500	0.112	0.127(0.011)	0.125	0.125(0.010)	0.099	0.123(0.010)
IN	300	0.192	0.202(0.020)	0.187	0.202(0.021)	0.193	0.193(0.021)
	500	0.141	0.156(0.011)	0.147	0.157(0.012)	0.136	0.154(0.012)

Chapter 7

Composite Likelihood for Complex Survey Data

7.1 Introduction

Large-scale sample surveys often use multi-stage sampling designs. At the first stage, primary sampling units (PSUs) or clusters are selected randomly from a population of PSUs. At the next stage, secondary sampling units (SSUs) are selected within the sampled PSUs, and so on until the last stage where elements of interest are selected. Stratification and unequal selection probabilities may be used at any stage of sampling. This design has several advantages. First, it only requires a list of clusters and lists of elements within sampled clusters, instead of a full list of all elements. Sometimes it is difficult or very expensive to create and maintain an element frame. Second, directly sampling elements may result in a widely spread sample. If personal interview is required, data collection can be costly due to high

travel expenses; also, field work can be difficult to manage, resulting in high non-response rates and severe measurement errors (Särndal, Swensson and Wretman, 1992).

Data collected from multi-stage sample surveys is hierarchically structured. Multilevel models are usually used to model a hierarchical population, taking account of the within-cluster correlations. Maximum Likelihood (ML) is the standard procedure for multilevel model analysis; it produces consistent parameter estimates. Restricted maximum likelihood (REML) is a modification of ML where parameters have been estimated and may reduce the bias of ML for small sample sizes. Numerical algorithms are needed to compute the ML or REML estimates, for example, the iterative generalized least squares method (IGLS) is used for two-level linear models (Goldstein, 1986), and different approximations to log-likelihood for general multilevel models (Pinheiro and Bates, 1995).

When making inference about model parameters from sample survey data, one needs to take account of the sampling design. If the design is informative in the sense that the sampling probabilities are correlated with the responses even after conditioning on covariates, ignoring the sampling design may lead to biased estimates and result in erroneous conclusions. Survey weighted methods are often used to account for the sampling design under a design-model framework (Rubin-Bleuer and Schiopu-Kratina, 2005). Under this framework, the finite population is assumed to be a realization of the underlying model and the survey sample is viewed as a sub-sample of the finite population. The finite population parameter is defined as a consistent estimator of the model parameter if the whole population is observable. The survey

sample estimator is a consistent estimator of the census or finite population parameter; in turn, it is a consistent estimator of the model parameter. Binder (1983) and Skinner (1989) developed the pseudo maximum likelihood (PML) method for single-level models. The PML estimates are found by maximizing a survey weighted log-likelihood function. Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) proposed the PML method for multilevel models (MPML). However, the MPML procedure is much more complicated due to the loss of independence. First, survey weights need to be scaled to obtain consistent estimates for model parameters under an informative sampling design. Different scaling methods were proposed; their performance depends on several factors such as cluster size, intra-cluster correlation and invariance. Asparouhov (2006) compared their performance under different setups. No single method outperformed the others in all situations. Second, for linear models, consistency for variance/covariance estimators requires large within-cluster sample sizes, which may not be true in practice. For nonlinear models, the estimation of regression parameters may depend on the estimation of the variance/covariance components in the model. As a result, the estimation of regression parameters may also require large within-cluster sample size (Rabe-Hesketh and Skrondal, 2006).

Alternatively, Rao, Verret and Hidiroglou (2013) considered inference for two-level linear regression models from a two-stage survey sample, and the case where the model hierarchy matches sampling hierarchy. They proposed a unified estimating procedure based on a weighted pairwise log composite likelihood (WCL). The resulting estimators are design-model consistent even for small within cluster sizes. Yi, Rao, and Li (2016) proved consistency of the WCL estimator and extended the method to nonlinear two-level models.

In this chapter, we consider testing hypotheses about model parameters for two-level models based on weighted pairwise log-likelihood. Our work is a natural extension of Rao, Verret and Hidiroglou (2013) and Yi, Rao and Li (2016). The rest of the chapter is organized as follows. In Section 2, we introduce the two-stage sampling design and two-level models. In Section 3, we review existing methods for multilevel survey data analysis and introduce the WCL method. In Section 4, we study the asymptotic normality of the WCL estimator and propose test statistics based on the WCL. In Section 5, we conduct simulation studies to assess the performance of the proposed test statistics.

7.2 Two-stage sampling and two-level models

Consider a two-level finite population, and two-stage sampling design. Assume that sampling hierarchy matches model hierarchy. We adopt the notation from Särndal, Swensson and Wretman (1992). Suppose the finite population U is composed of N PSUs (level-2 units, or clusters) and in the ith PSU there are M_i SSUs (level-1 units, or elements). The total number of SSUs in the finite population is $M = \sum_{i=1}^{N} M_i$. Let y_{ij} be the response value and $\mathbf{x}_{ij}$ the covariate vector associated with the jth SSU in the ith PSU. Let $\mathcal{F} = \{(Y_{ij}, \mathbf{x}_{ij}), i = 1, \cdots, N, j = 1, ..., M_i\}$ be the set of vectors for the finite population.

Let $U^{(2)} = \{1, ..., N\}$ denote the index set of PSUs and $U_i^{(1)}$ denote the index set of SSUs in the i^{th} PSU for $i \in U^{(2)}$. A survey sample is selected under a general two-stage sampling design.

First stage: a sample $S^{(2)}$ of n PSUs is drawn from $U^{(2)}$ according to the

first-stage design;

Second stage: for every $i \in S^{(2)}$, a sample $S_i^{(1)}$ of size m_i is drawn from $U_i^{(1)}$ according to the second-stage design.

Thus, the total sample size is $m = \sum_{i \in S^{(2)}} m_i$. Denote the first and second order inclusion probabilities for stage 1 by

$$\pi_i = \Pr(i \in S^{(2)}) \text{ for } i \in U^{(2)},$$

$$\text{and } \pi_{ii'} = \Pr(i, i' \in S^{(2)}) \text{ for } i, i' \in U^{(2)}, i \neq i',$$

and for stage 2 by

$$\pi_{k|i} = \Pr(k \in S_i^{(1)} | i \in S^{(2)}) \text{ for } k \in U_i^{(1)}$$

$$\text{and } \pi_{jk|i} = \Pr(j, k \in S_i^{(1)} | i \in S^{(2)}) \text{ for } j, k \in U_i^{(1)}, j \neq k.$$

Suppose the values of $(Y_{ij}, \mathbf{x}_{ij})$ are generated from a general two-level super-population model with parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ given by

$$\begin{aligned} \textit{Level 1: } Y_{ij} &\sim f(y_{ij}; \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1), \\ \textit{Level 2: } \mathbf{v}_i &\sim f(\mathbf{v}_i; \boldsymbol{\theta}_2), \text{ for } i = 1, ..., N \text{ and } j = 1, ..., M_i, \end{aligned} \tag{58}$$

where $\mathbf{v}_i$ is the vector of random effects associated with the ith PSU, $\boldsymbol{\theta}_1$ is the parameter of level-1 model and $\boldsymbol{\theta}_2$ is the parameter of level-2 model. Assume that $\mathbf{v}_i$ and $\mathbf{v}_{i'}$ for $i \neq i'$ are independent. Then, responses among different PSUs are

independent, i.e. $Y_{ik} \perp Y_{i'k'}$ for $i \neq i'$. Responses Y_{ij} and Y_{ik} within PSU i are correlated by sharing $\mathbf{v}_i$, but conditionally independent if $\mathbf{v}_i$ is given.

7.3 General review of methods for multilevel models for survey data

Suppose the entire finite population U is observed. The parameter estimator for model (58) can then be found by maximizing the finite population (or "census") log-likelihood function

$$l(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i\in U^{(2)}} \log \int \exp\{ \sum_{j\in U_i^{(1)}} \log f(y_{ij}|\mathbf{x}_{ij},\mathbf{v}_i,\boldsymbol{\theta}_1)\} f(\mathbf{v}_i;\boldsymbol{\theta}_2) d\mathbf{v}_i.$$

When only survey sample data are available, Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) extended the pseudo maximum likelihood (PML) method from single-level to multilevel models, and proposed the multilevel pseudo maximum likelihood (MPML) method. The MPML estimate is found by maximizing the pseudo log-likelihood function defined as

$$pl(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i\in S^{(2)}} w_i^* \log \int \exp\{ \sum_{j\in S_i^{(1)}} w_{j|i}^* \log f(y_{ij}|\mathbf{x}_{ij},\mathbf{v}_i,\boldsymbol{\theta}_1)\} f(\mathbf{v}_i;\boldsymbol{\theta}_2) d\mathbf{v}_i,$$

where $w_{j|i}^* = w_{j|i} \cdot k_{1i}$ with $w_{j|i} = 1/\pi_{j|i}$, and $w_i^* = w_i \cdot k_{2i}$ with $w_i = 1/\pi_i$. Here k_{1i} and k_{2i} are scaling factors. The MPML estimator is biased if only design weights are used. Different methods of scaling the design weights have been proposed to reduce the bias. For example, (method A) $k_{1i} = m_i / \sum_{j\in S_i^{(1)}} w_{j|i}$ and $k_{2i} = 1$, or

(method A1) $k_{1i} = m_i / \sum_{j \in S_i^{(1)}} w_{j|i}$ and $k_{2i} = 1/k_{1i}$, see Asparouhov (2006). As seen in the above formula, the weights for both sampling stages are needed. Note that the MPML requires knowledge of the sampling weights for both stages, which may not be available. In this work, we assume that weights for both sampling stages are provided.

Composite likelihood is a product of marginal or conditional likelihoods (Besag 1974, Lindsay 1988), and pairwise likelihood is a special case of composite likelihood. The "census" or finite population pairwise log-likelihood under the assumed two-level model (58) is given by

$$cl(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i \in U^{(2)}} \sum_{j>k \in U_i^{(1)}} \log f(y_{ij}, y_{ik}; \boldsymbol{\theta}),$$

where $f(y_{ij}, y_{ik}; \boldsymbol{\theta})$ is the marginal probability density function of Y_{ij} and Y_{ik}. "Census" parameter $\boldsymbol{\theta}_N \in \Omega$ is the maximiser of $cl(\boldsymbol{\theta}; \boldsymbol{y})$ and can be found by solving the "census" estimating equations (EE)

$$\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv \sum_{i \in U^{(2)}} \boldsymbol{U}_i(\boldsymbol{\theta}) = 0,$$

where $\boldsymbol{U}_i(\boldsymbol{\theta}) = \sum_{j>k \in U_i^{(1)}} \boldsymbol{u}_{ijk}(\boldsymbol{\theta})$ and $\boldsymbol{u}_{ijk}(\boldsymbol{\theta}) = \partial \log f(y_{ij}, y_{ik}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

The weighted pairwise composite log-likelihood function for survey sample data is given as

$$wcl(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i \in S^{(2)}} w_i \sum_{j>k \in S_i^{(1)}} w_{jk|i} \log f(y_{ij}, y_{ik}; \boldsymbol{\theta}), \tag{59}$$

where $w_i = 1/\pi_i$ and $w_{jk|i} = 1/\pi_{jk|i}$. The WCL estimates $\hat{\boldsymbol{\theta}} \in \Omega$ maximizing $wcl(\boldsymbol{\theta}; \boldsymbol{y})$

can be found by solving the survey weighted EE

$$\hat{\boldsymbol{U}}(\boldsymbol{\theta};\boldsymbol{y}) \equiv \sum_{i\in S^{(2)}} w_i \hat{\boldsymbol{U}}_i(\boldsymbol{\theta}) = 0, \tag{60}$$

where $\hat{\boldsymbol{U}}_i(\boldsymbol{\theta}) = \sum_{j>k\in S_i^{(1)}} w_{jk|i}\boldsymbol{u}_{ijk}(\boldsymbol{\theta})$. Unlike the MPML, the WCL method alsorequires the joint selection probabilities within clusters $\pi_{jk|i}$. For simple random sampling (SRS) or stratified sampling within clusters, these probabilities may be easily calculated. In the case where $\pi_{jk|i}$ cannot be easily calculated, an approximation method may be used. Haziza, Mecatti, and Rao (2008) reviewed the methods of approximating $\pi_{jk|i}$ by simple functions of $\pi_{j|i}$ and $\pi_{k|i}$ for the Rao-Sampford unequal probability selection approach.

Note that the estimating function $\hat{\boldsymbol{U}}(\boldsymbol{\theta};\boldsymbol{y})$ evaluated at the true parameter value $\boldsymbol{\theta}_0$ has mean zero under joint model ξ and sampling design d random mechanism, i.e. $E_\xi E_d[\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0;\boldsymbol{y})] = 0$. Under regularity conditions, $\hat{\boldsymbol{\theta}}_0$ is a model-design consistent estimator of $\boldsymbol{\theta}$. The consistency requires only a large number of clusters, n, in the sample and does not necessarily need a large number of sampled elements, m_i, within sampled clusters. Yi, Rao and Li (2016) provided the conditions for consistency and also generalized this method to non-linear two-level models.

Theorem 7. *(Yi, Rao and Li, 2016) Let $\boldsymbol{\theta}_0$ be the true parameter value in model (58). If conditions A1 - A8 in Appendix A hold, then*

$$\hat{\boldsymbol{\theta}} \to^p \boldsymbol{\theta}_0,$$

where "p" denotes convergence in probability with respect to joint model ξ and design

d as N and as n increase.

Through Taylor linearization, a sandwich variance estimator of $\hat{\boldsymbol{\theta}}$ was provided as

$$v_L(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{U}}'(\hat{\boldsymbol{\theta}}))^{-1}\boldsymbol{cov}(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))[(\hat{\boldsymbol{U}}'(\hat{\boldsymbol{\theta}}))^{-1}]^T, \tag{61}$$

where $\hat{\boldsymbol{U}}'(\hat{\boldsymbol{\theta}}) = \partial\hat{\boldsymbol{U}}(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$, the first derivative of $\hat{\boldsymbol{U}}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and $\boldsymbol{cov}(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ is a consistent estimator of $\boldsymbol{COV}(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$. To estimate $\boldsymbol{COV}(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$, we decompose it into two parts

$$\boldsymbol{COV}(\hat{\boldsymbol{U}}(\boldsymbol{\theta})) = \boldsymbol{COV}_\xi E_d[\hat{\boldsymbol{U}}(\boldsymbol{\theta})] + E_\xi \boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta})). \tag{62}$$

where E_ξ and $\boldsymbol{COV}_\xi$ represent the expectation and covariance under the model ξ, and E_d and $\boldsymbol{COV}_d$ represent the expectation and covariance under the sampling design d. The first term in (62) represents for the variability due to model, while the second term represents the variability due to sampling design. As discussed by Carrillo, Chen and Wu (2010), if the sampling fraction for the first stage sampling, n/N, is small, we may ignore the first term and approximate $\boldsymbol{COV}(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ by $E_\xi\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$, such that, $\boldsymbol{COV}(\hat{\boldsymbol{U}}(\boldsymbol{\theta})) \approx E_\xi\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$. As $\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ is the sampling variance of $\hat{\boldsymbol{U}}(\boldsymbol{\theta})$ given the finite population $\mathcal{F}$, if we can find an estimator, $\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$, that is asymptotically unbiased for $\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ with respect to the sampling design, then $\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ is also an asymptotically unbiased estimator of $E_\xi[\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta})]$, i.e.,

$$E_{\xi d}[\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))] = E_\xi\{E_d[\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))]\} = E_\xi[\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))].$$

To find $\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$, we write $\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ as the sum of two terms by conditioning

on the stage-1 sample $S^{(2)}$:

$$\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta})) = \boldsymbol{COV}_{PSU} + \boldsymbol{COV}_{SSU}. \tag{63}$$

where $\boldsymbol{COV}_{PSU} = \sum_{i,i' \in U^{(2)}} \Delta_{ii'} \frac{\boldsymbol{U}_i(\boldsymbol{\theta})}{\pi_i} \frac{\boldsymbol{U}_{i'}^T(\boldsymbol{\theta})}{\pi_{i'}}$ with $\Delta_{ii'} = \pi_{ii'} - \pi_i \pi_{i'}$ and $\pi_{ii} = \pi_i$. An estimator of $\boldsymbol{COV}_{PSU}$ is taken as $\boldsymbol{cov}_{PSU} = \sum_{i,i' \in S^{(2)}} \frac{\Delta_{ii'}}{\pi_{ii'}} \frac{\hat{\boldsymbol{U}}_i(\hat{\boldsymbol{\theta}})}{\pi_i} \frac{\hat{\boldsymbol{U}}_{i'}^T(\hat{\boldsymbol{\theta}})}{\pi_{i'}} - \sum_{i \in S^{(2)}} \frac{1}{\pi_i}(\frac{1}{\pi_i} - 1)\boldsymbol{cov}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta}))$ where $\boldsymbol{cov}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta}))$ is an estimator of $E_d[\boldsymbol{COV}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta})|S^{(2)})]$. Assuming invariance and independence (Särndal et al., 1992) for the stage-2 design, we have $\boldsymbol{COV}_{SSU} = \sum_{i \in U^{(2)}} w_i E_d[\boldsymbol{COV}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta})|S^{(2)})]$, which can be estimated as $\boldsymbol{cov}_{SSU} = \sum_{i \in S^{(2)}} w_i^2 \boldsymbol{cov}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta}))$. The estimator of $\boldsymbol{COV}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta}))$ is given by

$$\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta})) = \sum_{i,i' \in S^{(2)}} \frac{\Delta_{ii'}}{\pi_{ii'}} \frac{\hat{\boldsymbol{U}}_i(\hat{\boldsymbol{\theta}})}{\pi_i} \frac{\hat{\boldsymbol{U}}_{i'}^T(\hat{\boldsymbol{\theta}})}{\pi_{i'}} + \sum_{i \in S^{(2)}} \frac{1}{\pi_i} \boldsymbol{cov}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta})).$$

Note that $\boldsymbol{cov}_d(\hat{\boldsymbol{U}}_i(\boldsymbol{\theta}))$ requires the fourth-order selection probabilities within cluster. In practice, the within-cluster fourth-order selection probabilities might be difficult to calculate for a complex design. As a result, Yi, Rao and Li (2016) considered treating stage-1 samples as if they were selected with replacement, which leads to a simple function

$$\boldsymbol{cov}_d(\hat{\boldsymbol{U}}(\boldsymbol{\theta})) = \frac{n}{n-1} \sum_{i \in S^{(2)}} w_i^2 \hat{\boldsymbol{U}}_i(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{U}}_{i'}^T(\hat{\boldsymbol{\theta}}).$$

7.4 Hypothesis testing and confidence interval estimation based on weighted pairwise likelihood

In this section, we examine some analytical properties of the WCL method; in particular, confidence interval estimation and hypothesis testing. First, we establish the asymptotic normality of the WCL estimator $\hat{\boldsymbol{\theta}}$. Second, we construct three classes of tests based on the WCL, the Wald-type test, score test, and likelihood ratio test.

Let $\boldsymbol{\theta}_0$ be the true parameter in model (58). We first list some regularity conditions.

C1 There exists a neighbourhood of $\boldsymbol{\theta}_0$, $N_{\boldsymbol{\theta}_0}$, such that $wcl(\boldsymbol{y};\boldsymbol{\theta})$ is thrice differentiable for all $\boldsymbol{y}$ and $\boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}$.

C2 The second order derivative of $f(y_{ij}, y_{ik};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ satisfies

$$\sup_{i,j,k,p,q} E\left[\frac{\partial^2 \log f(y_{ij}, y_{ik};\boldsymbol{\theta})}{\partial\theta_p \partial\theta_q}\right]^4 < \infty. \tag{64}$$

C3 The third order derivative of $f(y_{ij}, y_{ik};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ is bounded for all $\boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}$,

$$\sup_{i,j,k,l,p,q}\left|\frac{\partial^3 \log f(y_{ij}, y_{ik};\boldsymbol{\theta})}{\partial\theta_l \partial\theta_p \partial\theta_q}\right| < M, \text{ almost surely under } \xi \text{ and } d$$

for all y and $M < \infty$.

Lemma 9. *Let* $\bar{T} = T/N$ *with* $T = \sum_{=1}^{N} T_i$ *and* $T_i = M_i(M_i - 1)/2$. *Assume* $\max_i T_i/\bar{T} < c_1$ *for some* $c_1 > 0$. *Define* $\bar{V} = T^{-1}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i}\|\boldsymbol{V}_{ijk}\|^2$ *where*

$\boldsymbol{V}_{ijk} = \frac{\partial^2 \log f(y_{ij}, y_{ik}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$. *Suppose that* $E[\bar{V}] = O(1)$. *If condition C2 above holds, then* $\bar{V} = O_p(1)$ *where the probability, "p" is taken w.r.t.* ξ.

Proof. Let $V_i = T_i^{-1} \sum_{1 \le j < k \le M_i} \|\boldsymbol{V}_{ijk}\|^2$ and rewrite $\bar{V} = N^{-1} \sum_{i=1}^{N} (\bar{T}^{-1} T_i) V_i$. Note that V_i satisfies

$$\begin{aligned} V_i^2 &= T_i^{-2} \left(\sum_{1 \le j < k \le M_i} \|\boldsymbol{V}_{ijk}\|^2 \right)^2 \\ &\le T_i^{-1} \sum_{1 \le j < k \le M_i} \|\boldsymbol{V}_{ijk}\|^4, \text{ by Jensen's inequality.} \end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned} \Pr\{|\bar{V} - E[\bar{V}]| > \epsilon\} &\le \frac{Var(\bar{V})}{\epsilon^2} \\ &\le \max_i (T_i^2 / \bar{T}^2) \frac{\sum_{i=1}^{N} Var(V_i)}{N^2 \epsilon^2} \\ &\le \max_i (T_i^2 / \bar{T}^2) \frac{\sum_{i=1}^{N} E[V_i^2]}{N^2 \epsilon^2} \\ &\le \max_i (T_i^2 / \bar{T}^2) \frac{\sum_{i=1}^{N} T_i^{-1} \sum_{1 \le j < k \le M_i} E[\|\boldsymbol{V}_{ijk}\|^4]}{N^2 \epsilon^2} \\ &\le \max_i (T_i^2 / \bar{T}^2) \frac{\sup_{i,j,k} E[\|\boldsymbol{V}_{ijk}\|^4]}{N \epsilon^2}. \end{aligned}$$

From Condition 2, $\sup_{i,j,k} E[\|\boldsymbol{V}_{ijk}\|^4] < \infty$. Thus,

$$\Pr\{|\bar{V} - E[\bar{V}]| > \epsilon\} \to 0 \text{ as } N \to \infty,$$

as $\max_i (T_i^2 / \bar{T}^2) < c_1 < \infty$. □

Lemma 10. *Let* $\boldsymbol{V}_{ijk} = \frac{\partial^2 \log f(y_{ij}, y_{ik}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$. *Define* $\boldsymbol{I}_Y(\boldsymbol{\theta}_0) =$

$T^{-1}\sum_{i\in S^{(2)}} w_i \sum_{j<k\in S_i^{(1)}} w_{jk|i}\boldsymbol{V}_{ijk}$, $\boldsymbol{I}_U(\boldsymbol{\theta}_0) = T^{-1}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i}\boldsymbol{V}_{ijk}$ *and* $\boldsymbol{I}_E(\boldsymbol{\theta}_0) = T^{-1}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i} E[\boldsymbol{V}_{ijk}]$. *Assume that there exists a non-stochastic matrix* $I_{\boldsymbol{\theta}_0}$ *such that* $\|I_E(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}_0}\| = o(1)$. *Suppose that the following conditions hold*

1. *The second order derivative of* $f(y_{ij}, y_{ik};\boldsymbol{\theta})$ *w.r.t.* $\boldsymbol{\theta}$ *satisfies*

$$\sup_{i,j,k,r,s} E\left[\frac{\partial^2 \log f(y_{ij}, y_{ik};\boldsymbol{\theta})}{\partial\theta_r\partial\theta_s}\right]^2 < \infty. \tag{65}$$

2. $\max_i T_i/\bar{T} < c_1$ *for some* $c_1 > 0$ *where* $\bar{T} = T/N$.

3. *For any random matrices* V_{ijk} *where* $(r,s)^{th}$ *components* V_{ijk}^{rs} *satisfy* $T^{-1}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i}(V_{ijk}^{rs})^2 = O_\xi(1)$, *we have*

$$\frac{1}{T}\sum_{i\in S^{(2)}} w_i \sum_{j>k\in S_i^{(1)}} w_{jk|i}V_{ijk} - \bar{V} \to^p 0$$

under d *as* $n\to\infty$, *where* $\bar{V} = T^{-1}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i} V_{ijk}$.

Then

$$\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0} = o_p(1)$$

under ξ *and* d, *as* $n\to\infty$.

Proof. Write

$$\|\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0}\| \le \|\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_U(\boldsymbol{\theta}_0)\| + \|\boldsymbol{I}_U(\boldsymbol{\theta}_0) - \boldsymbol{I}_E(\boldsymbol{\theta}_0)\| + \|\boldsymbol{I}_E(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0}\|.$$

By condition 3, the first term $\|\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_U(\boldsymbol{\theta}_0)\| = o_p(1)$ under d.

By definition, the third term is $\|\boldsymbol{I}_E(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0}\| = o(1)$.

We now show that $\|\boldsymbol{I}_U(\boldsymbol{\theta}_0) - \boldsymbol{I}_E(\boldsymbol{\theta}_0)\| = o_p(1)$ under ξ as $N \to \infty$. Let $V_{ijk}^{r,s}$ be the (r,s) component of the $p \times p$ matrix V_{ijk}. Now we show that the (r,s) component of $\boldsymbol{I}_U(\boldsymbol{\theta}_0) - \boldsymbol{I}_E(\boldsymbol{\theta}_0)$ converges to 0 in probability w.r.t. ξ. For any $\epsilon > 0$,

$$
\begin{aligned}
\Pr\{|\frac{1}{T}\sum_{i\in U^{(2)}}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])| > \epsilon\} &= \Pr\{|\frac{1}{N}\sum_{i\in U^{(2)}}\frac{N}{T}\frac{T_i}{T_i}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])| > \epsilon\} \\
&\leq \frac{E\left(\sum_{i\in U^{(2)}}\frac{N}{T}\frac{T_i}{T_i}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])\right)^2}{N^2\epsilon^2} \\
&\leq \max_i(\frac{T_i}{\bar{T}})^2\frac{E\left(\sum_{i\in U^{(2)}}\frac{1}{T_i}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])\right)^2}{N^2\epsilon^2} \\
&\leq c_1^2\frac{\sum_{i\in U^{(2)}}E\left(\frac{1}{T_i}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])\right)^2}{N^2\epsilon^2} \\
&\leq \frac{c_1^2}{N\epsilon^2}\max_{i=1,\ldots,N}E\left(\frac{1}{T_i}\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])\right)^2 \\
&\leq \frac{c_1^2}{N\epsilon^2}\max_{i=1,\ldots,N}\frac{1}{T_i^2}E\left(\sum_{j<k\in U_i^{(1)}}(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])^2\sum_{j<k\in U_i^{(1)}}1\right) \\
&= \frac{c_1^2}{N\epsilon^2}\max_{i=1,\ldots,N}\frac{1}{T_i}\sum_{j<k\in U_i^{(1)}}E(V_{ijk}^{r,s} - E_\xi[V_{ijk}^{r,s}])^2 \\
&\leq \frac{c_1^2}{N\epsilon^2}\max_{i=1,\ldots,N}\frac{1}{T_i}\sum_{j<k\in U_i^{(1)}}E(V_{ijk}^{r,s})^2 \text{ (by condition 1)} \\
&\to 0 \text{ as } N \to \infty.
\end{aligned}
$$

Then, $\|\boldsymbol{I}_U(\boldsymbol{\theta}_0) - \boldsymbol{I}_E(\boldsymbol{\theta}_0)\| = o_p(1)$ w.r.t. ξ.

Thus, we have

$$\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0} = o_p(1)$$

under ξ and d, as $n, N \to \infty$.

□

Lemma 11. *Let* $\boldsymbol{u}_{ijk}(\boldsymbol{\theta}) = \partial \log f(y_{ij}, y_{ik}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$. *Define* $\bar{\boldsymbol{U}}_i = \bar{\boldsymbol{U}}_i(\boldsymbol{\theta}_0) = \bar{T}^{-1} \sum_{j>k \in U_i^{(1)}} \boldsymbol{u}_{ijk}(\boldsymbol{\theta}_0)$ *and* $\hat{\bar{\boldsymbol{U}}}_i = \hat{\bar{\boldsymbol{U}}}_i(\boldsymbol{\theta_0}) = \bar{T}^{-1} \sum_{j>k \in S_i^{(1)}} w_{jk|i} \boldsymbol{u}_{ijk}(\boldsymbol{\theta}_0)$. *Assume that*

1. $\lim_{N,n\to\infty} n/N = f$, *where* $0 \le f \le 1$.
2. $\lim_{N\to\infty} N^{-1} \sum_{i \in U^{(2)}} \boldsymbol{COV}(\bar{\boldsymbol{U}}_i) = \boldsymbol{V}_1$, $nN^{-2} \sum_{i,i' \in U^{(2)}} (\pi_{ii'}/\pi_i\pi_{i'} - 1) \bar{\boldsymbol{U}}_i \bar{\boldsymbol{U}}_{i'}^T \to^p \boldsymbol{V}_2$ *and* $nN^{-2} \sum_{i \in S^{(2)}} w_i^2 \boldsymbol{COV}\left(\hat{\bar{\boldsymbol{U}}}_i | \mathcal{F}, S^{(2)}\right) \to^p \boldsymbol{V}_3$. *Here* $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, $\boldsymbol{V}_3$ *are positive-definite non-stochastic matrices.*
3. $N^{-1/2} \sum_{i \in U^{(2)}} \bar{\boldsymbol{U}}_i \to^d N(0, \boldsymbol{V}_1)$, *as* $N \to \infty$
4. *Given* $\mathcal{F}$, $n^{1/2}N^{-1}\left(\sum_{i \in S^{(2)}} w_i \bar{\boldsymbol{U}}_i - \sum_{i \in U^{(2)}} \bar{\boldsymbol{U}}_i\right) \to^d N(0, \boldsymbol{V}_2)$ *as* $n \to \infty$,
5. *Given* $\mathcal{F}$ *and* $S^{(2)}$, $n^{1/2}N^{-1} \sum_{i \in S^{(2)}} w_i \left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i\right) \to^d N(0, \boldsymbol{V}_3)$, *as* $n \to \infty$.

Then

$$n^{1/2}N^{-1}\bar{T}^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) \to^d N(0, \boldsymbol{J}_{\boldsymbol{\theta}_0})$$

where $\boldsymbol{J}_{\boldsymbol{\theta}_0} = f\boldsymbol{V}_1 + \boldsymbol{V}_2 + \boldsymbol{V}_3$.

Proof. See Appendix B. □

Lemma 11 shows the normality of estimating function, $\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)$, under model ξ and two-stage sampling design d. To do this, we write $n^{1/2}T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)$ as a sum of three

terms, such as,

$$\bar{T}^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) = \sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i + \left(\sum_{i\in S^{(2)}} w_i \bar{\boldsymbol{U}}_i - \sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i \right) + \sum_{i\in S^{(2)}} w_i \left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i \right).$$

Assumption 3 provides the normality of $N^{-1/2}\sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i$ under the model ξ by the survey-weighted CLT of $\bar{\boldsymbol{U}}_i$ under ξ. Assumption 4 provides the CLT under the stage-1 sampling design, given the finite population. Assumption 5 requires the independence among $w_i \left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i \right)$ conditioning on the finite population and stage-1 sample.

Theorem 8. *Let $\boldsymbol{\theta}_0$ be the true parameter in model (58) and $\hat{\boldsymbol{\theta}}$ be the solution of (60). If regularity conditions C1, C2, and C3 hold together with A6 and A8 in Appendix A, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to^d N(0, \boldsymbol{V}_{\boldsymbol{\theta}_0}) \text{ under } \xi \text{ and } d, \text{ as } n \to \infty$$

where $\boldsymbol{V}_{\boldsymbol{\theta}_0} = \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{J}_{\boldsymbol{\theta}_0} \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}$.

Proof. From Theorem 7, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, i.e. $\hat{\boldsymbol{\theta}} \to^p \boldsymbol{\theta}_0$ w.r.t. ξ and d as $n \to \infty$.

Linearizing $\hat{\boldsymbol{U}}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$, we have

$$0 = \hat{\boldsymbol{U}}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) + \frac{\partial \hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \frac{\partial^2 \hat{\boldsymbol{U}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$.

Since $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = o_p(1)$, $T^{-1}\frac{\partial^2 \hat{\boldsymbol{U}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = O_p(1)$ by condition C3, and $T^{-1}\frac{\partial \hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \to^p \boldsymbol{I}_{\boldsymbol{\theta}_0}$

as $n \to \infty$ (Lemma 10), it follows that

$$T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) = -\boldsymbol{I}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|), \tag{66}$$

Thus, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\sqrt{n}T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) + o_p(1).$$

Using Lemma 11 and Slutsky's theorem, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to^d N(0, \boldsymbol{V}_{\boldsymbol{\theta}_0}) \text{ as } n \to \infty.$$

□

Now partition $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, \boldsymbol{\phi}^T)^T$ where $\boldsymbol{\psi}$ is of dimension q and $\boldsymbol{\phi}$ is of dimension $p-q$. We are interested in testing $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ vs. $H_a : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0$. Denote $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\psi}}^T, \hat{\boldsymbol{\phi}}^T)^T$ and $\hat{\boldsymbol{U}} = (\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}, \hat{\boldsymbol{U}}_{\boldsymbol{\phi}})$ correspondingly. Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\psi}_0^T, \tilde{\boldsymbol{\phi}}^T)^T$ be the constrained maximum pairwise likelihood estimator of $\boldsymbol{\theta}$ obtained when $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Consider the following partition

$$\boldsymbol{I}_{\boldsymbol{\theta}_0} = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\psi}\boldsymbol{\psi}} & \boldsymbol{I}_{\boldsymbol{\psi}\boldsymbol{\phi}} \\ \boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\psi}} & \boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}} \end{pmatrix} \text{ and } \boldsymbol{V}_{\boldsymbol{\theta}_0} = \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{\psi}\boldsymbol{\psi}} & \boldsymbol{V}_{\boldsymbol{\psi}\boldsymbol{\phi}} \\ \boldsymbol{V}_{\boldsymbol{\phi}\boldsymbol{\psi}} & \boldsymbol{V}_{\boldsymbol{\phi}\boldsymbol{\phi}} \end{pmatrix}.$$

Let $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}_0}$ and $\hat{\boldsymbol{J}}_{\boldsymbol{\theta}_0}$ be consistent estimators of $\boldsymbol{I}_{\boldsymbol{\theta}_0}$ and $\boldsymbol{J}_{\boldsymbol{\theta}_0}$ respectively.

Proposition 3. *(Wald test) Under the null hypothesis, if conditions C1, C2, C3, and A6 and A8 in Appendix A hold, then,*

$$T_W = n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \hat{\boldsymbol{V}}_{\boldsymbol{\psi}\boldsymbol{\psi}}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \to^d \chi_q^2, \text{ as } n, N \to \infty,$$

where $\hat{\boldsymbol{V}}_{\boldsymbol{\psi\psi}}$ *is the* $q \times q$ *left corner submatrix of* $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}_0}$ *and* $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}_0} = \hat{\boldsymbol{I}}_{\boldsymbol{\theta}_0}^{-1}\hat{\boldsymbol{J}}_{\boldsymbol{\theta}_0}\hat{\boldsymbol{I}}_{\boldsymbol{\theta}_0}^{-1}$.

Proof. The proof is a direct consequence of Theorem 8. □

Let $\boldsymbol{H}_{\boldsymbol{\psi\psi}} = \left(\boldsymbol{I}^{\boldsymbol{\psi\psi}}\boldsymbol{V}_{\boldsymbol{\psi\psi}}^{-1}\boldsymbol{I}^{\boldsymbol{\psi\psi}}\right)^{-1}$, where $\boldsymbol{I}^{\boldsymbol{\psi\psi}} = (\boldsymbol{I}_{\boldsymbol{\psi\psi}} - \boldsymbol{I}_{\boldsymbol{\psi\phi}}\boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}})^{-1}$.

Proposition 4. *(Generalized score test) Under the null hypothesis, if conditions C1, C2, and C3 hold, then,*

$$T_{GS} = nT^{-2}\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}^{T}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})\hat{\boldsymbol{H}}_{\boldsymbol{\psi\psi}}^{-1}\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}}) \to^{d} \chi_q^2, \text{ as } n, N \to \infty$$

where $\hat{\boldsymbol{H}}_{\boldsymbol{\psi\psi}}$ *is a consistent estimator of* $\boldsymbol{H}_{\boldsymbol{\psi\psi}}$*, replacing* $\boldsymbol{I}_{\boldsymbol{\theta}_0}$ *by* $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}_0}$ *and* $\boldsymbol{V}_{\boldsymbol{\theta}_0}$ *by* $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}_0}$.

Proof. Since $\tilde{\boldsymbol{\phi}}$ satisfies $\hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}}) = 0$ and $\sqrt{n}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) = O_p(1)$ by Theorem 8, using a Taylor's expansion for $\hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})$, we have

$$0 = \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}}) = \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) + \frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + o_p(n^{-1/2}).$$

That is,

$$\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 = -\left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right)^{-1} \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) + o_p(n^{-1/2}).$$

Also, using a Taylor's expansion for $\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})$ about $\boldsymbol{\phi}_0$ and Theorem 8, we have

$$\begin{aligned}
\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}}) &= \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) + \frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + o_p(n^{-1/2}) \\
&= \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) - \frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right)^{-1} \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) + o_p(n^{-1/2}) \\
&= \left(\mathbf{I},\ -\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}}\left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}}\right)^{-1}\right) \hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) + o_p(n^{-1/2}).
\end{aligned}$$

Since $T^{-1}\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}} \to^p \boldsymbol{I}_{\boldsymbol{\psi\phi}}$ and $T^{-1}\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}} \to^p \boldsymbol{I}_{\boldsymbol{\phi\phi}}$ as $n \to \infty$, the following asymptotic approximations hold

$$
\begin{aligned}
T_{GS} &\approx nT^{-2}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)^T \begin{pmatrix} \mathbf{I} \\ -\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}} \left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}}\right)^{-1} \end{pmatrix}^T \hat{\boldsymbol{H}}^{-1}_{\boldsymbol{\psi\psi}} \begin{pmatrix} \mathbf{I} \\ -\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}} \left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\phi}}\right)^{-1} \end{pmatrix} \hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) \\
&\approx n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \boldsymbol{I}^T_{\boldsymbol{\theta}_0} \begin{pmatrix} \mathbf{I} \\ -\boldsymbol{I}_{\boldsymbol{\psi\phi}}\boldsymbol{I}^{-1}_{\boldsymbol{\phi\phi}} \end{pmatrix} \boldsymbol{H}^{-1}_{\boldsymbol{\psi\psi}} \begin{pmatrix} \mathbf{I} \\ -\boldsymbol{I}_{\boldsymbol{\psi\phi}}\boldsymbol{I}^{-1}_{\boldsymbol{\phi\phi}} \end{pmatrix}^T \boldsymbol{I}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\approx n(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \boldsymbol{V}^{-1}_{\boldsymbol{\psi\psi}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \\
&\to^d \chi^2_q.
\end{aligned}
$$

□

Proposition 5. *(Likelihood-ratio type test) If conditions C1, C2 and C3 hold, then,*

$$T_{LR} = -2nT^{-1}\left(wcl(\tilde{\boldsymbol{\theta}}) - wcl(\hat{\boldsymbol{\theta}})\right) \to^d \sum_k c_k \chi^2_{1k} \text{ as } n, N \to \infty$$

where χ^2_{1k} are iid χ^2_1 variables and c_k are eigenvalues of $\boldsymbol{V}_{\boldsymbol{\psi\psi}}(\boldsymbol{I}^{\boldsymbol{\psi\psi}})^{-1}$, $k = 1, ..., q$.

Proof. Since $\hat{\boldsymbol{\theta}} \to^p \boldsymbol{\theta}_0$ as $n \to \infty$, we expand $wcl(\boldsymbol{\theta}_0)$ around $\hat{\boldsymbol{\theta}}$, giving

$$wcl(\boldsymbol{\theta}_0) = wcl(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \frac{\partial wcl(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \frac{\partial^2 wcl(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$. Due to the existence and boundedness of the third derivative of $wcl(\boldsymbol{\theta})$, $\boldsymbol{I}_Y(\boldsymbol{\theta}^*) - \boldsymbol{I}_Y(\boldsymbol{\theta}_0) = o_p(1)$ where $\boldsymbol{I}_Y(\boldsymbol{\theta}^*) = T^{-1}\frac{\partial^2 wcl(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$; Also by

Lemma 10, $\|\boldsymbol{I}_Y(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\theta}_0}\| = o_p(1)$ and so $\|\boldsymbol{I}_Y(\boldsymbol{\theta}^*) - \boldsymbol{I}_{\boldsymbol{\theta}_0}\| = o_p(1)$.

Since $\frac{\partial wcl(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = 0$, we have

$$\begin{aligned} nT^{-1}[wcl(\hat{\boldsymbol{\theta}}) - wcl(\boldsymbol{\theta}_0)] &= -\frac{1}{2}n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T T^{-1} \frac{\partial^2 wcl(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \\ &= -\frac{1}{2}n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \boldsymbol{I}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) + o_p(1). \end{aligned}$$

Similarly, expanding $wcl(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})$ around $\boldsymbol{\phi}_0$ yields

$$\begin{aligned} nT^{-1}[wcl(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}}) - wcl(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)] &= -\frac{1}{2}n(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^T T^{-1} \frac{\partial^2 wcl(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + o_p(1) \\ &= -\frac{1}{2}n(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^T \boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + o_p(1) \\ &= -\frac{1}{2}n \begin{pmatrix} 0 \\ \\ \tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \end{pmatrix}^T I_{\boldsymbol{\theta}_0} \begin{pmatrix} 0 \\ \\ \tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \end{pmatrix} + o_p(1). \end{aligned}$$

In Theorem 8, we have shown that, as $n \to \infty$, $T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) \approx -\boldsymbol{I}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Then we have the following approximations, as $n \to \infty$,

$$\begin{aligned} \tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 &\approx -\left(\frac{\partial \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right)^{-1} \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) \\ &\approx -\boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1} T^{-1} \hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) \\ &= \begin{pmatrix} 0 \\ \\ -\boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1} \end{pmatrix}^T \begin{pmatrix} T^{-1}\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) \\ \\ T^{-1}\hat{\boldsymbol{U}}_{\boldsymbol{\phi}}(\boldsymbol{\psi}_0, \boldsymbol{\phi}_0) \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} 0 \\ -\boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1} \end{pmatrix}^T T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0) \\
&\approx \begin{pmatrix} 0 \\ -\boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1} \end{pmatrix}^T (-\boldsymbol{I}_{\boldsymbol{\theta}_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&= \begin{pmatrix} 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\psi\psi}} & \boldsymbol{I}_{\boldsymbol{\psi\phi}} \\ \boldsymbol{I}_{\boldsymbol{\phi}\psi} & \boldsymbol{I}_{\boldsymbol{\phi\phi}} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi}\psi} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \end{pmatrix}.
\end{aligned}
$$

Then, we have

$$
\begin{pmatrix} 0 \\ \tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \end{pmatrix} \approx \begin{pmatrix} 0 & 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & \mathbf{I} \end{pmatrix} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).
$$

Using the above fact, we may write T_{LR} as

$$
\begin{aligned}
T_{LR} &= -2nT^{-1}\left(wcl(\tilde{\boldsymbol{\theta}}) - wcl(\hat{\boldsymbol{\theta}})\right) \\
&= n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \boldsymbol{I}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \\
&- n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \begin{pmatrix} 0 & 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & \mathbf{I} \end{pmatrix}^T \boldsymbol{I}_{\boldsymbol{\theta}_0} \begin{pmatrix} 0 & 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & \mathbf{I} \end{pmatrix} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) + o_p(n^{-1/2})
\end{aligned}
$$

$$
\begin{aligned}
&= n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T \left[\boldsymbol{I}_{\boldsymbol{\theta}_0} - \begin{pmatrix} 0 & 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & \mathbf{I} \end{pmatrix}^T \boldsymbol{I}_{\boldsymbol{\theta}_0} \begin{pmatrix} 0 & 0 \\ \boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & \mathbf{I} \end{pmatrix} \right] (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + o_p(n^{-1/2}) \\
&= n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\psi\psi}} - \boldsymbol{I}_{\boldsymbol{\psi\phi}}\boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi\psi}} & 0 \\ 0 & 0 \end{pmatrix} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + o_p(n^{-1/2}) \\
&= n(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0)^T (\boldsymbol{I}_{\boldsymbol{\psi\psi}} - \boldsymbol{I}_{\boldsymbol{\psi\phi}}\boldsymbol{I}_{\boldsymbol{\phi\phi}}^{-1}\boldsymbol{I}_{\boldsymbol{\phi}\psi})(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0) + o_p(n^{-1/2}) \\
&= n(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0)^T (\boldsymbol{I}^{\boldsymbol{\psi\psi}})^{-1} (\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0) + o_p(n^{-1/2}).
\end{aligned}
$$

As $\sqrt{n}(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0) \to^d N(0, V_{\boldsymbol{\psi\psi}})$ and $(\boldsymbol{I}^{\boldsymbol{\psi\psi}})^{-1}$ is symmetric, by 20.28(a) in Seber (2007),

$$T_{LR} \to^d \sum_{k=1}^{q} \lambda_k z_k$$

where the z_k are i.i.d. χ_1^2 random variables and the λ_k are eigenvalues of $\boldsymbol{V}_{\boldsymbol{\psi\psi}}(\boldsymbol{I}^{\boldsymbol{\psi\psi}})^{-1}$. In reality, the λ_k are unknown and replaced by $\hat{\lambda}_k$ that are eigenvalues of $\hat{\boldsymbol{V}}_{\boldsymbol{\psi\psi}}(\hat{\boldsymbol{I}}^{\boldsymbol{\psi\psi}})^{-1}$ where $\hat{\boldsymbol{I}}^{\boldsymbol{\psi\psi}} = (\hat{\boldsymbol{I}}_{\boldsymbol{\psi\psi}} - \hat{\boldsymbol{I}}_{\boldsymbol{\psi\phi}}\hat{\boldsymbol{I}}_{\boldsymbol{\phi\phi}}^{-1}\hat{\boldsymbol{I}}_{\boldsymbol{\phi}\psi})^{-1}$. □

When the effective sample size that depends on the number of PSUs is small and q is large, $\hat{V}_{\boldsymbol{\psi\psi}}^{-1}$ and $\hat{Var}^{-1}(\hat{\boldsymbol{U}}_{\boldsymbol{\psi}})$ become unstable; hence, both the Wald and generalized score tests suffer instability. Moreover, the Wald test is not invariant to a reparametrization, while the score test and likelihood ratio test are invariant. T_{LR} requires estimation under both H_0 and H_a; on the other hand, T_W only needs estimates under H_a and T_{GS} needs only estimates under H_0. In some cases it is easier to find estimates under H_0, which makes T_{GS} more favourable.

Pace, Salvan and Sartori (2011) proposed a parameterization invariant adjustment on T_{LR} that leads to the usual asymptotic chi-square distribution. Following their work, we propose adjusted composite likelihood ratio test by scaling T_{LR}, as $n, N \to \infty$,

$$T_{adjLR} = \frac{T_{GS}}{nT^{-2}\hat{\boldsymbol{U}}^T(\boldsymbol{\theta})(\boldsymbol{I}(\boldsymbol{\theta}))^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta})} T_{LR} \to^d \chi^2_p$$

for the case of no nuisance parameters. Similarly, we propose

$$T_{adjLR} = \frac{T_{GS}}{nT^{-2}\hat{\boldsymbol{U}}^T_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})(\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}))^{\boldsymbol{\psi}\boldsymbol{\psi}}\hat{\boldsymbol{U}}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0, \tilde{\boldsymbol{\phi}})} T_{LR} \to^d \chi^2_q$$

for the case of nuisance parameters. The adjusted CLR test can be shown to be invariant to reparametrization .

7.5 Simulation study

We conduct a simulation study on the asymptotic normality of the WCL estimator and evaluate the performance of proposed test statistics based on the WCL.

We generate $B = 1000$ finite populations from a two-level random intercept model

$$y_{ij} = \beta_0 + \beta_1 x^{(1)}_{ij} + \beta_2 x^{(2)}_{ij} + \beta_3 x^{(3)}_{ij} + \beta_4 x^{(4)}_{ij} + v_i + e_{ij};$$

where $\boldsymbol{\beta} = (0.5, 1, 0, 0, 0)^T$, $e_{ij} \sim N(0, \sigma^2_e)$, $v_i \sim N(0, \sigma^2_v)$ and e_{ij}, v_i are independent. We set $\sigma^2_e = 2.0$ and $\sigma^2_v = 0.5$. Each finite population consists of $N = 1,000$ clusters of equal size and each cluster has $M = 100$ elements.

A probability sample is drawn from each finite population under a two-stage

sampling design:

- First draw $n = 50, 100$ clusters under a simple random sampling design,
- second, within each selected cluster, draw $m = 5$ elements using the Rao-Sampford PPS sampling design. The size variable z_{ij} is observable for all elements within the selected clusters. Both invariant and non-invariant selections are considered. For invariant selection, we set

$$z_{ij} = \left(1 + \exp\{-0.5 * [\frac{e_{ij}}{\alpha} + e^*_{ij}(1 - \alpha^{-2})^{-1/2}]\}\right)^{-1},$$

  where $\alpha = 1$, 3 or 20 and e^*_{ij} is independent of e_{ij}, but generated from the same distribution $N(0, \sigma^2_e)$. As α increases, the design becomes less informative. For non-invariant selection, we replace e_{ij} by $v_i + e_{ij}$ and e^*_{ij} by $v^*_i + e^*_{ij}$ where v^*_i is independent of v_i and generated from the same distribution. A similar set-up was also used in Asparouhov (2006), Rao, Verret, and Hidiroglou (2013), and Yi, Rao, and Li (2016).

Under this set-up, stage-1 sampling is non-informative while stage-2 sampling is informative since the size variable and response share the random errors. The informativeness decreases as α increases.

To study the asymptotic normality of the WCL estimator $\hat{\boldsymbol{\theta}}$, we produce 95% confidence intervals for each sample, i.e. $\hat{\boldsymbol{\theta}} \pm 1.96 se(\hat{\boldsymbol{\theta}})$ and evaluate the coverage rates for both regression and random effect parameters. Under asymptotic normality, we expect the coverage rates to be close to 95%. In addition to the WCL, other methods were also used to construct the confidence intervals, REML without weighting and MPML with scaled weights (A) and (A1) as described in Section (7.3).

To assess the performance of T_W, T_{GS}, T_{LR}, ant T_{adjLR} we consider testing the hypothesis: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_a : *at least one of them is not zero.* We calculated the rejection rates at nominal level 0.05 and 0.10. As T_{LR} requires the calculation of the distribution of a linear combination of independent χ^2_1 random variables, we consider the first-order and second-order Rao-Scott corrections (RS1 , RS2), Rao and Scott (1984).

Simulation results are presented in Tables 23 to 24. Under informative design (α =1), the WCL produces confidence intervals with more than 90% coverage rate for regression parameters and more than 90% for random error parameters. In contrast, the REML produce confidence intervals with an under-coverage for β_0 and random error parameters, and the methods A and A1 produce the confidence intervals with under-coverage for random error parameters (Similar results can be seen in Rabe-Hesketh and Skrondal, 2006). The under-coverage becomes more serious when the sample size n increases. When the design is less informative (α =3), all methods produce satisfactory results. For non-informative design, REML has the best performance because weighting used in the other methods increases variability. All the methods perform similarly under invariant and non-invariant designs.

Turning to the performance of statistical tests based on the WCL, we note that Wald test tends to over-reject the null hypothesis under all set-ups. In general, score test and likelihood ratio test under Rao-Scott adjustment perform similarly and produce the rejection rates close to the nominal level. Adjusted LR by Pace, Salvan and Sartori (2011) has better performance when $\alpha = 1$ but slightly under-rejects H_0 at level 0.05 under $\alpha = 20$. The performance of all tests improves as sample size, n, increases.

Table 22: Empirical coverage of 95% C.I. based on different estimation methods under invariant design

α	Parameter	n=50				n=100			
		WCL	**REML**	**A**	**A1**	**WCL**	**REML**	**A**	**A1**
1	β_0	0.94	0.19	0.89	0.94	0.93	0.02	0.83	0.93
	β_1	0.92	0.95	0.94	0.93	0.94	0.94	0.94	0.94
	β_2	0.93	0.95	0.95	0.94	0.93	0.95	0.95	0.95
	β_3	0.93	0.95	0.93	0.93	0.94	0.96	0.95	0.94
	β_4	0.93	0.93	0.92	0.93	0.94	0.95	0.93	0.93
	σ_v^2	0.86	0.92	0.90	0.91	0.90	0.93	0.90	0.90
	σ_e^2	0.87	0.70	0.61	0.75	0.90	0.58	0.52	0.72
3	β_0	0.94	0.85	0.93	0.94	0.94	0.75	0.94	0.94
	β_1	0.94	0.95	0.94	0.94	0.96	0.95	0.95	0.94
	β_2	0.94	0.95	0.95	0.94	0.94	0.95	0.95	0.95
	β_3	0.95	0.97	0.96	0.96	0.94	0.95	0.94	0.95
	β_4	0.93	0.94	0.92	0.93	0.93	0.95	0.94	0.94
	σ_v^2	0.89	0.91	0.94	0.93	0.93	0.93	0.95	0.95
	σ_e^2	0.91	0.92	0.89	0.89	0.93	0.93	0.87	0.88
20	β_0	0.94	0.94	0.94	0.94	0.95	0.94	0.95	0.94
	β_1	0.93	0.95	0.94	0.94	0.95	0.94	0.94	0.94
	β_2	0.94	0.96	0.95	0.95	0.94	0.95	0.94	0.94
	β_3	0.95	0.96	0.95	0.95	0.94	0.94	0.95	0.94
	β_4	0.95	0.96	0.95	0.94	0.94	0.95	0.94	0.94
	σ_v^2	0.88	0.92	0.92	0.92	0.92	0.94	0.95	0.95
	σ_e^2	0.93	0.93	0.91	0.90	0.95	0.95	0.91	0.91

Table 23: Empirical coverage of 95% C.I. based on different estimation methods under non-invariant design

α	Parameter	n=50				n=100			
		WCL	**REML**	**A**	**A1**	**WCL**	**REML**	**A**	**A1**
1	β_0	0.92	0.17	0.87	0.93	0.95	0.03	0.85	0.96
	β_1	0.92	0.94	0.93	0.92	0.93	0.96	0.95	0.94
	β_2	0.93	0.95	0.94	0.94	0.94	0.95	0.95	0.94
	β_3	0.92	0.95	0.94	0.94	0.93	0.93	0.93	0.93
	β_4	0.91	0.94	0.92	0.91	0.94	0.95	0.94	0.94
	σ_v^2	0.87	0.81	0.93	0.93	0.91	0.81	0.91	0.90
	σ_e^2	0.87	0.70	0.64	0.75	0.90	0.57	0.47	0.71
3	β_0	0.94	0.83	0.93	0.94	0.95	0.74	0.94	0.95
	β_1	0.93	0.94	0.94	0.94	0.95	0.94	0.95	0.95
	β_2	0.93	0.94	0.93	0.94	0.93	0.96	0.94	0.94
	β_3	0.93	0.95	0.94	0.95	0.95	0.96	0.95	0.94
	β_4	0.93	0.96	0.94	0.93	0.94	0.95	0.94	0.95
	σ_v^2	0.90	0.92	0.93	0.94	0.92	0.93	0.94	0.95
	σ_e^2	0.92	0.92	0.87	0.86	0.94	0.94	0.89	0.90
20	β_0	0.93	0.94	0.93	0.93	0.95	0.94	0.95	0.94
	β_1	0.93	0.95	0.94	0.93	0.95	0.95	0.96	0.96
	β_2	0.93	0.94	0.94	0.94	0.93	0.94	0.94	0.94
	β_3	0.93	0.94	0.93	0.93	0.95	0.95	0.95	0.94
	β_4	0.93	0.95	0.92	0.92	0.93	0.95	0.94	0.94
	σ_v^2	0.90	0.92	0.93	0.93	0.92	0.95	0.96	0.96
	σ_e^2	0.92	0.94	0.89	0.89	0.93	0.94	0.89	0.89

Table 24: Rejection rate at level 0.05 and 0.10 for testing $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ under different test statistics based on the WCL

α	Test	Non-invariant				Invariant			
		n=50		n=100		n=50		n=100	
		0.05	**0.1**	**0.05**	**0.1**	**0.05**	**0.1**	**0.05**	**0.1**
1	WALDT	0.13	0.19	0.09	0.15	0.11	0.18	0.08	0.15
	SCORE	0.08	0.14	0.08	0.13	0.06	0.13	0.06	0.12
	LRT-RS1	0.08	0.15	0.07	0.12	0.08	0.13	0.06	0.12
	LRT-RS2	0.08	0.14	0.06	0.12	0.06	0.12	0.06	0.11
	adjLR	0.06	0.13	0.07	0.12	0.04	0.11	0.05	0.12
3	WALDT	0.10	0.18	0.08	0.13	0.09	0.15	0.06	0.12
	SCORE	0.05	0.12	0.06	0.11	0.04	0.09	0.04	0.09
	LRT-RS1	0.07	0.14	0.07	0.12	0.06	0.11	0.06	0.11
	LRT-RS2	0.06	0.13	0.06	0.11	0.06	0.10	0.05	0.10
	adjLR	0.04	0.10	0.05	0.10	0.03	0.08	0.04	0.09
20	WALDT	0.10	0.17	0.10	0.14	0.08	0.14	0.08	0.14
	SCORE	0.04	0.10	0.06	0.12	0.03	0.09	0.05	0.11
	LRT-RS1	0.07	0.13	0.08	0.13	0.05	0.11	0.06	0.13
	LRT-RS2	0.06	0.13	0.07	0.13	0.04	0.10	0.06	0.12
	adjLR	0.03	0.09	0.06	0.12	0.03	0.08	0.04	0.11

Appendix A

Suppose $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$ satisfying $E_\xi E_d[\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0)] = 0$. We assume that the following conditions hold.

A1 Ω is a compact subset of the Euclidean space R^p,

A2 $\sup_{i,j,k} E_\xi[h^2_{ijk}(Y_{ij}, Y_{ik})] < \infty$ and $\sup_{i \le i \le N} E_\xi[\|\boldsymbol{Y}_i\|] < \infty$, where $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_i})^T$ and $h_{ijk}(Y_{ij}, Y_{ik}) = \sup_{\boldsymbol{\theta} \in \Omega} \|\boldsymbol{u}_{ijk}(\boldsymbol{\theta})\|$;

A3 For any given $c > 0$ and a given sequence $\{\boldsymbol{y}_i\}$ satisfying $\|\boldsymbol{y}_i\| < c$, the sequence of functions in $\boldsymbol{\theta}$, $\{\boldsymbol{u}_{ijk}(\boldsymbol{\theta})\}$, is equicontinuous on any open subset of Ω,

A4 Define $\Delta_T(\boldsymbol{\theta}) = E_\xi E_d[T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta})]$, where T is the design expectation of $E\{\sum_{i \in S^{(2)}} w_i E[\sum_{j>k \in S_i(2) w_{jk|i}}]\}$, i.e. $T = \sum_{i=1}^N T_i$ with $T_i = \frac{M_i(M_i-1)}{2}$. For any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\epsilon} \|\Delta_T(\boldsymbol{\theta})\| > \delta_\epsilon$,

A5 There exists a $\hat{\boldsymbol{\theta}} \in \Omega$ such that $\hat{\boldsymbol{U}}(\hat{\boldsymbol{\theta}}) = 0$,

A6 For any variable V_{ijk}, write $\bar{V} = T^{-1} \sum_{i=1}^N \sum_{1 \le j < k \le M_i} V_{ijk}$. If variables V_{ijk} satisfy $T^{-1} \sum_{i=1}^N \sum_{1 \le j < k \le M_i} V^2_{ijk} = O_\xi(1)$, then

$$\frac{1}{T} \sum_{i \in S^{(2)}} w_i \sum_{j>k \in S_i^{(1)}} w_{jk|i} V_{ijk} - \bar{V}$$

converges to 0 in design probability as $n \to \infty$;

A7 When the number of clusters in a sample approaches to infinity, the number of clusters in the corresponding population will tend to infinity as well. That is, if $n \to \infty$ then $N \to \infty$;

A8 $\max_i T_i/\bar{T} < c_1$ for some $c_1 > 0$ where $\bar{T} = T/N$.

Appendix B

Model based inference from survey data often proceeds under a model-design joint randomization framework. Suppose there is a sequence of two-level populations $\mathcal{F}_\nu$, indexed by $\nu = 1, 2, 3, ...$, generated from the model of interest ξ. Among them, finite population $\mathcal{F}_\nu$ consists of N_ν PSUs and there are $M_{\nu i}$ SSUs in the ith PSU. A two-stage sample is taken from $\mathcal{F}_\nu$, at the first stage n_ν PSUs are selected randomly from $\mathcal{F}_\nu$ and then $m_{\nu i}$ SSUs are selected in the ith sampled PSU at the second stage. For simplicity, we suppress the index ν.

We consider the following decomposition

$$\begin{aligned} Q &= n^{1/2}T^{-1}\hat{\boldsymbol{U}}(\boldsymbol{\theta}_0, Y) \\ &= n^{1/2}N^{-1}\sum_{i\in S^{(2)}} w_i\hat{\bar{\boldsymbol{U}}}_i \\ &= n^{1/2}N^{-1}\sum_{i\in^{(2)}} \bar{\boldsymbol{U}}_i \\ &+ n^{1/2}N^{-1}(\sum_{i\in S^{(2)}} w_i\bar{\boldsymbol{U}}_i - \sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i) \\ &+ n^{1/2}N^{-1}\sum_{i\in S^{(2)}} w_i\left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i\right) \\ &= Q_I + Q_{II} + Q_{III} \end{aligned}$$

where $\bar{\boldsymbol{U}}_i = \bar{T}^{-1}\sum_{j>k\in U_i^{(1)}} \boldsymbol{u}_{ijk}$ and $\hat{\bar{\boldsymbol{U}}}_i = \bar{T}^{-1}\sum_{j>k\in S_i^{(1)}} w_{jk|i}\boldsymbol{u}_{ijk}$.

The first component, $Q_I = n^{1/2}N^{-1}\sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i$, depends only on the finite population $\mathcal{F}$, the second component $Q_{II} = n^{1/2}N^{-1}\left(\sum_{i\in S^{(2)}} w_i\bar{\boldsymbol{U}}_i - \sum_{i\in U^{(2)}} \bar{\boldsymbol{U}}_i\right)$ depends

on the finite population $\mathcal{F}$ and the stage-1 sample $S^{(2)}$, and the last component $Q_{III} = n^{1/2}N^{-1}\sum_{i\in S^{(2)}} w_i \left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i\right)$ depends on the finite population $\mathcal{F}$ and samples of both stages, $S^{(2)}$ and $S_i^{(1)}, i \in S^{(2)}$. Also, given $\mathcal{F}$ and $S^{(2)}$, by the sampling design the elements in Q_{III}, $w_i \left(\hat{\bar{\boldsymbol{U}}}_i - \bar{\boldsymbol{U}}_i\right)$, are independent. It is easy to show all three terms have zero expectation under ξ and d, i.e., $E[Q_I] = 0$, $E[Q_{II}|\mathcal{F}] = 0$ and $E[Q_{III}|\mathcal{F}, S^{(2)}] = 0$.

Following Schenker and Welsh (1988), Chen and Rao (2007), and Fuller (2009), we establish the asymptotic normality of Q. For simplicity, we only consider the single-parameter case θ.

Lemma 12. *Let Q_I, Q_{II}, and Q_{III} be the three sequences of random variables defined as above. Let $\delta_1^2 = Var(Q_I)$, $\delta_2^2 = Var(Q_{II}|\mathcal{F})$, $\delta_3^2 = Var(Q_{III}|\mathcal{F}, S^{(2)})$, and $\delta^2 = \delta_1^2 + \delta_2^2 + \delta_3^2$ where $\delta_1, \delta_2, \delta_3 \geq 0$. Assume that*

1. *$\delta_1^{-1}Q_I \to^d N(0,1)$ as $N \to \infty$;*
2. *as $\nu \to \infty$, $\sup_{t\in R} |\Pr\{\delta_2^{-1}Q_{II} \leq t|\mathcal{F}\} - \Phi(t)| \to^p 0$ under d;*
3. *as $\nu \to \infty$, $\sup_{t\in R} |\Pr\{\delta_3^{-1}Q_{III} \leq t|\mathcal{F}, S^{(2)}\} - \Phi(t)| \to^p 0$ under d;*
4. *as $\nu \to \infty$, $\gamma_\nu = \delta/\delta_3 \to^p \gamma$, $\gamma_{1\nu} = \delta_1/\delta_3 \to^p \gamma_1$ and $\gamma_{2\nu} = \delta_2/\delta_3 \to^p \gamma_2$ under ξ and d; then $\gamma^2 = 1 + \gamma_1^2 + \gamma_2^2$.*

Then,

$$\delta^{-1}Q \to^d N(0,1), \text{ as } n \to \infty.$$

Proof.

$$|\Pr\{\delta^{-1}Q \leq t\} - \Phi(t)| \;=\; |E[\Pr\{\delta^{-1}(Q_I + Q_{II} + Q_{III}) \leq t|\mathcal{F}, S^{(2)}\}] - \Phi(t)|$$

$$
\begin{aligned}
&= |E[\Pr\{\delta_3^{-1}Q_{III} \le \delta_3^{-1}(\delta t - Q_I - Q_{II})|\mathcal{F}, S^{(2)}\}] - \Phi(t)| \\
&\le |E[\Pr\{\delta_3^{-1}Q_{III} \le S|\mathcal{F}, S^{(2)}\}] - E[\Phi(S)]| \\
&+ |E[\Phi(S)] - \Phi(t)|,
\end{aligned}
$$

where $S = \delta_3^{-1}(\delta t - Q_I - Q_{II}) = \gamma_\nu - \gamma_{1\nu}\delta_1^{-1}Q_I - \gamma_{2\nu}\delta_2^{-1}Q_{II}$.

Using assumption 3, $\sup_{t\in R} |\Pr\{\delta_3^{-1}Q_{III} \le t|\mathcal{F}, S^{(2)}\} - \Phi(t)| \to^p 0$ uniformly in t; then, $|\Pr\{\delta_3^{-1}Q_{III} \le S|\mathcal{F}, S^{(2)}\} - \Phi(S)|\mathcal{F}, S^{(2)}| \to^p 0$. Since $|\Pr\{\delta_3^{-1}Q_{III} \le S|\mathcal{F}, S^{(2)}\} - \Phi(S)|$ is bounded, $E[|\Pr\{\delta_3^{-1}Q_{III} \le S|\mathcal{F}, S^{(2)}\} - \Phi(S)|] \to 0$.

Let $S^* = \gamma t - \gamma_1\delta_1^{-1}Q_I - \gamma_2\delta_2^{-1}Q_{II}$. Using assumption 4, $\gamma_\nu \to^p \gamma$, $\gamma_{1\nu} \to^p \gamma_1$ and $\gamma_{2\nu} \to^p \gamma_2$ as $\nu \to \infty$. Further, if $\Phi(\cdot)$ is a bounded and continuous function, then $E[\Phi(S^*)] - E[\Phi(S)] \to 0$.

Let Z denote a standard normal random variable, then

$$
\begin{aligned}
E[\Phi(S^*)] &= E[P(Z \le \gamma t - \gamma_1\delta_1^{-1}Q_I - \gamma_2\delta_2^{-1}Q_{II})] \\
&= E[P(\gamma_2\delta_2^{-1}Q_{II} \le \gamma t - \gamma_1\delta_1^{-1}Q_I - Z)] \\
&= E[P(\delta_2^{-1}Q_{II} \le \gamma_2^{-1}(\gamma t - \gamma_1\delta_1^{-1}Q_I - Z)|\mathcal{F}, Z)] \\
&= E[P(\delta_2^{-1}Q_{II} \le R|\mathcal{F}, Z)]
\end{aligned}
$$

where $R = \gamma_2^{-1}(\gamma t - \gamma_1\delta_1^{-1}Q_I - Z)$.

Similarly, using assumption 2, we may show that

$$
\lim_{\nu\to\infty} E|P(\delta_2^{-1}Q_{II} \le R|\mathcal{F}, Z) - \Phi(R)| = 0.
$$

Let V, U denote independent standard normal random variables independent of

Z. As $\Phi(\cdot)$ is continuous and bounded,

$$\begin{aligned} E[\Phi(R)] &= E[\Phi(\gamma_2^{-1}(\gamma t - \gamma_1\delta_1^{-1}Q_I - Z))] \\ &= E[\Phi(\gamma_2^{-1}(\gamma t - \gamma_1 V - Z))] \text{ (by assumption 1)} \\ &= \Pr(U \leq \gamma_2^{-1}(\gamma t - \gamma_1 V - Z)) \\ &= \Pr(\gamma_2 U + \gamma_1 V + Z \leq \gamma t) \\ &= \Phi(\frac{\gamma}{\sqrt{1+\gamma_2^2+\gamma_1^2}}t) \\ &= \Phi(t) \end{aligned}$$

□

Remark:

1. Consider the case of non-random and finite cluster sizes M_i. If $f = \lim_{\nu\to\infty} n_\nu/N_\nu = 0$, assumption (1) can be ignored since $Q_I = O_p(\sqrt{n/N})$ and Q_{II}, Q_{III} are $O_p(1)$. When $f > 0$, under regularity conditions on the component log-densities $u_{ijk}(\theta) = \partial \log f(y_{ij}, y_{ik}; \boldsymbol{\theta})/\partial\theta$, we may have a central limit theorem for the composite likelihood score statistic of Q_I as $N \to \infty$.

2. Since cumulative distribution functions are monotone and bounded, and the distribution of a standard normal random variable is continuous, $\delta_2^{-1}Q_{II} \to^d N(0,1)$ as $n \to \infty$ implies assumption (2) and $\delta_3^{-1}Q_{III} \to^d N(0,1)$ as $n \to \infty$ implies assumption (3) by Lemma 3.2 of R.R. Rao (1962) Page 662.

3. Assumption (2) requires asymptotic normality of Q_{II} based on single-stage

sampling. For some particular(single-stage) procedures, conditions for asymptotic normality can found in the literature. For simple random sampling without replacement, conditions are found in e.g. Hajek (1960, 1961), for rejective sampling, in Hajek (1964) and for "random replacement sampling", in Rosen (1967).

Here we only give sufficient conditions for assumption (3).

Lemma 13. *Given $\mathcal{F}$ and $S^{(2)}$, assume that $w_k\left(\hat{\bar{U}}_k - \bar{U}_k\right)$ for $k \in S^{(2)}$ are independent. If the following holds*

$$\lim_{\nu\to\infty} \frac{\sum_{k\in S^{(2)}} u_k^{(4)} w_k^4}{(\sum_{k\in S^{(2)}} \delta_k^2 w_k^2)^2} = 0, \tag{67}$$

where $\mu_k^{(4)} = E\left[\{\hat{\bar{U}}_k - \bar{U}_k\}^4 | \mathcal{F}, S^{(2)}\right]$ and $\delta_k^2 = E\left[\{\hat{\bar{U}}_k - \bar{U}_k\}^2 | \mathcal{F}, S^{(2)}\right]$, then,

$$\sum_{k\in S^{(2)}} Z_k | \mathcal{F}, S^{(2)} \to^d N(0,1), \text{ as } n \to \infty,$$

where $Z_k = (\sum_{k\in S^{(2)}} \delta_k^2 w_k^2)^{-1/2} w_k \left(\hat{\bar{U}}_k - \bar{U}_k\right)$.

Proof.

$$\begin{aligned} E[e^{it\sum_{k\in S^{(2)}} Z_k}] &= E\left(E[e^{it\sum_{k\in S^{(2)}} Z_k} | \mathcal{F}, S^{(2)}]\right) \\ &= E\left(\prod_{k\in S^{(2)}} E[e^{itZ_k} | \mathcal{F}, S^{(2)}]\right). \end{aligned}$$

By (26.4) on page 343 of Billingsley (1995),

$$|e^{itZ_k} - (1 + itZ_k - \frac{1}{2}t^2Z^2)| \leq \min\{|tZ_k|^2, \frac{1}{6}|tZ_k|^3\}.$$

Then, for any $\epsilon > 0$

$$\begin{aligned}
|E[e^{itZ_k}|\mathcal{F}, S^{(2)}] - (1 - \frac{1}{2}t^2E[Z_k^2])| &\leq E[\min\{|tZ_k|^2, |tZ_k|^3\}|\mathcal{F}, S^{(2)}] \\
&\leq E[|tZ_k|^3 I\{|Z_k| < \epsilon\}|\mathcal{F}, S^{(2)}] \\
&+ E[|tZ_k|^2 I\{|Z_k| \geq \epsilon\}|\mathcal{F}, S^{(2)}] \\
&\leq \epsilon|t|^3 \frac{w_k^2\delta_k^2}{\sum_{k\in S^{(2)}} w_k^2\delta_k^2} + t^2E[|Z_k|^2 I\{|Z_k| \geq \epsilon\}|\mathcal{F}, S^{(2)}]
\end{aligned}$$

For fixed t and arbitrary ϵ and by the Lyapounov condition (67) that

$$\begin{aligned}
\sum_{k\in S^{(2)}} |E[e^{itZ_k}|\mathcal{F}, S^{(2)}] - (1 - \frac{1}{2}t^2E[Z_k^2])| &\leq \sum_{k\in S^{(2)}} E[\min\{|tZ_k|^2, |tZ_k|^3\}|\mathcal{F}, S^{(2)}] \\
&\leq \epsilon|t|^3 + t^2 \sum_{k\in S^{(2)}} E[|Z_k|^2 I\{|Z_k| \geq \epsilon\}|\mathcal{F}, S^{(2)}] \\
&\to 0.
\end{aligned}$$

Therefore, by (27.5) in Billingsley (1995),

$$|\prod_{k\in S^{(2)}} E[e^{itZ_k}|\mathcal{F}, S^{(2)}] - \prod_{k\in S^{(2)}} (1 - \frac{1}{2}t^2E[Z_k^2])| \leq \sum_{k\in S^{(2)}} |E[e^{itZ_k}|\mathcal{F}, S^{(2)}] - (1 - \frac{1}{2}t^2E[Z_k^2])| \to 0$$

and by (27.5) and (27.15) in Billingsley (1995),

$$|e^{-t^2/2} - \prod_{k=1}^{n} (1 - \frac{1}{2}t^2E[Z_k^2])| = |\prod_{k=1}^{n} e^{-t^2/2E[Z_k^2]} - \prod_{k=1}^{n} (1 - \frac{1}{2}t^2E[Z_k^2])|$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{n} |e^{-t^2/2E[Z_k^2]} - (1 - \frac{1}{2}t^2 E[Z_k^2])| \\
&\leq \sum_{k=1}^{n} t^4 (E[Z_k^2])^2 e^{t^2/2E[Z_k^2]} \\
&\leq t^4 e^{t^2} \sum_{k \in S^{(2)}} (E[Z_k^2])^2 \\
&\leq t^4 e^{t^2} \sum_{k \in S^{(2)}} E[Z_k^4]) \text{ (by Jensen's inequality)} \\
&\to 0.
\end{aligned}
$$

Thus,

$$|\prod_{k=1}^{n} E[e^{itZ_k}|\mathcal{F}, S^{(2)}] - e^{-t^2/2}| \to 0.$$

This completes our proof. □

Chapter 8

Discussion

In this chapter, we summarize the findings from our work and discuss some directions for future research.

8.1 Summary

We have focused on two methods for analysing correlated data, the QIF for marginal models and the pairwise composite likelihood for two-level models. Under a joint model-design framework, we extended these two methods to survey data. Survey weights were used to take account of sampling design.

For marginal models, we proposed estimation procedures based on the survey weighted QIF and studied their large sample properties. Through simulation studies, we showed that weighting can correct the bias of the unweighted QIF estimator due to an informative sampling design. Using the weighted QIF, we constructed Wald, alternative Wald, and likelihood-ratio-type tests for testing hypotheses about the regression parameter, as well as test statistic for goodness of fit. Different from

the QIF for non-survey data, the likelihood-ratio-type and goodness-of-fit test statistics no longer follow χ^2 distribution under the null hypothesis; instead that they are asymptotically equivalent to some linear combinations of independent χ^2 random variables. The distribution of linear combinations of χ^2 random variables can be approximated under Rao-Scott first- and second-order adjustments. Variable selection of marginal models for survey sample data was studied via the survey weighted QIF. The proposed variable selectors combine the survey weighted QIF with SCAD penalty and hold the oracle properties. The results from simulation studies showed the selection procedures based on the penalized survey-weighted QIF consistently selects the true subset of covariates and provides consistent and efficient estimates simultaneously. Both linearization and bootstrap methods can yield reasonable standard deviation estimates.

For two-level models, we followed Yi, Rao, and Li (2016) and further studied the inference properties of the WCL estimator. The WCL can provide an asymptotically unbiased estimate for both regression and random effects parameters, while the MPML methods fail to estimate the random error parameter consistently. We showed the asymptotic normality of the WCL estimator and proposed test statistics for both regression coefficients and random effects parameters. We also constructed different test statistics: Wald test, Score test, and composite likelihood ratio test and studied their performance. We show that the method possesses good coverage properties and significantly outperforms other approaches such as REML.

8.2 Future work

There are many issues that remain for future work. First, so far we only consider complete response case, while non-response is unavoidable in practice. Carrilllo, Chen, and Wu (2011) studied the GEE for survey data with missing value. We would like to explore the QIF for survey data with missing values. Second, we have proposed different test statistics based on the QIF and WCL, but we did not study the power of those tests. We would like to investigate the power of obtained tests. Third, joint selection probabilities are required for the WCL. Many approximation methods are available in literature. We will explore their impact on the performance of the WCL.

List of References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. in Petrov, B.N.; Csaki, F., *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akademiai Kiado, 267 - 281.

[2] Asgari,F., Biglarian, A., Seifi, B., Bakhshi, A., Miri, H., Bakhshi,E. (2013). Using quadratic inference functions to determine the factors associated with obesity: findings from the STEPS Survey in Iran. *Annals of Epidemiology*, 23(9): 534–538.

[3] Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, 5(3): 439–460.

[4] Bai, Y., Fung, W.K. and Zhu, Z.Y. (2009) Penalized Quadratic Inference Functions for Single-Index Models with Longitudinal Data. *Journal of Multivariate Analysis*, 100: 152–161.

[5] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society: Series B*, 36: 192-236.

[6] Billingsley, P. (1995). *Probability and Measure* (3rd ed.). Wiley, New York.

[7] Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51: 279–292.

[8] Binder, D. A. and Roberts, G. R. (2003) Design-based and model-based methods for estimating model parameters. *Analysis of survey data* (Southampton, 1999), 29-48, Wiley Series in Survey Methodology. Wiley, Chichester.

[9] Binder, D.A. and G. Roberts. (2009). "Design and Model Based Inference for Model Parameters." In *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Pfeffermann, D. and Rao, C.R. (eds.) Vol. 29B. Chapter 24. Amsterdam.Elsevier. 666 p.

[10] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6): 2350–2383.

[11] Cantoni, E., Flemming, J.M. and Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linear Models. *Biometrics*, 61: 507-514.

[12] Carrillo, I.A., Chen, J. and Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *Canadian Journal of Statistics*, 38: 540-554.

[13] Carrillo, I.A., Chen, J. and Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics*, 27: 255-277.

[14] Chen, J., Rao, J.N.K. (2007). Asymptotic Normality under Two-phase Sampling Designs. *Statistica Sinica*, 17: 1047–1064.

[15] Cho, H. and Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica*, 23: 901–927.

[16] Cochran, W. G. (1963). *Sampling techniques*, 2nd ed. Wiley, New York.

[17] Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61: 384–390.

[18] Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, (82)2: 407–410.

[19] Dziak, J.J. (2006). Penalized quadratic inference function for variable selection in longitudinal research. Ph.D. Thesis, the Pennsylvania State University.

[20] Dziak, J.J. and Li, R. (2006). An overview on variable selection for longitudinal data. In Hong, D. editor, *Quantitative Medical Data Analysis*. World Sciences Publisher, Singapore.

[21] Emrich, L.J. and Piedmonte, M. R. (1991). A Method for Generating High-Dimensional Multivariate Binary Variables. *The American Statistician*, 45, p. 302–304

[22] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 13481360.

[23] Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2011). *Applied Longitudinal Analysis* (2^{nd} ed.) John Wiley & Sons. New Jersey.

[24] Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya C*, 37: 117–132.

[25] Fuller, W.A. (2009). *Sampling Statistics*, John Wiley & Sons, Inc.

[26] Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B*. 17: 269–278.

[27] Godambe, V.P. (1965). A review of the contributions towards a unified theory of sampling from finite populations. *International Statistical Review*, 33: 242–258.

[28] Godambe, V.P. (1975). A reply to my critics. *Sankhyd C*, 37: 53–76.

[29] Godambe, V.P. and Thompson, M.E. (2009). Estimating functions and survey sampling. In C. R. Rao and D. Pfeffermann, Eds. *Handbook of Statistics. Sample Surveys: Inference and Analysis Volume 29b*. Elsevier, 669–687.

[30] Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1): 43–56

[31] Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. Publications of the Mathematics Institute of the Hungarian Academy of Science, 5: 361–74.

[32] Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4): 14911523.

[33] Hansen, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4): 1029–1054.

[34] Han, P. and Song, P. (2011). A note on improving quadratic inference functions using linear shrinkage approach. *Statistics and Probability Letters* 81, 438–445.

[35] Hartley, H.O. and Sielken, R.L. (1975). A superpopulation viewpoint for finite population sampling. *Biometrics*, 31: 411422.

[36] Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Approximate variance estimators under the Rao-Sampford design. *Metron*, 66: 91–108.

[37] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47: 663685.

[38] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12: 55-67.

[39] Horn, R.A., Johnson, C.R. (1985). *Matrix Analysis*, New York.

[40] Hu, Y. and Song, P. (2012). Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes. *Statistics in Medicine*, 31: 787–800.

[41] Johnson, B.A., Lin, D.Y., and Zeng, D. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, 103(482): 672680.

[42] Korn, E.L. and B.I. Graubard. (1999). *Analysis of Health Surveys*, Wiley, New York.

[43] Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, Jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9: 1010–1019.

[44] Leisch, F., Weingessel, A., and Hornik, K. (2015). package "bindata". Available at `https://cran.r-project.org/web/packages/bindata/bindata.pdf&context=cssmwp`.

[45] Liang K-Y, Zeger SL. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13-22.

[46] Lindsay, B.G. (1988). Composite likelihood method. *Contemporary Mathematics*, 80: 221–39.

[47] Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1): 1–18.

[48] Mallows, C. L. (1973). Some Comments on C_P. *Technometrics*, 15(4): 661675.

[49] McCullagh, P. (1983). Quasi-Likelihood Functions. *The Annals of Statistics*, 11(1): 59–67.

[50] Molina, E.A., and Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis*, 13: 395–405.

[51] Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3: 169175.

[52] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(30): 370–384.

[53] Odueyungbo A., Browne D., Akhtar-Danesh N., Thabane L. (2008). Comparison of generalized estimating equations and quadratic inference functions using data from the National Longitudinal Survey of Children and Youth (NLSCY) database. *BMC Medical Research Methodology*, 8:28.

[54] Pace, L., Salvan A., and Sartori N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica* 21: 129-48.

[55] Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, 57: 120-125.

[56] Petersen, K.B., Pedersen, M.S. (2012). The Matrix Cookbook. Available at `http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf`.

[57] Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, 61: 317–337.

[58] Pilla, R.S. (2005). Inference under convex cone alternatives for correlated data. Eprint: arxiv.org/pdf/math/0506522.pdf.

[59] Pilla R.S., Qu, A., and Loader, C. (2006). Testing for order-restricted hypotheses in longitudinal data. *Journal of the Royal Statistical Society: Series B*, 68:437-455.

[60] Pinheiro, J. and Bates, D. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1): 12–35.

[61] Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics*, 62(2): 379–391.

[62] Qu A, Lindsay B.G., Li B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87:823-836.

[63] Qu, A. and Lindsay, B.G. (2003), Building adaptive estimating equations when inverse of covariance estimation is difficult. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65: 127 -142.

[64] Qu, A. and Song, P. (2004). Assessing robustness of generalized estimating equations and quadratic inference functions. *Biometrika*, 91(2):447-59.

[65] Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169: 805-827.

[66] Rao, R.R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics*, 33:659–680.

[67] Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statistical Association*, 3: 173–180.

[68] Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, 76: 221-230.

[69] Rao, J.N.K. and Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Properties Estimated from Survey Data. *The Annals of Statistics*, 12: 46-60.

[70] Rao, J.N.K. and Scott, A.J. (1987). On Simple Adjustments to Chi-Square Tests with Survey Data. *The Annals of Statistics*, 15: 385-397.

[71] Rao, J.N.K., Tausi, M. (2004). Estimating function variance estimation under stratified multistage sampling. *Communications in statistics*, 33: 2087–2095.

[72] Rao, J.N.K., Verret, F., and Hidiroglou, M.A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, 39(2): 263–282.

[73] Rao, J.N.K., Wu, C.F. (1988). Resampling inferences with complex survey data. *Journal of the American Statistical Association*, 83: 231–241.

[74] Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on re-sampling methods for complex surveys. *Survey Methodology*, 18: 209–217.

[75] Rosen, B. (1972a). Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I. *The Annals of Mathematical Statistics*, 43(2): 373–397.

[76] Rosen, B. (1972b). Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, II. *The Annals of Mathematical Statistics*, 43(3): 748–776.

[77] Rosen B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62(2): 159–191.

[78] Rotnitzky, A. and Jewell, N. (1990). Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data. *Biometrika*, 77(3): 485–497.

[79] Royall, R.M. (1971a). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference* (ed. V. P. Godambe and D. A. Sprott), 259–274. Holt, Rinehart & Winston, Toronto.

[80] Royall, R.M. (1971b). Discussion of paper by D. Basu. *In Foundations of Statistical Inference* (ed. V. P. Godambe and D. A. Sprott), 238–239. Holt, Rinehart & Winston, Toronto.

[81] Royall, R.M. and Eberhardt, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyd C*, 37: 43–52.

[82] Royall, R.M. and Cumberland, W.G. (1977). An empirical study of prediction theory in finite population sampling I: Simple random sampling and the ratio estimator. *Symposium on Survey Sampling*, Chapel Hill, N.C.

[83] Rubin-Bleuer, S. and Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6): 2789–2810.

[84] Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54: 499–513.

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

[86] Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2(6): 110-114.

[87] Schenker, N. and Welsh, A.H. (1988). Asymptotic Results for Multiple Imputation. *Annals of Statistics*, 16(4): 1550–1566.

[88] Schwarz, G. E. (1978), Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461-464,

[89] Seber, G.A.F. (2007). *A Matrix Handbook for Statisticians*, Wiley-Interscience.

[90] Shao, J. (2003). *Mathematical Statistics*, SpringerVerlag.

[91] Skinner, C.J. (1989). Domain Means, Regression and Multivariate Analysis. In Skinner, C. J., Holt, D. and Smith, T. M. F. eds. *Analysis of Complex Surveys*, Chichester, John Wiley & Sons.

[92] Song, P., Jiang, Z., Park, E. and Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. S*tatistics in Medicine*, 28: 3683–3696.

[93] Song, P.. (2007). *Correlated Data Analysis*, Springer, New York.

[94] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B, 58: 267-288.

[95] Wang, H., Li, G., and Tsai, C.L. (2007a). Regression coefficient and autoregressive order shrinkage and selection via LASSO. *Journal of Royal Statistical Society*, Series B. 69: 63-78.

[96] Wang, H., Li, R., and Tsai, C.L. (2007b). On the Consistency of SCAD Tuning Parameter Selector. *Biometrika*, 94: 553-558.

[97] Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser. B Stat. Methodol*, 71: 177–190.

[98] Wang, F., Wang, L., and Song, P. (2012). Quadratic inference function approach to merging longitudinal studies: validation and joint estimation. *Biometrika*, 99: 755–762.

[99] Wang, L., Wang, S., and Wang, G. (2014). Variable selection and estimation for longitudinal survey data. *Journal of Multivariate Analysis*, 130: 409–424.

[100] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the GaussNewton method. *Biometrika*, 61(3): 439–447.

[101] Westgate, P.M. and Braun, T.M. (2012). The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions. *Statistics in Medicine*, 31(20):2209–22.

[102] Wicklin, R., 2013, *Simulating Data with SAS*, SAS Institute Inc., Cary NC, pp. 154–157.

[103] Xu, C., Chen, J. and Mantel, H. (2013). Pseudo-likelihood-based Bayesian Information Criterion in Analysis of Survey Data. *Survey Methodology*, 39: 303–321.

[104] Yi, G., Rao, J.N.K., and Li, H. (2016). A Weighted Composite Likelihood Approach for Analysis of Survey Data under Two-Level Models. *Statistica Sinica*, 26: 569–587.

[105] Yuan, K.H. and Jennrich, R.I. (1998). Asymptotics of Estimating Equations under Natural Conditions. *Journal of Multivariate Analysis*, 65: 245–260.

[106] Zhang, F. (1999). *Matrix Theory*, Springer-Verlag, New York.

[107] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101: 1418-1429.