2018

Improved Methods and Selecting Classification Types for Time-Dependent Covariates in the Marginal Analysis of Longitudinal Data

I-Chen Chen
University of Kentucky, i.chen.chen@uky.edu
Author ORCID Identifier:
https://orcid.org/0000-0001-6764-8395

Digital Object Identifier: https://doi.org/10.13023/ETD.2018.098

I-Chen Chen, Student

Dr. Philip M. Westgate, Major Professor

Dr. Steven R. Browning, Director of Graduate Studies

Improved Methods and Selecting Classification Types for Time-Dependent Covariates in the Marginal Analysis of Longitudinal Data

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Public Health at
the University of Kentucky

By

I-Chen Chen

Lexington, Kentucky

Director: Dr. Philip M. Westgate, Associate Professor of Biostatistics
Lexington, Kentucky 2018

ABSTRACT OF DISSERTATION

Improved Methods and Selecting Classification Types for Time-Dependent Covariates in the Marginal Analysis of Longitudinal Data

Generalized estimating equations (GEE) are popularly utilized for the marginal analysis of longitudinal data. In order to obtain consistent regression parameter estimates, these estimating equations must be unbiased. However, when certain types of time-dependent covariates are presented, these equations can be biased unless an independence working correlation structure is employed. Moreover, in this case regression parameter estimation can be very inefficient because not all valid moment conditions are incorporated within the corresponding estimating equations. Therefore, approaches using the generalized method of moments or quadratic inference functions have been proposed for utilizing all valid moment conditions. However, we have found that such methods will not always provide valid inference and can also be improved upon in terms of finite-sample regression parameter estimation. Therefore, we propose a modified GEE approach and a selection method that will both ensure the validity of inference and improve regression parameter estimation.

In addition, these modified approaches assume the data analyst knows the type of time-dependent covariate, although this likely is not the case in practice. Whereas hypothesis testing has been used to determine covariate type, we propose a novel strategy to select a working covariate type in order to avoid potentially high type II error rates with these hypothesis testing procedures. Parameter estimates resulting from our proposed method are consistent and have overall improved mean squared error relative to hypothesis testing approaches.

Finally, for some real-world examples the use of mean regression models may be sensitive to skewness and outliers in the data. Therefore, we extend our approaches from their use with marginal quantile regression to modeling the conditional quantiles of the response variable. Existing and proposed methods are compared in simulation studies and application examples.

KEYWORDS: Generalized Estimating Equations, Time-Dependent Covariate, Empirical Covariance Matrix, Working Correlation Structure, Mean Squared Error, Marginal Quantile Regression

Author's signature: I-Chen Chen

Date: April 19, 2018

Improved Methods and Selecting Classification Types for Time-Dependent Covariates in the Marginal Analysis of Longitudinal Data

By
I-Chen Chen

Director of Dissertation: Dr. Philip M. Westgate

Director of Graduate Studies: Dr. Steven R. Browning

Date: April 19, 2018

ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Dr. Philip Westgate, for all of his guidance and advice throughout my dissertation study. Thanks for always being patient and promptly helping with any questions I had. I would also like to thank Dr. David Fardo, Dr. Erin Abner, and Dr. Solomon Harrar for serving on my disseration committee. I really appreciate their feedback and comments for the dissertation. I would finally like to thank my wife, Yu-Ru Chu (朱昱儒), for all of her support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Significance

Longitudinal studies in which independent participants contribute repeated measurements over time are common in practice. Generalized estimating equations (GEEs) [3] are popularly used for the marginal analysis of longitudinal data. The main feature of the GEE approach is that when the mean structure is assumed to be correctly specified, consistent regression parameter estimates can be obtained regardless of if the working correlation structure is correctly given. However, when certain types of time-dependent covariates are presented, the estimating equations, as well as the regression parameter estimates, can be biased due to the use of invalid moment conditions, which are functions of the parameters in statistical models and the data. Although invalid moment conditions do not occur when an independence working correlation structure is incorporated [4], resulting regression parameter estimation can be very inefficient because all valid moment conditions may not be used when employing this structure [5, 6].

In this dissertation, we therefore focus on utilizing all valid moment conditions, with the goal of improving estimation efficiency over GEE with an independence working structure, and comparing our proposed approach with the existing approaches in the presence of time-dependent covariates, as presented in Chapter 2. Chapter 3 discusses a strategy to select a working type of time-dependency because in practice it may not be the case that the data analyst knows the type of time-dependent covariate. In Chapter 4, we note that the application example regularly used in existing literature with respect to marginal mean regression for longitudinal data analysis may not be ideal due to highly skewed response distribution. Therefore, we extend our approaches presented in Chapters 2 and 3 to marginal quantile regression for modeling the conditional quantiles of the response variable. Chapter 5 summarizes the findings of this dissertation, discusses their importance and future work.

Specifically, Chapter 2 concentrates on improving estimation efficiency relative to the generalized method of moments (GMM) [7] approach proposed by Lai and Small [1] and the modified version of the quadratic inference functions (QIF) method [8] proposed by Zhou *et al.* [9]. Furthermore, since limited attention has been given to both approaches' validity of inference in finite-sample settings, we first propose a modified GEE approach to improve the validity of inference and regression parameter estimation of the two existing approaches. The resulting approach will be more efficient than GEE with an independence working structure, yet practically it still takes advantage of GEE's accessibility to analysts. Moreover, it has the potential to perform better than the GMM approach of Lai and Small [1] and the modified QIF of Zhou *et al.* [9] in small-sample settings due to potential variance inflations. However, which combination of method and working structure will result in the smallest variances of regression parameter estimates will be unknown to the data analyst, and therefore we extended the applicability of the correlation information criterion (CIC) [10] in order to select a combination to use for inference.

The current estimation methods require the researcher to specify the classification type of the time-dependent covariate, but this will often be unknown in practice. Therefore, multiple approaches have been proposed to choose a type of time-dependency. In short, Lai and Small [1] proposed hypothesis testing based on GMM, and Lalonde *et al.* [2] proposed an alternative approach using correlations that conducts the testing of each individual moment condition. However, these approaches can lead to too many moments being deemed valid, thus preferring biased regression parameter estimation. As a result, Chapter 3 introduces a criterion that accounts for the impacts moment conditions have on both the efficiency and bias of regression parameter estimation corresponding to time-dependent covariates, with the goal of minimizing mean squared error (MSE). Additionally, the proposed approach provides consistent estimation. We note that Leung *et al.* [11] considered an empirical likelihood (EL) approach [12] in which moment conditions that are not guaranteed to provide consistent estimation are weighted, relying upon their estimated likelihoods of being valid, and linearly combined. Although this approach avoids having to select a covariate type, and is no less efficient than GEE with an independence working structure, we later demonstrate that this approach can be inefficient relative to our proposed approach.

Existing methods for the marginal analysis of longitudinal data in the presence of time-dependent covariates have only been developed for the modeling of the mean. Nonetheless, for some real-world datasets the use of mean regression models may be sensitive to skewed response distribution and outliers in the data. Therefore, Chapter 4 first focuses on the use of marginal quantile regression and combines the estimating equations approach of Fu and Wang [13], which has been shown to improve estimation performance in marginal quantile regression and is robust to different error distributions, with our estimation method from Chapter 2. In consequence, the proposed approach can achieve notable gains in efficiency when compared with estimating equations under an independence correlation structure. Second, we extend the use of our selection approach from Chapter 3 to choose a working classification type such that consistent regression parameter estimation is a result.

Chapter 2 Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates

2.1 Introduction

Longitudinal studies in which subjects contribute repeated measurements over time are popular in practice. Generalized estimating equations (GEEs) [3] are routinely used for the marginal analysis of correlated data arising from such studies. When the mean structure is assumed to be correctly specified, consistent regression parameter estimates can often be obtained regardless of whether or not the working correlation structure is correctly given. However, accurately modeling the correlation structure can be very important with respect to estimation efficiency [6]. In addition, when utilizing the empirical sandwich estimator of the covariance matrix of the regression parameter estimates, valid large-sample inference can be attained.

Only a limited number of studies have addressed the validity of GEE when covariates are time-dependent. Although GEE requires unbiased estimating equations in order to produce consistent regression parameter estimates, certain types of time-dependent covariates can violate this requirement and result in invalid moment conditions when GEE incorporates arbitrary working correlation structures, particularly when the non-diagonal elements of the correlation matrix are non-zero [4]. Therefore, Pepe and Anderson [4] suggested that the use of GEE with an independence working correlation structure, which will yield unbiased estimating equations, may be a safe approach in the presence of time-dependent covariates. However, when a marginal analysis contains time-varying covariates, using an independence working structure can lead to a considerable loss of parameter estimation efficiency because not all valid moment conditions are utilized by the corresponding estimating equations [5, 6].

To improve estimation efficiency in the presence of time-dependent covariates by making use of all valid moment conditions, Lai and Small [1] proposed the use of generalized method of moments (GMM) [7]. They showed that their GMM approach maintains or improves upon the efficiency of GEE with an independence working structure. An alternative approach that has been proposed to improve efficiency is a modified version of the quadratic inference functions (QIF) method [8]. This method has the potential to improve efficiency relative to GEE, and therefore Zhou *et al.* [9] modified the QIF approach such that it includes all valid moment conditions, thus theoretically resulting in greater efficiency relative to GEE with an independence working structure. They showed via simulation that their modified QIF and the GMM approach of Lai and Small [1] performed similarly in terms of regression parameter estimation when subjects contributed 5 repeated measurements, whereas their QIF approach performed better when subjects contributed 15 repeated measurements.

Although the advantages of the GMM of Lai and Small [1] and the modified QIF approach of Zhou *et al.* [9] have been demonstrated, limited attention has been given to their validity and utility in finite-sample settings. In previous empirical work, it has been shown that general methods based on GMM and QIF can result in liberal inference, i.e., inflated test size and sub-nominal confidence interval (CI) coverage probability (CP), due to the need for finite-sample corrections to standard error estimators [14, 15, 16, 17]. The reason for this is because these approaches utilize an empirical estimator for the optimal weighting matrix, and the use of this estimator can increase the variances of regression parameter estimates relative to their theoretical variances. The degree of variance inflation increases with the number of moment conditions [17] and as the number of subjects decreases. The variance inflation can, at least partially, offset any efficiency gains due to the use of the modified QIF and GMM approaches. Additionally, in results to be presented later, we found the GMM approach of Lai and Small [1] can also result in biased standard error estimates, and thus invalid inference, due to the singularity of the approach's weighting matrix. We note that Lai and Small [1] pointed out certain instances from their simulation study in which the empirical CPs resulting from the use of 95% CIs were low, even as small as 65.5%.

To improve upon the validity of inference and regression parameter estimation of the modified QIF and GMM approaches in the presence of time-dependent covariates, we first propose a modified GEE approach. With this approach, we modify the inverse of any working correlation structure such that any components that create invalid moment conditions are removed. Therefore, the resulting approach will be more efficient than utilizing GEE with an independence working structure. Furthermore, it also has the potential to perform better than the GMM approach of Lai and Small [1] and the modified QIF in small-sample settings due to potential variance inflations. Second, we propose an approach to select a method to use for inference. In the GEE literature, criteria such as the correlation information criterion (CIC) [10] can be used to select a working correlation structure, and Westgate [18] proposed simultaneously selecting a working correlation structure and either GEE or the QIF approach by utilizing the trace of the empirical covariance matrix (TECM). In this chapter, we extend the use of the popular CIC to choose a method, either our modified GEE, the GMM of Lai and Small [1], or the modified QIF approach, and a working structure within the modified GEE or QIF. We note that the bias induced by the singularity of the weighting matrix employed by the GMM of Lai and Small [1] can have a detrimental impact if allowed to be selected by our extended CIC approach. Therefore, we propose for consideration a modified GMM approach that removes the singularity.

Section 2.2 introduces notation and issues with time-dependent covariates, and discusses GEE with an independence working structure, the GMM of Lai and Small [1], and the modified QIF approach. We also consider bias corrections for the empirical estimators of the covariance matrix of regression parameter estimates. In Section 2.3, we propose the modified GMM and GEE approaches in the presence of time-dependent covariates. Furthermore, we introduce the extended CIC selection criterion used for selecting the best combinations of approach and structure. In Section 2.4, we carry out a simulation study to compare the estimation performances of the proposed methods and to assess the CIC's utility. In Section 2.5, we make comparisons in application to anthropometric screening data to evaluate the association between body mass index (BMI) and morbidity among children in the Philippines. We give concluding remarks in Section 2.6. Finally, supplementary material is presented in Section 2.7.

2.2 Time-Dependent Covariates and Current Methods

2.2.1 Notation and Time-Dependent Covariates

Assume we conduct a longitudinal study in which there are N independent subjects, and these subjects are measured at each of T distinct time points. However, participants need not have the same number of time points. The observed outcome vector for the ith subject is denoted by $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iT}]^T$, which has a marginal mean given by $E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = \boldsymbol{\mu}_i$ that is linked to covariates via a function, f, such that $f(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ for $\mathbf{x}_{ij} = [1, x_{1ij}, \ldots, x_{(p-1)ij}]^T$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^T$. The corresponding working covariance matrix for $\boldsymbol{Y}_i$ is given by $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\left(\boldsymbol{\alpha}\right)\boldsymbol{A}_i^{1/2}$, $i = 1, \ldots, N$. Here, $\boldsymbol{A}_i = diag[\phi\nu(\mu_{i1}), \ldots, \phi\nu(\mu_{iT})]$ is a diagonal matrix representing the working marginal variances, ϕ is a scale parameter assuming common dispersion, ν is a known function, and $\boldsymbol{R}_i\left(\boldsymbol{\alpha}\right)$ is a symmetric positive definite working correlation matrix with 1 along the diagonal and one or more unknown parameters given by $\boldsymbol{\alpha}$.

With the GEE approach [3] to marginal modeling, let $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^T$. To obtain the estimate of the regression parameters, $\hat{\boldsymbol{\beta}}$, we iteratively solve

$$\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}\left(\boldsymbol{\alpha}\right)\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{2.1}$$

In Equation (2.1), the $(k+1)$th row corresponds to the estimating equation for β_k and is given by

$$\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{j=1}^{T}\frac{\partial\mu_{is}}{\partial\beta_k}v_i^{sj}(Y_{ij}-\mu_{ij})=0,$$

where v_i^{sj}, $i=1,...,N$ and $s,j=1,...,T$, is the (s,t)th element of $\boldsymbol{V}_i^{-1}$. If β_k corresponds to certain types of time-dependent covariates, as specified in the following paragraph, then we may not have $E\left[\frac{\partial\mu_{is}}{\partial\beta_k}(Y_{ij}-\mu_{ij})\right]=0$ for all s, j, $1\leqslant s,j\leqslant T$. We note that if GEE incorporates a working independence structure, then the only moment conditions that are used are the ones such that $j=s$, and hence all corresponding expected values of these moment conditions are 0 regardless of covariate type. Therefore, Pepe and Anderson [4] advocated the use of GEE with a working independence structure.

Lai and Small [1] presented three types of time-dependent covariates. The kth covariate is classified as a Type I time-dependent covariate if it satisfies

$$E\left[\frac{\partial\mu_{is}}{\partial\beta_k}(Y_{ij}-\mu_{ij})\right]=0 \text{ for all } s,\ j,\quad s=1,...,T,\quad j=1,...,T. \tag{2.2}$$

A common example of a Type I covariate is time itself; i.e., age, grade levels, or educational stages. We note that time-independent covariates also satisfy Equation (2.2). A Type II time-dependent covariate satisfies

$$E\left[\frac{\partial\mu_{is}}{\partial\beta_k}(Y_{ij}-\mu_{ij})\right]=0 \text{ for all } s\geqslant j,\quad j=1,...,T.$$

Specifically, Y_{ij} given $\mathbf{x}_{ij}$ does not influence the future time-dependent covariate process, $\mathbf{x}_{i,j+1}$,...,$\mathbf{x}_{iT}$. In words, there is no feed-back cycle from the outcomes to the covariate process. A time-dependent covariate is defined to be of Type III if it is not of Type II or IV and it satisfies

$$E\left[\frac{\partial\mu_{is}}{\partial\beta_k}(Y_{ij}-\mu_{ij})\right]\neq 0 \text{ for some } s>j,\quad j=1,...,T,$$

such that there does exist a feed-back cycle in which the covariate value affects the outcome, and that outcome influences future covariate values. Lalonde *et al.* [2]

defined a Type IV time-dependent covariate, which is the opposite of a Type II covariate in that it satisfies

$$E\left[\frac{\partial \mu_{is}}{\partial \beta_k}(Y_{ij} - \mu_{ij})\right] = 0 \text{ for all } s \leqslant j, \quad s = 1, ..., T.$$

Specifically, Y_{ij} given $\mathbf{x}_{ij}$ does affect the future time-dependent covariate process, $\mathbf{x}_{i,j+1}$,...,$\mathbf{x}_{iT}$, but the previous covariates have no impact on future outcomes, and therefore the feed-back cycle is ruled out.

2.2.2 Existing Methods

Generalized Estimating Equations with Independence

Unbiased estimating equations can be obtained by using GEE with an independence working correlation structure regardless of the types of time-dependent covariates that are utilized within the marginal model. Therefore, this approach was advocated by Pepe and Anderson [4]. However, this safe approach may have a great loss in efficiency since the non-diagonal elements of $\boldsymbol{V}_i^{-1}$, $i = 1, \ldots, N$, are not used [5]. Specifically, information from the estimation equations $\partial\mu_{is}/\partial\beta_k(Y_{ij} - \mu_{ij})$, $s \neq j$, $i = 1, \ldots, N$, in the GEE approach is eliminated, and therefore ignoring these additional moment conditions, when valid, can result in a relative loss in efficiency with respect to estimation of parameters corresponding to Type I and II time-dependent covariates [1].

Generalized Method of Moments

Lai and Small [1] utilized GMM [7] in order to take advantage of all valid estimating equations. Specifically, they created a vector, $\boldsymbol{g}_i(\boldsymbol{\beta})$, comprised of all valid moment conditions from subject i, $i = 1, \ldots, N$, corresponding to the estimation of the p parameters such that $E[\boldsymbol{g}_i(\boldsymbol{\beta})] = \mathbf{0}$. With respect to the kth covariate, or $(k+1)$th parameter, the T^2 available moment conditions are $\partial\mu_{is}/\partial\beta_k(Y_{ij} - \mu_{ij})$, $j, s = 1, \ldots, T$, and only a subset are utilized for Type II-IV time-dependent covariates. There are T^2 valid moment conditions for a Type I time-dependent covariate or a time-independent covariate, $T(T+1)/2$ valid moment conditions for a Type II or IV time-dependent covariate, and T valid moment conditions for a Type III time-dependent covariate. To create $\boldsymbol{g}_i(\boldsymbol{\beta})$, all valid moments corresponding to each parameter are stacked such that the maximum length of $\boldsymbol{g}_i(\boldsymbol{\beta})$ is $T^2 \times p$.

Define

$$\bar{\boldsymbol{g}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\beta}).$$

This is used to create a quadratic form given by

$$\boldsymbol{Q}_N(\boldsymbol{\beta}) = N\bar{\boldsymbol{g}}_N^T(\boldsymbol{\beta})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\bar{\boldsymbol{g}}_N(\boldsymbol{\beta}) \tag{2.3}$$

in which $\boldsymbol{C}_N(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\beta})\boldsymbol{g}_i^T(\boldsymbol{\beta})$ is an empirical covariance matrix that is consistent for the optimal weighting matrix, $E[\boldsymbol{C}_N(\boldsymbol{\beta})] = \boldsymbol{\Sigma}_N = (1/N)\sum_{i=1}^{N} Cov[\boldsymbol{g}_i(\boldsymbol{\beta})]$. The GMM estimator, $\hat{\boldsymbol{\beta}}_{GMM}$, obtained by minimizing the quadratic form in Equation (2.3) asymptotically solves the estimating equations given by

$$N\dot{\boldsymbol{g}}_N^T(\boldsymbol{\beta})\boldsymbol{C}_N^{-1}(\boldsymbol{\beta})\bar{\boldsymbol{g}}_N(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.4}$$

in which $\dot{\boldsymbol{g}}_N(\boldsymbol{\beta}) = E[\partial\bar{\boldsymbol{g}}_N(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T]$. We note that the estimating equations in Equation (2.4) are asymptotically equivalent to the optimal estimating equations given by $N\dot{\boldsymbol{g}}_N^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_N^{-1}\bar{\boldsymbol{g}}_N(\boldsymbol{\beta}) = \mathbf{0}$ because $\boldsymbol{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$ [8, 19]. Optimality is with respect to minimizing the asymptotic variances of the regression parameter estimates out of all possible estimating equations which take linear combinations of $\bar{\boldsymbol{g}}_N(\boldsymbol{\beta})$ [20].

Using $\boldsymbol{C}_N$ in place of $\boldsymbol{\Sigma}_N$ in the estimating equations increases estimation variability, thus inflating $Cov(\hat{\boldsymbol{\beta}})$ relative to its theoretical value [14, 15]. As a result, the estimation performance of the GMM approach may not be as ideal as expected. Furthermore, if the number of parameters is large in a model, the total number of valid moment conditions might become large as well, particularly for a Type I time-dependent covariate that can utilize all valid moment conditions. This leads to some potential questions with respect to high dimensional and non-invertible issues [7, 19, 21], and increases estimation variability even further. As a result, the finite-sample validity and utility of inference with this approach can be questionable.

Modified Quadratic Inference Functions

The QIF method proposed by Qu *et al.* [8] is based on the GMM and GEE approaches. Rewrite $\boldsymbol{R}_i^{-1} = \sum_{r=1}^{m} \alpha_{ri}\boldsymbol{M}_{ri}$ in Equation (2.1), where $\boldsymbol{M}_{ri}$, $r = 1, \ldots, m$, $i = 1, \ldots, N$, are known basis matrices and α_{ri}, $r = 1, \ldots, m$, $i = 1, \ldots, N$, are functions of correlation parameters [8]. This rewrites GEE as a linear combination of m sets of unbiased estimating equations. For example, two basis matrices are typically used for exchangeable and AR-1 working structures. For both structures, $\boldsymbol{M}_{1i}$ is an identity matrix, whereas $\boldsymbol{M}_{2i}$ is a matrix with 0 on the diagonal and 1 elsewhere for exchangeable and $\boldsymbol{M}_{2i}$ is a matrix with 1 on the sub-diagonal and 0 elsewhere for AR-1.

Utilizing GMM, define

$$\bar{\boldsymbol{g}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\beta}) = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} \boldsymbol{g}_{1i}(\boldsymbol{\beta}) \\ \vdots \\ \sum_{i=1}^{N} \boldsymbol{g}_{mi}(\boldsymbol{\beta}) \end{bmatrix}, \tag{2.5}$$

where $\boldsymbol{g}_{ri}(\boldsymbol{\beta}) = \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_{ri} \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, \ldots, m$, $i = 1, \ldots, N$, and the estimation of correlation parameters, α_{ri}, is not necessary. We note that $\bar{\boldsymbol{g}}_N(\boldsymbol{\beta})$ here is defined differently than in the GMM approach of Lai and Small [1], and therefore optimality is not with respect to the same linear combination as in the GMM approach of Lai and Small [1]. Therefore, theoretical efficiencies can differ for these two approaches. However, regression parameter estimates are obtained by utilizing the same form for the estimating equations.

In order to utilize the QIF approach in the presence of time-dependent covariates, we have to ensure that $\bar{\boldsymbol{g}}_N(\boldsymbol{\beta})$, as given in Equation (2.5), only incorporates valid moment conditions, depending on the type of time-dependent covariate. Therefore, Zhou *et al.* [9], and similarly Cho and Dashnyam [22], modified the QIF and constrained $\boldsymbol{M}_{2i}$, denoted as $\boldsymbol{M}^*_{2i}$, to be a lower triangular matrix for a Type II time-dependent covariate, and therefore there are $T(T+1)/2$ estimating equations for $s \geqslant j$ to be used in $\boldsymbol{g}_{2i}(\boldsymbol{\beta})$. Take three time points, for example, such that $\boldsymbol{M}^*_{2i} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ for exchangeable and AR-1 working structures, respectively. For a Type IV time-dependent covariate, $\boldsymbol{M}^*_{2i} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ for exchangeable and AR-1, respectively. Only the identity matrix, $\boldsymbol{M}_{1i}$, is used for a Type III time-dependent covariate such that T valid estimating equations are utilized, and therefore $\boldsymbol{M}^*_{2i}$ is a 3 by 3 matrix of 0's for both structures. Additionally, no modifications to $\boldsymbol{M}_{2i}$ are needed for a Type I time-dependent covariate because all T^2 moment conditions are valid. These valid estimating equations then can be optimally combined using the GMM approach of Hansen [7].

As with the GMM approach of Lai and Small [1], finite sample covariance inflation occurs due to the use of $\boldsymbol{C}_N(\tilde{\boldsymbol{\beta}})$ in place of $\boldsymbol{\Sigma}_N$ within the estimating equations [16, 17, 23]. As a result, the small-sample estimation performance of the modified QIF approach may not be as ideal as expected. However, it typically will not have singularity issues as with the GMM approach of Lai and Small [1], although there are a few exceptions [19].

The asymptotic estimator for $Cov(\hat{\boldsymbol{\beta}})$ based on either the GMM or modified QIF approach is given by $(1/N)(\dot{\boldsymbol{g}}_N^T \boldsymbol{C}_N^{-1} \dot{\boldsymbol{g}}_N)^{-1} = (1/N)\boldsymbol{J}_N^{-1}$, in which the components of these formulas depend on which method is utilized. However, this formula does not account for the covariance inflation due to the use of $\boldsymbol{C}_N(\tilde{\boldsymbol{\beta}})$ in place of $\boldsymbol{\Sigma}_N$ in the estimating equations. After accounting for this covariance inflation, we have

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \frac{1}{N}(\boldsymbol{I}_p + \boldsymbol{G})\boldsymbol{J}_N^{-1}(\boldsymbol{I}_p + \boldsymbol{G})^T, \tag{2.6}$$

in which $\boldsymbol{G} = -\frac{\partial}{\partial \boldsymbol{\beta}^{*T}} \left[\boldsymbol{J}_N^{-1} \dot{\boldsymbol{g}}_N^T \boldsymbol{C}_N^{-1}(\boldsymbol{\beta}^*)\bar{\boldsymbol{g}}_N(\boldsymbol{\beta})\right] |_{\boldsymbol{\beta}^*=\boldsymbol{\beta}}$ [17]. However, the estimated empirical covariances, $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$ or $\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T$, $i = 1, \ldots, N$, can still be negatively biased for small sample size because the estimated residuals, $\hat{\boldsymbol{e}}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i$, $i = 1, \ldots, N$, are too small on average [24]. After utilizing a correction for this bias, such as the correction of Mancl and DeRouen [24], for the GMM approach of Lai and Small [1] and the modified QIF approach we propose estimating $Cov(\hat{\boldsymbol{\beta}})$ with

$$\widehat{\boldsymbol{\Sigma}}_{BC,QIF} = (1/N)(\boldsymbol{I}_p + \boldsymbol{G})\boldsymbol{J}_N^{-1}\dot{\boldsymbol{g}}_N^T \boldsymbol{C}_N^{-1}(\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}_N(\hat{\boldsymbol{\beta}})\boldsymbol{C}_N^{-1}(\tilde{\boldsymbol{\beta}})\dot{\boldsymbol{g}}_N \boldsymbol{J}_N^{-1}(\boldsymbol{I}_p + \boldsymbol{G})^T, \tag{2.7}$$

as proposed by Westgate [17] for use with the typical QIF approach.

2.3 Proposed Methods

2.3.1 Modified GMM

The GMM approach of Lai and Small [1] was proposed because it theoretically is equally or more efficient than GEE incorporating an independence working structure. However, as we will demonstrate in Supplemental Material, resulting standard error (SE) estimates can be biased, thus resulting in invalid inference, due to the fact that the empirical estimator, $\boldsymbol{C}_N(\boldsymbol{\beta})$, of the optimal weighting matrix in Equation (2.3) is singular because of the large number of moment conditions. The objective quadratic form (3) and its corresponding inference then becomes unobtainable. Additionally, we found the unique Moore-Penrose generalized inverse [25, 26] fails to solve the weighting matrix. Therefore, we propose to incorporate a linear shrinkage approach, originally proposed by Han and Song [19] to resolve potential singularity problems with QIF in special cases, with the GMM approach of Lai and Small [1]. This shrinkage approach theoretically leads to a consistent estimator and has the same efficiency, but avoids singularity, thus allowing appropriate SE estimates to be obtained and valid inference to be attained.

In short, we replace the original $\boldsymbol{C}_N(\boldsymbol{\beta})$ of the GMM approach with $\boldsymbol{S}_N(\boldsymbol{\beta}) = \rho_N\mu_N\boldsymbol{I} + (1-\rho_N)\boldsymbol{C}_N(\boldsymbol{\beta})$, in which μ_N is the mean of the diagonal elements of $\boldsymbol{\Sigma}_N$, $\boldsymbol{I}$ is the identity matrix, and ρ_N can be obtained by minimizing $E[||\boldsymbol{S}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$. Formulas for estimates of ρ_N and μ_N can be obtained from [19]. Furthermore, a bias-corrected estimate for $Cov(\hat{\boldsymbol{\beta}})$ can be obtained by modifying Equation (2.7), using $\boldsymbol{S}_N^{-1}$ in place of $\boldsymbol{C}_N^{-1}$.

2.3.2 Modified GEE

Due to the popularity of GEE and the limitations of the previously discussed methods, we propose a modified GEE approach in which elements in the inverse of the working correlation matrix are replaced with 0 whenever their use will yield biased equations. Replacement is done for each individual estimating equation, depending on the type of covariate. Specifically, our proposed estimating equation for β_k, $k = 0, 1, \ldots, p-1$, is given by

$$\sum_{i=1}^{N} \boldsymbol{D}_i^{k+1}\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_{ik}^{*-1}\left(\boldsymbol{\alpha}\right)\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.8}$$

where $\boldsymbol{D}_i^{k+1}$ is the $(k+1)$th row of $\boldsymbol{D}^T$, and the elements of $\boldsymbol{R}_{ik}^{*-1}$ depend on the covariate type. The modified GEE approach then stacks these individual estimating equations, and regression parameter, correlation parameter, and SE estimation are done in the same manner as with GEE. We note that small-sample SE corrections [27, 28], such as the ones discussed for the modified QIF and GMM approaches, can be applied with GEE and thus our modified GEE approach as well.

We propose to construct $\boldsymbol{R}_{ik}^{*-1}$, $k = 0, 1, \ldots, p-1$, given in Equation (2.8) by modifying $\boldsymbol{R}_i^{-1}$, the inverse of any given working correlation structure used in Equation (2.1), according to the specific type of time-dependent covariate. Specifically, if parameter k corresponds to a Type I time-dependent or time-independent covariate, then all T^2 moment conditions are valid and therefore $\boldsymbol{R}_{ik}^{*-1} = \boldsymbol{R}_i^{-1}$, implying that the estimating equation is the same for GEE and our modified GEE. With respect to the estimating equation for the parameter corresponding to a Type II time-dependent covariate, $\boldsymbol{R}_{ik}^{*-1}$ is restricted to be a lower triangular matrix such that the information from the $T(T+1)/2$ valid moment conditions for $s \geqslant j$ is included. Specifically, $\boldsymbol{R}_{ik}^{*-1}$ is obtained by taking $\boldsymbol{R}_i^{-1}$ and making all upper non-diagonal elements equal to 0. The opposite is done with respect to a Type IV time-dependent covariate, such that $\boldsymbol{R}_{ik}^{*-1}$ is obtained by taking $\boldsymbol{R}_i^{-1}$ and making all lower non-diagonal elements equal to 0. Finally, $\boldsymbol{R}_{ik}^{*-1}$ is an identity matrix in the estimating equation when the parameter corresponds to a Type III time-dependent covariate. We note that with our modified GEE, the working structure is technically no longer a true working correlation structure because some non-zero elements of $\boldsymbol{R}_i^{-1}$ are replaced with 0 such that invalid moment conditions are not utilized, and therefore $\boldsymbol{R}_{ik}^{*-1}$ will not be the inverse of a true working correlation matrix when β_k corresponds to a Type II or IV time-dependent covariate.

2.3.3 Analysis Method Selection

We have discussed multiple analysis methods to use in the presence of time-dependent covariates: GEE with independence, the GMM approach of Lai and Small [1] or our modified GMM approach to ensure valid inference, a modified QIF approach, and our modified GEE. Unfortunately, none of these methods are guaranteed to always perform best; i.e., produce the least variable regression parameter estimates. The modified GMM and QIF approaches are both efficient but are with respect to different optimalities, and therefore one is not guaranteed to outperform the other. Furthermore, finite-sample inflations of the variances of regression parameter estimates, relative to the theoretical asymptotic variances, occurs with both of these methods. As a result, use of our modified GEE may result in smaller realized variances of regression parameter estimates. Therefore, an approach to select a single method, with corresponding working correlation structure if applicable, is needed.

Our goal is to choose an analysis method and corresponding structure that results in the least variable regression parameter estimates. To do so, we take a similar approach to Westgate [18] who proposed the use of the TECM [29] to choose between the typical QIF and GEE approaches. Although the TECM is simple to obtain by summing up the estimates of the variances of the parameter estimates, a potential disadvantage in practice is that the variance(s) of any given parameter(s) might dominate the overall criterion value. Therefore, we will extend the CIC [10] for use in our setting, as it has found popularity in the GEE correlation structure selection literature. Specifically, let $\widehat{\boldsymbol{\Sigma}}_{BC}$ denote our finite-sample corrected estimate of $Cov(\hat{\boldsymbol{\beta}})$ for any given method under consideration for selection. The CIC value for that particular method is given by $tr(\widehat{\boldsymbol{\Sigma}}_I^{-1}\widehat{\boldsymbol{\Sigma}}_{BC})$, in which $\widehat{\boldsymbol{\Sigma}}_I = (\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1} \boldsymbol{D}_i)^{-1}$. The single method, with corresponding working correlation structure if applicable, that yields the smallest CIC value is then selected for conducting inference.

2.4 Simulation Study

2.4.1 Study Description

We now compare the finite-sample regression parameter estimation performances of GEE incorporating the independence working correlation structure, the modified QIF, and our modified GEE when time-dependent covariates are presented, in addition to assessing how well the CIC works in terms of selecting a single method and corresponding structure. Five modeling options regarding the combinations of analysis approaches and working structures are GEE with an independence working correlation structure, and combinations of either modified GEE or QIF approach with an exchangeable or AR-1 working structure. We note that our modified GMM approach performed poorly in terms of regression parameter estimation due to weight being assigned to an identity matrix, which can be inefficient for parameter estimation, and therefore we initially do not consider it for selection for results presented within this chapter. However, in Supplementary Material, we do consider it for selection and we include results from the use of this approach and the original GMM approach of Lai and Small [1] to show their relatively poor finite-sample performances in terms of estimation and validity of inference.

Each setting of our simulation study consists of either 25, 50, 100, or 500 subjects representing small (25/50), moderate (50/100), and large (100/500) sample sizes. Each subject contributes 5 or 15 repeated measurements at the same time points. Each setting is conducted through 1000 simulations using R version 3.1.2 [30]. Furthermore, we utilize two scenarios motivated by the time-dependent covariate literature.

Scenario 1 comes from Diggle *et al.* [31], which is also used by previous studies [1, 9, 11], and uses one Type II time-dependent covariate such that $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{1i,j-1} + \gamma_i + \epsilon_{ij}$, and $x_{1ij} = \kappa x_{1i,j-1} + e_{ij}$, $j = 1, \ldots, 5$ or 15, where $\boldsymbol{\beta} = [0, 1, 1]^T$ and random effects, γ_i, ϵ_{ij}, and e_{ij}, are mutually independent and normally distributed with mean 0 and variance 1. We note that $Var(e_{ij}) = \sigma_e^2$. Additionally, because the time process, x_{1ij}, is stationary, x_{1i0} follows a normal distribution with mean 0 and variance $\sigma_e^2/(1-\kappa^2)$. Here we let $\kappa = 0.5$. The marginal mean is given by $E[Y_{ij}|x_{1ij}] = \beta_0 + (\beta_1 + \kappa\beta_2)x_{1ij}$, which gives true values of $\beta_0 = 0$ for the marginal intercept and $\beta_1 + \kappa\beta_2 = 1.5$ for the marginal parameter corresponding to the time-dependent covariate. The covariance structures in Tables 2.1 and 2.3 are constructed via the random effects in the above data generating model, and thus the working correlation structures utilized in this study are all misspecified, whereas the structures in Tables 2.2 and 2.4 are constructed by eliminating the random effects and generating data using a true AR-1 correlation structure. We note that the derivation of this marginal mean can be found in Web-based Supplementary Materials of Leung *et al.* [11].

Scenario 2 extends scenario 1 by adding two additional types of covariates. Specifically, the marginal model now includes a time-independent binary indicator covariate, which utilizes the same moment conditions as a Type I time-dependent covariate, and a Type I time-dependent covariate corresponding to time itself. Therefore, data are generated from $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{1i,j-1} + \beta_3 x_{2i} + \beta_4 x_{3ij} + \gamma_i + \epsilon_{ij}$, where $\boldsymbol{\beta} = [0, 1, 1, 1.5, 0.5]^T$, x_{2i}, $i = 1, \ldots, N$, are independently sampled from $Bernoulli(0.6)$, and $x_{3ij} = j$, $i = 1, \ldots, N$, $j = 1, \ldots, 5$ or 15. The marginal mean becomes $E[Y_{ij}|x_{1ij}] = \beta_0 + (\beta_1 + \kappa\beta_2)x_{1ij} + \beta_3 x_{2i} + \beta_4 x_{3ij}$, and therefore the true marginal regression parameter values are $\beta_0 = 0$, $\beta_1 + \kappa\beta_2 = 1.5$, $\beta_3 = 1.5$, and $\beta_4 = 0.5$. The true covariance structures are the same as in scenario 1.

In order to assess the differences in estimation performances of the five modeling options, we present ratios of empirical mean squared error (MSE) from non-intercept parameters, which we refer to as relative efficiencies (REs), in Tables 2.1-2.4. For any given RE, the numerator is the MSE from the use of GEE with an independence working structure and the denominator is the MSE for the given method. The modeling option that performs best therefore has the largest ratio. In order to assess the utility of the CIC, the number of times each method with corresponding structure is selected out of 1,000 simulations are given in the tables, and REs as previously defined are also shown.

2.4.2 Results

When incorporating only one Type II time-dependent covariate in the settings of scenario 1 and the true structure was constructed by random effects (see Table 2.1), the REs indicated that our proposed modified GEE approach performed best overall in terms of regression parameter estimation for $N \leq 50$. However, in some settings when $N \geq 100$, the modified QIF approach did demonstrate an efficiency advantage. The reason for these findings is that the smaller N is, the greater the finite-sample variance inflation that occurs with the modified QIF approach, thus allowing our modified GEE approach to often work just as well or better. However, when the working and true structures were AR-1 (see Table 2.2), the modified QIF was able to outperform the modified GEE for $N \geq 50$. The reason for this efficiency gain is because the modified GEE and QIF technically do not utilize working correlation matrices after setting elements equal to 0 in order to ensure only valid moment conditions are utilized, and therefore QIF is theoretically more efficient than GEE when the working structure is not the true correlation structure. We note that neither the GMM approach of Lai and Small [1] nor our modified GMM approach performed well in these settings, as presented in Supplemental Material.

In scenario 2 (see Tables 2.3 and 2.4), RE results corresponding to the Type II time-dependent covariate were similar to results observed in scenario 1. Furthermore, results with respect to time (Type I) and the time-independent covariate were similar. Under the true structure constructed by random effects (Table 2.3), GEE with independence, the modified GEE and the modified QIF with either working structure all produced similar REs regarding the both covariate types. However, the modified QIF did not work as well when $N \leq 50$. Alternatively, when the true structure was AR-1 (Table 2.4), our modified GEE with AR-1 working structure worked best overall, with the exception of the setting of $N = 500$. Specifically, the modified GEE with AR-1 working structure resulted in the largest REs corresponding to time and the time-independent covariate. Furthermore, results in Supplementary Material showed that both the GMM and modified GMM approaches had relatively poor performances in all settings. We note that when the true structure was formed via random effects (Table 2.3) and the number of repeated measurements, T, was increased to 15, the

REs became larger even when the sample size was small. Similarly, as the number of repeat measurements was increased to 15, the REs increased when the true and working structures were AR-1 (Table 2.4).

As desired, the CIC tended to select most often the approach that resulted in the greatest REs for parameter estimates, and therefore the CIC performed well in terms of its resulting RE. When the true covariance structure was defined by random effects, the modified GEE and QIF approaches incorporating the exchangeable working structure were chosen most often in either scenario. Although the CIC seemed to have overselected the modified QIF approach when $N \leqslant 100$ in some settings when the true structure was constructed by random effects in either scenario, the differences were not notable overall and might be due to random error (Tables 2.1 and 2.3). When the true correlation structure was AR-1, the CIC was most likely to select the modified GEE or QIF approach incorporating the AR-1 working structure (Tables 2.2 and 2.4). We note that when $N \leqslant 50$ in scenario 1 (Table 2.2) and when $N = 50$ or 100 and $T = 15$ in scenario 2 (Table 2.4), the modified GEE with AR-1 structure was chosen more frequently than the modified QIF with AR-1 structure, although the latter method resulted in slightly larger REs corresponding to the Type II time-dependent covariate. The reason for this is because the REs corresponding to the intercept (results not shown), time, and the time-independent covariate were notably greater for the modified GEE with AR-1 and thus had greater influence on the CIC. However, the CIC still worked well in such situations in terms of its RE with regard to the Type II time-dependent covariate.

In summary, the modified GEE approach worked well for $N \leq 50$, while the modified QIF tended to work best when $N \geq 100$. However, realistically the data analyst will not know for sure which of these two methods will perform best for the analysis of any given dataset. Therefore, we proposed the use of the CIC to help to determine which method, and working structure, should be utilized. The CIC was found to perform quite well in terms of REs. Specifically, with respect to time and the time-independent covariate, REs resulting from the use of the CIC were close to 1 when the true structure was constructed by random effects, whereas the REs were often greater than 1 when the true structure was AR-1, especially with regard to time when $T = 15$. Furthermore, REs corresponding to the Type II time-dependent covariate showed that the CIC worked very well in each scenario. Specifically, the use of the CIC was notably better than sole use of GEE with independence, and for any given setting the CIC typically performed similarly to the sole use of the best method and working structure combination(s).

2.5 Application

We now compare our proposed approaches with the existing methods using data from the study of anthropometric screenings among children [32, 33]. In this study, the target is to explore the association between anthropometric covariates at a given survey time point and morbidity outcomes in the future. The data were originally collected from 1984 to 1985, obtaining survey information from 448 households [32]. Lai and Small [1] used a subset of data containing 370 children ($\leqslant$ 14 years) from Bhargava [33]. This data consists of repeated measurements from each child at three time points with four months between each subsequent measurement. Children with incomplete data were excluded, and only one child per household was chosen, which eliminates the need to account for statistical correlation due to household clustering [33].

We utilize the marginal model used by Lai and Small [1], Lueng *et al.* [11], and Zhou *et al.* [9] that is given by

$$\mu_{ij} = \beta_0 + \beta_1 BMI_{ij} + \beta_2 Age_{ij} + \beta_3 Female_i + \beta_4 SR2_{ij} + \beta_5 SR3_{ij}, \quad j = 1, 2, 3,$$

where μ_{ij} is the ith child's marginal mean morbidity index during the jth four-month interval. We note that the morbidity index is given by

$$y_{ij} = \log\left(\frac{\text{days child was sick in last 2 weeks prior to time } j + 0.5}{14.5 - \text{days child was sick in last 2 weeks prior to time } j}\right),$$

adopting the same logistic transformation made by Bhargava [33] and Lai and Small [1]. Age and the indicators for survey rounds 2 and 3, which are used to account for seasonality in morbidity, are Type I time-dependent covariates. Female indicator is a time-independent covariate. Furthermore, we treat body mass index (BMI) as a Type II time-dependent covariate based on the hypothesis testing done by Lai and Small [1].

As in the simulation study, we analyze this dataset using GEE with a working independence structure, our proposed modified GEE, and the modified QIF approach. The original and modified GMM approaches are excluded because of their low precision in estimation. Table 2.5 gives regression parameter estimates, bias-corrected empirical SE estimates, and CIC values. We note that the empirical SE estimates resulting from the use of GEE with independence are notably different from SE estimates presented in Lai and Small [1] and Zhou *et al.* [9], as these manuscripts utilized model-based SE estimates which are not valid unless independence truly is the correct structure.

The CIC values indicate that our proposed modified GEE approach is preferable for the analysis of this particular dataset (see Table 2.5). Furthermore, the CIC value is smallest for the modified GEE with AR-1 working structure, and hence this particular method and working structure is preferable. We note, however, that all the approaches actually produce similar results in terms of regression parameter and SE estimates, with some slight exceptions with the modified QIF approach. The reason for this is because the correlation among repeated measurements of morbidity outcomes is small, as explained by Lai and Small [1]. The estimated correlation parameters, $\hat{\alpha}$, from the modified GEE approach, as indicated at the bottom of Table 2.5, are presented to express the small correlation.

2.6 Concluding Remarks

When certain types of time-dependent covariates are included in a marginal model, the estimating equations used by GEE may be biased, thus resulting in biased regression parameter estimates unless the independence working correlation structure is used. However, GEE incorporating independence can be inefficient because not all valid moment conditions are utilized for the estimation of regression parameters corresponding to Type I and II time-dependent covariates. Therefore, GMM and modified QIF approaches that utilize all valid moment conditions have been proposed to improve efficiency. However, we found that these approaches may result in invalid inference (results in Supplementary Material). To improve upon the validity of inference with the GMM approach, we developed a modified, non-singular weighting matrix. Unfortunately, this modified GMM approach did not perform well in terms of regression parameter estimation (results in Supplementary Material), and therefore we do not advocate its use in practice. Furthermore, we applied previously developed small-sample corrections to estimators of the covariance matrix of regression parameter estimates. More notably, in order to improve regression parameter estimation while still attaining valid inference, we proposed a modified GEE approach which is meant to potentially improve upon the performance of the modified QIF when the number of subjects is not large. Which combination of method and working structure will result in the smallest variances of regression parameter estimates will be unknown to the data analyst, and therefore we extended the applicability of the CIC in order to select a combination under consideration. In short, the proposed modified GEE often outperformed all other methods that have been proposed for the marginal analysis of longitudinal data in the presence of time-dependent covariates. However, the modified QIF did perform best, in terms of estimating the regression parameter corresponding to a Type II time-dependent covariate, in some large-sample settings in our simulation study due to its theoretical efficiency advantage. Furthermore, the

CIC performed well in terms of selecting the best method and structure combinations and thus regression parameter estimation.

Although we only used working independence, exchangeable, and AR-1 correlation structures in our studies, other working structures are available as well, including less parsimonious Toeplitz forms and unstructured working matrices. We note that the modified conjugate gradient QIF approach of Zhou *et al.* [9] essentially assumes an unstructured working structure, and finite-sample bias corrections for the empirical covariance matrix of regression parameter estimates have been proposed for the original approach which can easily be incorporated with this modified approach [34]. Although the unstructured form can be included with the modified QIF [9, 17, 21, 34], other working structures cannot be used due to the need for $\boldsymbol{R}_i^{-1} \approx \sum_{r=1}^{m} \alpha_{ri} \boldsymbol{M}_{ri}$. Therefore, our modified GEE approach has an additional advantage in terms of its flexibility with respect to the working structures it can implement. We note with both the modified QIF and GEE, estimating equations corresponding to Type II and IV time-dependent covariates only utilize valid moment conditions. As a result, the working structure is technically no longer an actual correlation structure because some non-zero elements are replaced with 0 such that invalid moment conditions are not utilized.

The GMM approach of Lai and Small [1] was previously shown to have unreliable inference if a large number of moment conditions are used [35]. Lai and Small [1] empirically demonstrated that the coverage probability of a confidence interval for a parameter corresponding to a Type II time-dependent covariate, for instance, has a notable decline from the nominal 0.95 level when the number of moment conditions increases, which can occur, for instance, as the number of time points increases. To correct for this type of invalid inference, we utilized small-sample corrections. However, the novel finding in our chapter is that we found coverage probabilities can be non-nominal due to singularity of the empirical weighting covariance matrix, thus motivating us to propose the modified GMM approach.

In this chapter, we assumed the analyst knows the type of time-dependent covariate(s). However, in practice this may not be the case. In such situations, a conservative, but safe, approach would be to treat unknown types of time-dependent covariates as Type III in order to ensure that only valid moment conditions are utilized. An alternative option would be to conduct a test. To assess a specific type of time-dependent covariate, Lai and Small [1] proposed a hypothesis test to examine the validity of moment conditions using the GMM approach. However, due to the inadequacies we found with this approach, we feel further study is warranted. Additionally, Lalonde *et al.* [2] proposed an alternative testing approach for assessing the validity of moment conditions based on tests of correlation between moment conditions at different time points.

Our simulation study and application example utilized marginal models for continuous responses. However, the methodology addressed in this chapter is also applicable to marginal generalized linear models with different link functions, and unbalanced repeated measurements are permittable. Due to the added complexity of data generation with respect to time-dependent covariates, future research is needed on the simulation of such data for any type of outcome.

2.7 Supplementary Material

We now present simulation results that supplement the results presented in the chapter. We also study the GMM approach of Lai and Small [1] and our modified version of their GMM approach. In Table 2.6 we show that all studied methods, with the exception of the GMM approach, result in near-nominal inference. In Table 2.7 we further assess the invalidity of inference corresponding to the GMM approach. Finally, in Table 2.8 we demonstrate the loss in efficiency when using our modified version of the GMM approach relative to the other valid approaches. For conciseness of presentation and to avoid convergence issues resulting from the GMM approach when the number of repeated measurements is 15, we only present results for $T = 5$ corresponding to the settings of Scenario 1, as described in the chapter, for which the marginal model incorporates a Type II time-dependent covariate.

In Table 2.6, empirical coverage probabilities (CPs) of 95% confidence intervals are given for the five modeling options studied in the chapter, as well as the GMM approach and our modified GMM approach. Regardless of the given true covariance structure, the empirical CPs for the original GMM approach are low, and once the CP was even as small as 0.265. We note that CPs with this approach reduced with the number of subjects. On the other hand, CPs corresponding to all other methods are close to 0.95 regardless of the working structure and number of subjects.

Table 2.7 demonstrates that the invalidity of the GMM approach is due to biased standard error (SE) estimation, and to our knowledge this is the first study to actually assess the validity of SE estimation with this approach. Specifically, Table 2.7 presents empirical standard deviations (ESDs) of $\hat{\boldsymbol{\beta}}_1$ and empirical means of corresponding SE estimates. Ideally the empirical means of SE estimates should be similar to the corresponding ESDs. However, as the number of subjects decreased, the amount of bias in the SE estimates increases, as can be seen via the difference in ESDs and empirical mean SEs. As a result, empirical CPs were notably influenced. We again note that the reason for this bias in the SE estimates is due to the singularity of the empirical weighting covariance matrix, $\boldsymbol{C}_N$, because of the use of numerous moment conditions.

In Table 2.8, we present the relative efficiencies (REs) and correlation information criterion (CIC) selection frequencies of our modified GMM approach along with the five modeling options used within the chapter's simulation study. The REs show that our proposed modified GMM approach using linear shrinkage does not perform well with respect to regression parameter estimation. The reason for this result is that the empirical covariance matrix of this approach, although asymptotically optimal, is no longer necessarily optimal in finite-sample settings due to the need to assign weight to an identity matrix. Therefore, although we proposed this method such that valid inference can be attained, we do not advocate its use in practice. However, we do note that the CIC very rarely chose the modified GMM approach due to its poor performance, and therefore considering it for selection typically was not detrimental.

Table 2.1: Results from scenario 1 for settings in which one Type II time-dependent covariate is utilized. True structure is constructed by random effects.

			CIC	GEE	Modified GEE		Modified QIF	
T	N			Ind	Exch	AR-1	Exch	AR-1
5	25	RE	1.10	1.00	1.18	1.11	1.08	1.05
		CIC Selection Frequencies		48	288	129	326	209
	50	RE	1.13	1.00	1.16	1.10	1.13	1.06
		CIC Selection Frequencies		26	323	73	391	187
	100	RE	1.18	1.00	1.21	1.12	1.19	1.10
		CIC Selection Frequencies		5	351	37	438	169
	500	RE	1.17	1.00	1.18	1.10	1.18	1.11
		CIC Selection Frequencies		0	327	0	625	48
15	25	RE	1.30	1.00	1.42	1.15	1.42	1.11
		CIC Selection Frequencies		19	377	39	367	198
	50	RE	1.37	1.00	1.42	1.15	1.40	1.17
		CIC Selection Frequencies		3	411	25	385	176
	100	RE	1.34	1.00	1.39	1.17	1.42	1.20
		CIC Selection Frequencies		0	394	1	509	96
	500	RE	1.51	1.00	1.48	1.17	1.56	1.24
		CIC Selection Frequencies		0	270	0	728	2

T - number of repeated measurements; N - number of independent subjects;
Ind - independence; Exch - exchangeable; CIC - correlation information criterion;
GEE - generalized estimating equations; QIF - quadratic inference function;
RE - relative efficiency. For each setting, they compare the empirical
mean squared error (MSE) from the use of the GEE with independence structure
to the MSEs from the use of different modeling options or CIC;
CIC Selection - Number of times out of 1,000 simulations that CIC selected the given
method and corresponding structure.

Table 2.2: Results from scenario 1 for settings in which one Type II time-dependent covariate is utilized. True structure is AR-1.

			CIC	GEE	Modified GEE		Modified QIF	
T	N			Ind	Exch	AR-1	Exch	AR-1
5	25	RE	1.32	1.00	1.25	1.36	1.20	1.32
		CIC Selection Frequencies		17	92	471	124	296
	50	RE	1.30	1.00	1.27	1.36	1.24	1.38
		CIC Selection Frequencies		4	73	439	90	394
	100	RE	1.35	1.00	1.26	1.36	1.23	1.45
		CIC Selection Frequencies		1	23	458	29	489
	500	RE	1.41	1.00	1.27	1.34	1.28	1.43
		CIC Selection Frequencies		0	0	96	0	904
15	25	RE	1.38	1.00	1.10	1.43	1.07	1.46
		CIC Selection Frequencies		9	36	496	98	361
	50	RE	1.40	1.00	1.08	1.42	1.05	1.48
		CIC Selection Frequencies		0	6	512	30	452
	100	RE	1.55	1.00	1.11	1.49	1.10	1.67
		CIC Selection Frequencies		0	0	414	2	584
	500	RE	1.65	1.00	1.12	1.49	1.13	1.66
		CIC Selection Frequencies		0	0	71	0	929

T - number of repeated measurements; N - number of independent subjects;
Ind - independence; Exch - exchangeable; CIC - correlation information criterion;
GEE - generalized estimating equations; QIF - quadratic inference function;
RE - relative efficiency. For each setting, they compare the empirical
mean squared error (MSE) from the use of the GEE with independence structure
to the MSEs from the use of different modeling options or CIC;
CIC Selection - Number of times out of 1,000 simulations that CIC selected the given
method and corresponding structure.

Table 2.3: Results from scenario 2 for settings in which one time-independent, one Type I, and one Type II time-dependent covariate are utilized. True structure is constructed by random effects.

			CIC	GEE	Modified GEE		Modified QIF	
T	N			Ind	Exch	AR-1	Exch	AR-1
5	25	RE of Time-Independent	0.97	1.00	1.00	1.00	0.93	0.86
		RE of Type I Covariate	0.97	1.00	1.00	0.96	0.95	0.87
		RE of Type II Covariate	1.11	1.00	1.15	1.09	1.07	0.98
		CIC Selection Frequencies		49	257	169	322	203
	50	RE of Time-Independent	0.98	1.00	1.00	0.98	0.97	0.93
		RE of Type I Covariate	0.98	1.00	1.00	0.97	0.97	0.92
		RE of Type II Covariate	1.16	1.00	1.20	1.13	1.15	1.05
		CIC Selection Frequencies		37	320	88	371	184
	100	RE of Time-Independent	0.99	1.00	1.00	0.98	0.99	0.96
		RE of Type I Covariate	1.00	1.00	1.00	0.97	0.99	0.98
		RE of Type II Covariate	1.18	1.00	1.21	1.13	1.18	1.08
		CIC Selection Frequencies		16	366	45	419	154
	500	RE of Time-Independent	1.00	1.00	1.00	0.98	0.99	0.99
		RE of Type I Covariate	1.00	1.00	1.00	0.95	1.00	1.00
		RE of Type II Covariate	1.17	1.00	1.17	1.11	1.17	1.13
		CIC Selection Frequencies		0	363	1	576	60
15	25	RE of Time-Independent	0.97	1.00	1.00	1.00	0.96	0.87
		RE of Type I Covariate	0.98	1.00	1.01	0.96	0.97	0.88
		RE of Type II Covariate	1.28	1.00	1.45	1.15	1.38	1.07
		CIC Selection Frequencies		19	366	69	346	200
	50	RE of Time-Independent	1.00	1.00	1.00	1.00	0.97	0.94
		RE of Type I Covariate	1.00	1.00	1.00	0.93	0.96	0.94
		RE of Type II Covariate	1.37	1.00	1.47	1.16	1.46	1.10
		CIC Selection Frequencies		2	413	37	367	181
	100	RE of Time-Independent	1.00	1.00	1.00	1.00	0.98	0.97
		RE of Type I Covariate	1.00	1.00	1.00	0.94	0.98	0.96
		RE of Type II Covariate	1.37	1.00	1.39	1.15	1.43	1.13
		CIC Selection Frequencies		1	420	12	449	118
	500	RE of Time-Independent	0.99	1.00	1.00	0.99	0.99	0.99
		RE of Type I Covariate	1.00	1.00	1.00	0.95	1.00	0.99
		RE of Type II Covariate	1.51	1.00	1.48	1.17	1.54	1.21
		CIC Selection Frequencies		0	342	0	642	16

T - number of repeated measurements; N - number of independent subjects;
Ind - independence; Exch - exchangeable; CIC - correlation information criterion;
GEE - generalized estimating equations; QIF - quadratic inference function;
RE - relative efficiency. For each setting, they compare the empirical
mean squared error (MSE) from the use of the GEE with independence structure
to the MSEs from the use of different modeling options or CIC;
CIC Selection - Number of times out of 1,000 simulations that CIC selected the given method and corresponding structure.

Table 2.4: Results from scenario 2 for settings in which one time-independent, one Type I, and one Type II time-dependent covariate are utilized. True structure is AR-1.

			CIC	GEE	Modified GEE		Modified QIF	
T	*N*			Ind	Exch	AR-1	Exch	AR-1
5	25	RE of Time-Independent	1.00	1.00	1.00	1.04	0.97	0.96
		RE of Type I Covariate	1.03	1.00	1.00	1.05	0.97	0.91
		RE of Type II Covariate	1.32	1.00	1.27	1.35	1.23	1.25
		CIC Selection Frequencies		13	79	551	141	216
	50	RE of Time-Independent	1.02	1.00	1.00	1.03	0.97	0.95
		RE of Type I Covariate	1.03	1.00	1.00	1.05	0.99	0.97
		RE of Type II Covariate	1.34	1.00	1.33	1.38	1.30	1.35
		CIC Selection Frequencies		2	61	596	96	245
	100	RE of Time-Independent	1.05	1.00	1.00	1.08	0.98	1.02
		RE of Type I Covariate	1.03	1.00	1.00	1.05	0.99	1.00
		RE of Type II Covariate	1.29	1.00	1.26	1.32	1.23	1.32
		CIC Selection Frequencies		0	25	598	54	323
	500	RE of Time-Independent	1.06	1.00	1.00	1.06	0.99	1.06
		RE of Type I Covariate	1.05	1.00	1.00	1.05	1.00	1.04
		RE of Type II Covariate	1.40	1.00	1.27	1.34	1.27	1.45
		CIC Selection Frequencies		0	0	369	0	631
15	25	RE of Time-Independent	1.03	1.00	1.00	1.06	0.96	0.92
		RE of Type I Covariate	1.08	1.00	1.00	1.13	0.95	0.98
		RE of Type II Covariate	1.35	1.00	1.12	1.45	1.06	1.35
		CIC Selection Frequencies		12	31	620	89	248
	50	RE of Time-Independent	1.06	1.00	1.00	1.07	0.97	1.03
		RE of Type I Covariate	1.15	1.00	1.00	1.16	0.99	1.10
		RE of Type II Covariate	1.38	1.00	1.09	1.42	1.06	1.46
		CIC Selection Frequencies		2	12	673	32	281
	100	RE of Time-Independent	1.05	1.00	1.00	1.07	0.99	1.04
		RE of Type I Covariate	1.14	1.00	1.00	1.17	0.99	1.09
		RE of Type II Covariate	1.50	1.00	1.16	1.48	1.13	1.54
		CIC Selection Frequencies		0	0	661	2	337
	500	RE of Time-Independent	1.05	1.00	1.00	1.05	1.00	1.05
		RE of Type I Covariate	1.12	1.00	1.00	1.13	1.00	1.12
		RE of Type II Covariate	1.53	1.00	1.14	1.50	1.13	1.67
		CIC Selection Frequencies		0	0	359	0	641

T - number of repeated measurements; *N* - number of independent subjects;
Ind - independence; Exch - exchangeable; CIC - correlation information criterion;
GEE - generalized estimating equations; QIF - quadratic inference function;
RE - relative efficiency. For each setting, they compare the empirical mean squared error (MSE) from the use of the GEE with independence structure to the MSEs from the use of different modeling options or CIC;
CIC Selection - Number of times out of 1,000 simulations that CIC selected the given method and corresponding structure.

Table 2.5: Parameter estimates, bias-corrected standard error estimates (in parentheses), and CIC values resulting from analyses of the anthropometric screening dataset.

	GEE	Modified GEE		Modified QIF	
Covariate	Independence	Exch	AR-1	Exch	AR-1
BMI	-0.06 (0.05)	-0.05 (0.06)	-0.05 (0.05)	-0.08 (0.05)	-0.07 (0.05)
Age	-0.01 (0.003)	-0.01 (0.004)	-0.01 (0.003)	-0.01 (0.004)	-0.01 (0.003)
Gender	0.15 (0.11)	0.15 (0.11)	0.14 (0.11)	0.13 (0.11)	0.15 (0.11)
SR 2	-0.28 (0.11)	-0.28 (0.11)	-0.28 (0.11)	-0.26 (0.11)	-0.31 (0.11)
SR 3	0.02 (0.13)	0.02 (0.13)	0.03 (0.13)	0.05 (0.13)	0.01 (0.13)
CIC	7.05	5.87	5.60	7.06	7.00
$\hat{\alpha}$		0.12	0.15		

GEE - generalized estimating equations; QIF - quadratic inference function; Exch - exchangeable; SR - survey round; CIC - correlation information criterion; $\hat{\alpha}$ - estimated correlation parameter.

Table 2.6: Empirical coverage probabilities of 95% confidence intervals covering β_1 from the settings of scenario 1 and $T = 5$.

		GEE	Modified GEE		Modified QIF		GMM	
Structure	N	Independence	Exch	AR-1	Exch	AR-1	Lai & Small	Modified
1	25	0.959	0.954	0.953	0.954	0.958	0.336	0.954
	50	0.959	0.962	0.960	0.960	0.954	0.778	0.952
	100	0.949	0.951	0.947	0.946	0.942	0.829	0.945
	500	0.953	0.954	0.956	0.952	0.952	0.925	0.961
2	25	0.952	0.953	0.953	0.946	0.956	0.265	0.946
	50	0.947	0.955	0.953	0.953	0.951	0.793	0.945
	100	0.949	0.954	0.948	0.954	0.957	0.822	0.953
	500	0.958	0.957	0.954	0.956	0.955	0.922	0.955

Structure 1 - true structure is constructed by random effects,
Structure 2 - true structure is AR-1;
N - number of independent subjects;
GEE - generalized estimating equations; QIF - quadratic inference function;
GMM - generalized method of moments; Exch - exchangeable;
Lai & Small - the GMM approach of of Lai and Small [1];
Modified - the proposed modified GMM approach.

Table 2.7: Empirical standard deviations (ESDs) of $\hat{\beta}_1$ and empirical mean standard error (SE) estimates, along with empirical coverage probabilities (CPs) of 95% confidence intervals covering β_1, from the settings of scenario 1 and $T = 5$ for the GMM approach of Lai and Small [1].

Structure	N	ESD	Empirical Mean SE	CP
1	25	0.292	0.057	0.336
	50	0.119	0.075	0.778
	100	0.085	0.062	0.829
	500	0.033	0.031	0.925
2	25	0.185	0.029	0.265
	50	0.069	0.044	0.793
	100	0.048	0.033	0.822
	500	0.019	0.017	0.922

Structure 1 - true structure is constructed by random effects, and
Structure 2 - true structure is AR-1;
N - number of independent subjects.

Table 2.8: Relative efficiencies and CIC selection frequencies. Proposed modified GMM approach is added to the settings of scenario 1 and $T = 5$.

			CIC	GEE	Modified GEE		Modified QIF		Modified GMM
Structure	N			Ind	Exch	AR-1	Exch	AR-1	
1	25	RE	1.02	1.00	1.18	1.11	1.08	1.05	0.75
		CIC Selection		20	244	103	305	193	135
	50	RE	1.09	1.00	1.16	1.10	1.13	1.06	0.71
		CIC Selection		22	314	71	385	183	25
	100	RE	1.18	1.00	1.21	1.12	1.19	1.10	0.70
		CIC Selection		4	351	36	438	169	2
	500	RE	1.17	1.00	1.18	1.10	1.18	1.11	0.68
		CIC Selection		0	327	0	625	48	0
2	25	RE	1.25	1.00	1.25	1.36	1.20	1.32	0.70
		CIC Selection		9	85	449	118	292	47
	50	RE	1.30	1.00	1.27	1.36	1.24	1.38	0.68
		CIC Selection		3	72	436	90	393	6
	100	RE	1.35	1.00	1.26	1.36	1.23	1.45	0.66
		CIC Selection		1	23	458	29	489	0
	500	RE	1.41	1.00	1.27	1.34	1.28	1.43	0.63
		CIC Selection		0	0	96	0	904	0

Structure 1 - true structure is constructed by random effects,
Structure 2 - true structure is AR-1;
N - number of independent subjects; Ind - independence; Exch - exchangeable;
CIC - correlation information criterion; GEE - generalized estimating equations;
QIF - quadratic inference function; GMM - generalized method of moments;
RE - relative efficiency. For each setting, they compare the empirical mean squared error (MSE) from the use of the GEE with independence structure to the MSEs from the use of different modeling options or CIC;
CIC Selection - Number of times out of 1,000 simulations that CIC selected the given method and corresponding structure.

Chapter 3 A Novel Approach to Selecting Classification Types for Time-Dependent Covariates for the Marginal Analysis of Longitudinal Data

3.1 Introduction

Generalized estimating equations (GEE) [3] are popular for the marginal analysis of longitudinal data in which subjects contribute repeated measurements over time. Consistent regression parameter estimates, under a correctly given mean structure, can often be obtained even if the working correlation structure is incorrectly specified. However, in the presence of certain types of time-dependent covariates, the estimating equations, and therefore the regression parameter estimates, can be biased due to the use of invalid moment conditions. Although invalid moment conditions do not exist when using an independence working correlation structure [4], resulting regression parameter estimation can be very inefficient because all valid moment conditions may not be used when employing this structure [5, 6].

In order to use all valid moment conditions, with the goal of improving estimation efficiency relative to GEE with a working independence structure, multiple methods have been proposed. In short, Lai and Small [1] took a generalized method of moments (GMM) approach [7], and Zhou *et al.* [9] modified the quadratic inference function (QIF) method [8]. Furthermore, Chen and Westgate [36] proposed a modified GEE approach that potentially improves upon the performance of the modified QIF method, particularly when the number of independent subjects is not large. Furthermore, the modified GEE and QIF approaches have been shown to perform better than the GMM approach [9, 36].

Although these methods require the data analyst to specify the type of time-dependent covariate, this will often be unknown in practice. Therefore, Leung *et al.*

[11] considered an empirical likelihood (EL) approach [12] in which moment conditions that are not guaranteed to provide consistent estimation are weighted, depending upon their estimated likelihoods of being valid, and linearly combined. Although this approach avoids the need to choose a covariate type, and is no less efficient than GEE with a working independence structure, we later demonstrate that this approach can be inefficient relative to our proposed approach. Alternatively, a covariate type could be determined via hypothesis testing. Specifically, Lai and Small [1] proposed hypothesis testing based on GMM, and Lalonde *et al.* [2] proposed an alternative approach utilizing correlations that requires the testing of each individual moment condition. However, in results to be presented later, we show that these approaches can result in too many moments being deemed valid; i.e., high type II error rates, thus favoring biased regression parameter estimation.

Therefore, in this chapter we propose a novel approach to select a working covariate type. In short, we propose a criterion that accounts for the impacts moment conditions have on both the efficiency and bias of regression parameter estimation corresponding to time-dependent covariates, with the goal of minimizing mean squared error (MSE). Furthermore, the proposed approach provides consistent estimation.

This chapter is organized as follows. Section 3.2 reviews existing approaches in the presence of time-dependent covariates. In Section 3.3, we propose our approach to selecting a working classification type for time-dependent covariates. In Section 3.4, we carry out a simulation study to assess the utility of the proposed method relative to existing methods, and in Section 3.5 we demonstrate these methods in application to anthropometric screening data [33]. Finally, concluding remarks are given in Section 3.6.

3.2 Time-Dependent Covariates and Current Methods

3.2.1 Notation and Generalized Estimating Equations

For ease of illustration, assume a longitudinal study setting in which there are N independent subjects measured at each of T distinct time points. We denote the observed outcome vector for the ith subject as $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iT}]^T$, which has a marginal mean given by $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ linked to covariates through a function, f, such that $f(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ for $\mathbf{x}_{ij} = [1, x_{1ij}, \ldots, x_{pij}]^T$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$. The working covariance matrix for $\boldsymbol{Y}_i$ is given by $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}$, $i = 1, \ldots, N$, where $\boldsymbol{A}_i = diag[\phi\nu(\mu_{i1}), \ldots, \phi\nu(\mu_{iT})]$ is a diagonal matrix representing the marginal variances, ϕ is a scale parameter assuming common dispersion, ν is a known function, and $\boldsymbol{R}_i$ is a symmetric positive definite working correlation matrix.

Using the GEE approach [3] to marginal modeling and letting $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^T$, regression parameter estimates, $\hat{\boldsymbol{\beta}}$, can be obtained by iteratively solving

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{3.1}$$

The $(k+1)$th row in Equation (3.1) corresponds to the estimating equation for β_k and is given by

$$\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{j=1}^{T} \frac{\partial\mu_{is}}{\partial\beta_k} \upsilon_i^{sj}(Y_{ij} - \mu_{ij}) = 0,$$

where υ_i^{sj}, $i = 1, ..., N$; $s, j = 1, ..., T$, is the (s, j)th element of $\boldsymbol{V}_i^{-1}$.

3.2.2 Types of Time-Dependent Covariates

Four types of time-dependent covariates are known to exist [1, 2]. Types I-III are well-known [1], whereas Type IV is a newer addition to the literature [2]. The kth covariate is classified as a Type I time-dependent covariate if $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} - \mu_{ij}\}) = 0 \; \forall \; s, j$, a Type II if $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} - \mu_{ij}\}) = 0$ for $s \geqslant j$, a Type III if $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} -$

$\mu_{ij}\}) \neq 0$ for some $s > j$, and a Type IV, which is the opposite of a Type II, if $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} - \mu_{ij}\}) = 0$ for $s \leqslant j$.

If β_k corresponds to a time-dependent covariates classified as Type II, III, or IV, then $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} - \mu_{ij}\}) \neq 0$ for some s, j combinations, and hence these moments are invalid. If GEE incorporates a working independence correlation structure, then the only moment conditions used are the ones such that $s = j$ which are always valid, and therefore an unbiased estimating equation is used regardless of the type of time-dependency [4]. However, this safe approach can result in notable efficiency loss if the covariate is not of Type III because additional valid moment conditions exist but are not utilized [1, 5]. Therefore, methods have been proposed that allow the use of all valid moment conditions, thus yielding more efficient parameter estimation, yet requiring the type of time-dependency to be known.

3.2.3 Existing Estimation Methods

Generalized Method of Moments

Lai and Small [1] utilized GMM [7] to combine all valid moment conditions. In short, they created a vector, $\boldsymbol{g}_i(\boldsymbol{\beta})$, consisting of all valid moment conditions from subject i, $i = 1, \ldots, N$, corresponding to the estimation of the $p + 1$ parameters such that $E(\boldsymbol{g}_i(\boldsymbol{\beta})) = \mathbf{0}$. With respect to the kth covariate, or $(k+1)$th parameter, there are T^2 valid moment conditions corresponding to a covariate that is of Type I, $T(T + 1)/2$ valid moment conditions for Type II or IV, and T valid moment conditions for Type III. To create $\boldsymbol{g}_i(\boldsymbol{\beta})$, all valid moments corresponding to each parameter are stacked such that the maximum length of $\boldsymbol{g}_i(\boldsymbol{\beta})$ is $T^2 \times (p+1)$, and estimating equations are formed by optimally weighting the linear combinations of $(1/N)\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\beta})$ through GMM [20].

Modified Quadratic Inference Functions

The QIF method proposed by Qu *et al.* [8] is based on the GMM and GEE approaches. In short, using correlation structures such that $\boldsymbol{R}_i^{-1} \approx \sum_{r=1}^{m} \alpha_{ri}\boldsymbol{M}_{ri}$, Equation (3.1) can be viewed as a linear combination of m sets of unbiased estimating equations that can be stacked and optimally, linearly combined via GMM. We note that this method utilizes GMM, as does the method of Lai and Small [1], although different estimating equations are used. With this method, $\boldsymbol{M}_{ri}$, $r = 1, \ldots, m$; $i = 1, \ldots, N$, are known basis matrices and α_{ri}, $r = 1, \ldots, m$; $i = 1, \ldots, N$, are functions of correlation parameters that can be ignored [8]. Two basis matrices are typically utilized for exchangeable and AR-1 working structures. For both structures, $\boldsymbol{M}_{1i}$ is an identity matrix, while $\boldsymbol{M}_{2i}$ is a matrix with 0 on the diagonal and 1 elsewhere for exchangeable, and 1 on the sub-diagonal and 0 elsewhere for AR-1. Zhou *et al.* [9], and similarly Cho and Dashnyam [22], modified $\boldsymbol{M}_{2i}$, denoted as $\boldsymbol{M}_{2i}^{*}$, to be a lower triangular matrix for a Type II time-dependent covariate, and thus $T(T+1)/2$ estimating equations for $s \geqslant j$ are used in $\boldsymbol{g}_{2i}(\boldsymbol{\beta})$. Alternatively, $\boldsymbol{M}_{2i}^{*}$ is an upper triangular matrix for a Type IV covariate and also yields the use of all $T(T+1)/2$ valid estimating equations. Only the identity matrix, $\boldsymbol{M}_{1i}$, is used when the covariate is of Type III, such that only the T valid estimating equations are used. In addition, no constrains to $\boldsymbol{M}_{2i}$ are needed for Type I because all T^2 moment conditions are valid.

Modified Generalized Estimating Equations

Chen and Westgate [36] proposed a modified GEE method in which elements in the inverse of the working correlation matrix are replaced with 0 whenever their use yields biased equations. Specifically, they created $\boldsymbol{R}_{ik}^{*-1}$, $k = 0, 1, \ldots, p$, by modifying $\boldsymbol{R}_i^{-1}$, the inverse of any given working correlation structure in Equation (3.1), according to the specific type of time-dependent covariate. If parameter k corresponds to a Type

I time-dependent or time-independent covariate, then all T^2 moment conditions are valid, and therefore $\boldsymbol{R}_{ik}^{*-1} = \boldsymbol{R}_i^{-1}$, indicating that the estimating equation is the same as the typical GEE equation. In regards to a Type II covariate, $\boldsymbol{R}_{ik}^{*-1}$ is restricted to be a lower triangular matrix such that the information from the $T(T+1)/2$ valid moment conditions for $s \geqslant j$ is included. The opposite is done regarding a Type IV covariate, such that $\boldsymbol{R}_{ik}^{*-1}$ is obtained by making all lower non-diagonal elements of $\boldsymbol{R}_i^{-1}$ equal to 0. Finally, $\boldsymbol{R}_{ik}^{*-1}$ is an identity matrix in the estimating equation corresponding to a Type III covariate. This modified GEE approach works particularly well for small sample size settings, and the modified QIF has the potential to perform better with larger sample sizes [36].

3.2.4 Empirical Likelihood Approach and Hypothesis Testing

Empirical Likelihood with Shrinkage Parameters

The previously described methods require correct specification of the covariate type, whereas in practice the true type will likely be unknown. Therefore, Leung *et al.* [11] utilized an EL approach [12] in which moment conditions that are not guaranteed to provide consistent estimation are empirically weighted based on their estimated likelihoods of being valid. The authors proposed dividing the T^2 available moment conditions into two vectors, $\boldsymbol{S}^M(\boldsymbol{\beta})$ and $\boldsymbol{S}^A(\boldsymbol{\beta})$. $\boldsymbol{S}^M(\boldsymbol{\beta})$ is comprised of the T moments that are always valid, and $\boldsymbol{S}^A(\boldsymbol{\beta})$ consists of the remaining $T^2 - T$ moments whose validity depends on the covariate type. A vector, $\boldsymbol{\gamma}$, of shrinkage parameters with dimension $T^2 - T$ is multiplied by $\boldsymbol{S}^A(\boldsymbol{\beta})$ to form $\boldsymbol{S}^{A,\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \boldsymbol{\gamma}^T \boldsymbol{S}^A(\boldsymbol{\beta})$. Here the elements for $\boldsymbol{\gamma}$ can be viewed as non-negative weights in [0,1] that are supposed to shrink the contributions from moment conditions based on the degree of bias they are estimated to create. The EL method is then used to combine the estimating functions $\boldsymbol{S}^M(\boldsymbol{\beta})$ and $\boldsymbol{S}^{A,\boldsymbol{\gamma}}(\boldsymbol{\beta})$ and to obtain the regression parameter estimates, $\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}$. Although $\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}$ is consistent, reducing the $T^2 - T$ moment conditions under question to one dimension via $\boldsymbol{S}^{A,\boldsymbol{\gamma}}(\boldsymbol{\beta})$ can still be inefficient, as will be demonstrated later via simulation.

GMM Hypothesis Testing

An alternative approach that has been used in the literature is to conduct hypothesis testing to determine the covariate type [1, 2]. The hypothesis testing approach of Lai and Small [1] examines the validity of moment conditions and is based on their GMM approach. In short, assume there are u moment conditions that are considered, v moments are known to be valid, and thus $u - v$ conditions are to be tested. The resulting test statistic has an asymptotic χ^2_{u-v}-distribution under the null hypothesis that the $u - v$ moment conditions under question are valid [1, 37, 38]. Therefore, this approach can test the null hypothesis that a covariate is of Type I versus the alternative that it is of Type II, and if Type I is rejected, then a test for Type II against Type III can be conducted. Alternatively, the procedure can be reversed.

Hypothesis Testing Using Correlations

As opposed to GMM-based hypothesis testing in terms of determining grouped moment conditions, Lalonde *et al.* [2] proposed another hypothesis testing approach to simultaneously examine the ungrouped moment conditions for $s \neq j$. They propose a separate test for each moment condition, having a null hypothesis of $E(\partial\mu_{is}/\partial\beta_k\{Y_{ij} - \mu_{ij}\}) = 0$, which is based on the correlation between standardized residuals and values of the kth covariate. Furthermore, they utilized a multiple testing adjustment to stabilize the family type I error rate [39]. As with the GMM hypothesis testing approach, tests having non-significant results correspond to valid moment conditions.

3.3 Proposed Method

The previously described methods have notable limitations, as will be apparent later in the simulation study results. The EL shrinkage approach of Leung *et al.* [11] can be inefficient when the covariate is not of Type III, and the hypothesis testing approaches can result in biased estimation due to their potential to favor the null hypothesis of valid moment conditions. Another limitation of the hypothesis testing approaches is that a significance level must be specified.

To remove these limitations, we propose an approach to select a working classification type. We note that although types for more than one time-dependent covariate can be chosen, for simplicity of notation we assume there is only one covariate of unknown type. To choose a working type for this covariate, consider an estimated MSE given by

$$\widehat{MSE}(\hat{\boldsymbol{\beta}}_z) = \widehat{Cov}(\hat{\boldsymbol{\beta}}_z) + \{\hat{\boldsymbol{\beta}}_z - \hat{\boldsymbol{\beta}}_{III}\}\{\hat{\boldsymbol{\beta}}_z - \hat{\boldsymbol{\beta}}_{III}\}^T. \tag{3.2}$$

Here, $\widehat{Cov}(\hat{\boldsymbol{\beta}}_z)$ denotes an empirically estimated covariance matrix of $\hat{\boldsymbol{\beta}}_z$, the vector of regression parameter estimates obtained when assuming the time-dependent covariate is of Type z, $z = I, II, III$, or IV. We note that $\widehat{Cov}(\hat{\boldsymbol{\beta}}_z)$ can be obtained when using the modified GEE or QIF approaches, but such an empirical covariance estimate may not be valid when using the GMM approach [36]. Due to $\boldsymbol{\beta}$ being unknown, in Equation (3.2) we replace it with $\hat{\boldsymbol{\beta}}_{III}$ because $\hat{\boldsymbol{\beta}}_{III} - \boldsymbol{\beta} \xrightarrow{p} 0$, thus providing a consistent estimate for bias, given by $\{\boldsymbol{\beta}_z - \boldsymbol{\beta}\}$. Here, $\boldsymbol{\beta}_z$ is defined such that $\hat{\boldsymbol{\beta}}_z - \boldsymbol{\beta}_z \xrightarrow{p} 0$. Furthermore, $\widehat{Cov}(\hat{\boldsymbol{\beta}}_z) \to 0$ as $N \to \infty$, and therefore $\widehat{MSE}(\hat{\boldsymbol{\beta}}_z) \to \{\boldsymbol{\beta}_z - \boldsymbol{\beta}\}\{\boldsymbol{\beta}_z - \boldsymbol{\beta}\}^T$.

Utilizing the estimated MSE allows for the consideration of both the efficiency that results from the use of the moment conditions corresponding to Type z as well as the bias that may arise. In order to utilize this estimated MSE to choose a working covariate type, we propose selecting the type that results in the smallest value for $tr\big(\widehat{MSE}(\hat{\boldsymbol{\beta}}_z)\big)$. We note that we utilize the trace because the trace of the empirical covariance matrix has been shown to work well for the selection of a working correlation structure [18, 29].

As $N \to \infty$, $\widehat{MSE}(\hat{\boldsymbol{\beta}}_z) \to \{\boldsymbol{\beta}_z - \boldsymbol{\beta}\}\{\boldsymbol{\beta}_z - \boldsymbol{\beta}\}^T$. Therefore, if a given working covariate type causes bias, then asymptotically this type will not be selected when using our proposed approach. In short, the proposed method results in consistent regression parameter estimation, although the true type is not guaranteed to be chosen. Specifically, if the true type is I, then any working type yields consistent estimation and can be asymptotically selected through our approach. If the truth is Type II (IV), then our method will select either Type II (IV) or III. Finally, our approach will asymptotically select Type III if this is the true type.

When using the proposed approach, $\widehat{Cov}(\hat{\boldsymbol{\beta}}_z)$ can yield biased estimates of variances of the estimated parameters corresponding to any time-dependent covariates for which the type was selected. Specifically, this formula assumes only the given type can be selected. However, the true variance of a corresponding regression parameter estimate depends on the complex probabilities of each type being selected. As a result, cluster bootstrapped standard errors (SEs) should be utilized for statistical inference in practice [40, 41]. Although results are not presented in our simulation study in the following section, we do note that the empirical coverage probabilities of 95% confidence intervals using bootstrapped SEs resulted in near-nominal coverage.

3.4 Simulation Study

3.4.1 Study Description

We compare the finite-sample performances of our proposed covariate type selection approach to the use of hypothesis testing and the EL approach of Leung *et al.* [11]. The proposed approach is demonstrated with both the modified GEE and modified QIF methods, as is the hypothesis testing approach of Lalonde *et al.* [2]. The hypothesis testing approach of Lai and Small [1] is used with their GMM method. For simplicity, results are presented with respect to an exchangeable working correlation structure for the modified GEE and modified QIF, although similar results were found with respect to an AR-1 working structure. Furthermore, a nominal 0.05 significance level was utilized for hypothesis testing approaches.

Three scenarios are used in the simulation study, corresponding to true Type I, II, and III time-dependent covariates, with results presented in Tables 3.1-3.3, respectively. Each scenario has the same marginal model given by $Y_{ij} = \beta_0 + \beta_1 x_{ij}$, $j = 1, \ldots, 5$; $i = 1, \ldots, N$, $N = 100$ and 500, although data generation depends on the covariate type as described below. Each setting is conducted through 1,000 simulations using R version 3.1.2 [30]. Furthermore, models are based on previous literature for time-dependent covariates [1, 31]. Although extensions of these scenarios were also studied in which marginal models included multiple differing types of time-dependent covariates, results were similar and therefore are not presented.

When the time-dependent covariate is either Type I or II, data are generated from $Y_{ij} = \tilde{\beta}_0 + \tilde{\beta}_1 x_{ij} + \tilde{\beta}_2 x_{i,j-1} + \gamma_i + \epsilon_{ij}$ and $x_{ij} = \kappa x_{i,j-1} + e_{ij}$, $j = 1, \ldots, 5$, where $\tilde{\boldsymbol{\beta}} = [0, 1, 1]^T$, and random effects, γ_i, ϵ_{ij}, and e_{ij}, are mutually independent and normally distributed with mean 0 and variance 4 [1, 31]. Note that $Var(e_{ij}) = \sigma_e^2$. Furthermore, when the covariate is Type I, $\tilde{\beta}_2 = 0$. In addition, x_{i0} follows a normal distribution with mean 0 and variance $\sigma_e^2/(1 - \kappa^2)$ because the time process for x_{ij} is stationary. Here let $\kappa = 0.5$. The marginal mean is given by $E[Y_{ij}|x_{ij}] = \tilde{\beta}_0 + (\tilde{\beta}_1 + \kappa\tilde{\beta}_2)x_{ij}$, which gives true values of $\tilde{\beta}_0 = 0$ for the marginal intercept, and $\tilde{\beta}_1 = 1$ and $\tilde{\beta}_1 + \kappa\tilde{\beta}_2 = 1.5$ for the marginal parameters corresponding to the Type I and Type II covariates, respectively.

When the time-dependent covariate is Type III, the process of data generation is from $Y_{ij} = \alpha x_{ij} + \gamma y_{i,j-1} + u_{ij}$ and $x_{ij} = \rho y_{i,j-1} + v_{ij}$, $j = 1, \ldots, 5$, where $\alpha = 0.5$, $\gamma = 0.2$, $\rho = 0.4$, and random effects, u_{ij} and v_{ij}, are mutually independent and normally distributed with mean 0 and variance 1 [1]. Note that $Var(u_{ij}) = \sigma_u^2$ and $Var(v_{ij}) = \sigma_v^2$. Moreover, y_{i0} follows a normal distribution with mean 0 and variance $\{\sigma_u^2/[1-(\alpha\rho+\gamma)^2]\} + \{\alpha^2\sigma_v^2/[1-(\alpha\rho+\gamma)^2]\}$ due to the stationary time process of (x_{ij}, Y_{ij}). The marginal mean is given by $E[Y_{ij}|x_{ij}] = [\alpha + \gamma\rho(\sigma_u^2 + \alpha^2\sigma_v^2)/(\rho^2\sigma_u^2 + \sigma_v^2 - 2\sigma_v^2\alpha\gamma\rho - \gamma^2\sigma_v^2)]x_{ij}$, which provides true values of 0 and 0.03 for the marginal intercept and slope.

In order to examine differences in estimation performances, in Tables 3.1-3.3 we present empirical biases and ratios of empirical MSEs of estimates for β_1, which we refer to as relative efficiencies (REs). For any given RE, the numerator is the MSE resulting from the use of GEE with an independence working structure, and the denominator is the MSE resulting from use of the given approach. Furthermore, we present the number of times a working covariate type is chosen out of the 1,000 simulations. We note that we do not consider Type IV for selection, as it may not be realistic in practice because it assumes that current outcomes have an impact on future covariate values but the covariate values cannot affect future outcomes. Table 3.4 presents, for each scenario, the empirical mean proportions of moment conditions deemed valid by the hypothesis testing approach of Lalonde *et al.* [2], corresponding to lower and upper non-diagonal triangular matrices for $s > j$ and $s < j$, respectively.

3.4.2 Results

The RE results corresponding to a true Type I time-dependent covariate (Table 3.1) demonstrate that the methods of comparison are all notably more efficient than GEE with a working independence correlation structure. This was most evident with the hypothesis testing approaches, as they favor a working Type I specification. Although less efficient in this scenario, the proposed selection approach selected Type I in the majority of simulations and resulted in greater regression parameter estimation efficiency than the use of the EL approach.

Results corresponding to a true Type II or III time-dependent covariate (Tables 3.2 and 3.3, respectively) demonstrate the utility of the proposed selection approach and the potentially dangerous cost of taking a hypothesis testing approach. The proposed approach, in general, resulted in the greatest efficiency relative to all other methods when the covariate was Type II, and, as desired, was as efficient as GEE with independence when the truth was Type III. Alternatively, use of the correlation test on each moment condition or use of the GMM-based test resulted in REs ranging from 0.04 to 0.95 over these scenarios, with the majority being 0.67 or below. This is a result from the tendency for these methods to favor Type I specification, thus resulting in biased regression parameter estimates. In Scenarios 2 and 3 with the consideration of a multiple testing adjustment, the high mean proportions of moment conditions incorrectly deemed valid explains the preference for Type I (Table 3.4) and thus small REs (Tables 3.2 and 3.3). We note that the REs were not notably improved when not using a multiple testing adjustment (result not shown), although the proportions of valid moments decreased (Table 3.4) and therefore lowered the type II error rates.

Based on theoretical expectations, the proposed approach favors consistent regression parameter estimation. Specifically, when the truth was Type II, the number of times the approach selected Type II or III increased with N. Similarly, when the truth was Type III, the number of times the approach selected Type III increased with N. Furthermore, Tables 3.2 and 3.3 also explicitly demonstrate that the proposed approach results in reducing bias as N increases.

3.5 Application

We now use data from the study of anthropometric screenings among children in the Philippines [32, 33] to examine the association between anthropometric covariates and future morbidity outcome. The data obtained from surveying 448 households were originally collected from 1984 to 1985 [32]. Lai and Small [1] used a subset of data containing 370 children ($\leqslant$ 14 years) from Bhargava [33], and each child had repeated measurements at three time points with four months between each subsequent measurement. Children with incomplete information were excluded, and only one child per household was chosen in order to eliminate statistical correlation due to household clustering [33].

We adopt the marginal model used by Lai and Small [1], Leung *et al.* [11], and Zhou *et al.* [9], given by

$$\mu_{ij} = \beta_0 + \beta_1 BMI_{ij} + \beta_2 Age_{ij} + \beta_3 Female_i + \beta_4 SR2_{ij} + \beta_5 SR3_{ij}, \quad j = 1, 2, 3,$$

where μ_{ij} is the ith child's marginal mean morbidity index during the jth four-month interval. The morbidity index utilizing the same logistic transformation made by Bhargava [33] and Lai and Small [1] is given by

$$y_{ij} = \log\left(\frac{\text{days child was sick in last 2 weeks prior to time } j + 0.5}{14.5 - \text{days child was sick in last 2 weeks prior to time } j}\right),$$

The known Type I time-dependent covariates collected from the anthropometric data are age in months and two indicator variables for survey rounds 2 and 3 to present seasonality in morbidity, whereas the type for BMI is unknown and is therefore our focus.

As in the simulation study, we analyze this dataset using the modified GEE and QIF methods with an exchangeable structure, and select a classification type for BMI through the use of our proposed approach. We also conduct the hypothesis testing methods as well as the EL approach of Leung *et al.* [11]. Table 3.5 gives the estimates of regression parameters and corresponding bootstrapped SEs using 1,000 cluster bootstrap samples, as well as the working covariate type for BMI by method.

The hypothesis testing approach using correlations for $s \neq j$ determines that, given non-significant p-values for all moment conditions, BMI is of Type I. Similarly, the GMM-based hypothesis testing approach gives a non-significant p-value of 0.80 for testing the null hypothesis of BMI being a Type I. Although both hypothesis testing approaches tend to be biased toward Type I, our proposed approach selects BMI to be of Type I when using either the modified GEE or QIF, thus giving stronger support for the use of a working Type I specification. Specifically, the criterion values resulting from the use of working Type I, II, and III within the modified GEE were 0.00164, 0.00347, and 0.00402, respectively, and with the modified QIF they were 0.00165, 0.00171, and 0.00172, respectively. With both methods, the smallest criterion value corresponds to Type I. Furthermore, the proposed approach, as well as the hypothesis testing methods, produce notably smaller SE estimates than the EL approach, thus revealing its potential for inefficiency. We note that the working type chosen for BMI is different from previous work. Specifically, Lalonde *et al.* [2] misclassified one valid moment at a nominal 0.05 significance level and treated this covariate as Type II, and Lai and Small [1] did not test the null hypothesis of BMI being of Type I.

3.6 Concluding Remarks

The marginal analysis of data in the presence of time-dependent covariates can be challenging when the type of time-dependency is unknown. Existing methods are limited, as they have the potential to be inefficient or result in biased regression parameter estimation. Therefore, we proposed an approach to select a working time-dependency type, and via a simulation study we showed that our proposed method is preferable to existing methods. Although the proposed approach is conservative relative to the use of hypothesis testing when the true covariate is of Type I, it is superior under settings of true Type II or III as the hypothesis testing approaches can work poorly as they favor a Type I specification, thus resulting in biased regression parameter estimation.

We note that in small-sample settings, adjustments to covariance estimators may be needed to correct for negative bias. In short, use of the empirical covariance weighting matrix with GMM or estimation of correlation parameters with GEE may increase variability in finite-sample sizes, resulting in covariance inflation of regression parameter estimates [14, 15, 16, 17, 27, 28]. Furthermore, the estimated empirical covariances utilized in practice are too small on average due to the use of residuals as opposed to unknown errors [24]. Such corrections are available for the modified GEE and QIF approaches, as well as the GMM approach [36].

Our simulation study and application example analyzed marginal models with continuous outcomes. However, the selection approach proposed in this chapter is applicable to marginal generalized linear models in general, regardless of the outcome type, and subjects with unbalanced repeated measurements are allowable. Furthermore, because of the increased complexity of the data generating process regarding time-dependent covariates, future work accounting for other outcome types is needed.

Table 3.1: Results for settings in which one Type I time-dependent covariate is used.

		GEE - Ind	Modified GEE		Modified QIF		GMM	EL
N			Proposed	Corr Test	Proposed	Corr Test	LS Test	
100	Bias	0.0034	0.0036	-0.0002	0.0045	0.0001	0.0004	-0.0015
	RE	1.00	1.43	5.71	1.28	5.57	3.07	1.22
	Type I		588		601		1000	
	Type II		305		211		0	
	Type III		107		188		0	
500	Bias	0.0011	0.0009	-0.0002	0.0012	-0.0003	-0.0003	-0.0014
	RE	1.00	1.38	6.29	1.28	6.26	5.54	1.26
	Type I		524		534		1000	
	Type II		342		246		0	
	Type III		134		220		0	

GEE - generalized estimating equations; Ind - independence;
QIF - quadratic inference function; GMM - generalized method of moments;
EL - empirical likelihood approach of Leung *et al.* [11]; N - number of independent subjects;
Corr Test - hypothesis testing approach of Lalonde *et al.* [2] using correlations;
LS Test - GMM-based hypothesis testing approach of Lai and Small [1];
Bias - empirical bias of each approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the GEE with independence structure to the MSE from the given method;
Types I-III - The number of times out of 1,000 simulations that the given covariate type was chosen.

Table 3.2: Results for settings in which one Type II time-dependent covariate is used.

		GEE - Ind	Modified GEE		Modified QIF		GMM	EL
N			Proposed	Corr Test	Proposed	Corr Test	LS Test	
100	Bias	0.0004	-0.0111	-0.2840	-0.0096	-0.2561	-0.0507	-0.0092
	RE	1.00	1.13	0.19	1.02	0.22	0.95	1.03
	Type I		35		39		1000	
	Type II		828		682		0	
	Type III		137		279		0	
500	Bias	0.0006	0.0006	-0.2839	0.0008	-0.2632	-0.0668	-0.0042
	RE	1.00	1.18	0.04	1.09	0.05	0.47	1.04
	Type I		0		0		1000	
	Type II		848		706		0	
	Type III		152		294		0	

GEE - generalized estimating equations; Ind - independence;
QIF - quadratic inference function; GMM - generalized method of moments;
EL - empirical likelihood approach of Leung *et al.* [11]; N - number of independent subjects;
Corr Test - hypothesis testing approach of Lalonde *et al.* [2] using correlations;
LS Test - GMM-based hypothesis testing approach of Lai and Small [1];
Bias- empirical bias of each approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the GEE with independence structure to the MSE from the given method;
Types I-III - The number of times out of 1,000 simulations that the given covariate type was chosen.

Table 3.3: Results for settings in which one Type III time-dependent covariate is used.

		GEE - Ind	Modified GEE		Modified QIF		GMM	EL
N			Proposed	Corr Test	Proposed	Corr Test	LS Test	
100	Bias	-0.0026	-0.0026	-0.0313	-0.0026	-0.0299	-0.0256	-0.0042
	RE	1.00	1.00	0.64	1.00	0.60	0.63	0.91
	Type I		4		0		1000	
	Type II		2		0		0	
	Type III		994		1000		0	
500	Bias	-0.0006	-0.0006	-0.0136	-0.0006	-0.0148	-0.0328	-0.0009
	RE	1.00	1.00	0.67	1.00	0.61	0.24	0.99
	Type I		0		0		1000	
	Type II		0		0		0	
	Type III		1000		1000		0	

GEE - generalized estimating equations; Ind - independence;
QIF - quadratic inference function; GMM - generalized method of moments;
EL - empirical likelihood approach of Leung *et al.* [11]; N - number of independent subjects;
Corr Test - hypothesis testing approach of Lalonde *et al.* [2] using correlations;
LS Test - GMM-based hypothesis testing approach of Lai and Small [1];
Bias - empirical bias of each approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the GEE with independence structure to the MSE from the given method;
Types I-III - The number of times out of 1,000 simulations that the given covariate type was chosen.

Table 3.4: Mean proportions of moment conditions deemed to be valid by the hypothesis testing approach of Lalonde *et al.* [2].

		Scenario 1		Scenario 2		Scenario 3	
N		Lower	Upper	Lower	Upper	Lower	Upper
	Ideal Proportion	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
	Adjusted Method						
100	Mean Proportion	1.0000	1.0000	1.0000	1.0000	0.8325	0.9984
500	Mean Proportion	1.0000	1.0000	1.0000	0.9980	0.4714	0.9455
	Unadjusted Method*						
100	Mean Proportion	1.0000	1.0000	1.0000	0.9995	0.5564	0.9378
500	Mean Proportion	1.0000	1.0000	0.9592	0.1803	0.3003	0.7361

Lower - moment conditions for $s > j$ in a lower, non-diagonal triangular matrix;
Upper - moment conditions for $s < j$ in a upper, non-diagonal triangular matrix;
N - number of independent subjects;
Ideal Proportion - the ideal proportion of valid moment conditions corresponding to the specific type of time-dependent covariate;
Adjusted and Unadjusted Methods - whether a multiple testing adjustment was used;
Mean Proportion - the empirical mean proportion of moment conditions deemed to be valid by the hypothesis testing approach of Lalonde *et al.* [2].
*Note that Lalonde *et al.* [2] proposed using an adjustment, but for illustrative purposes we also present results from not using an adjustment.

Table 3.5: Parameter estimates, bootstrapped standard error estimates (in parentheses), and working covariate types for BMI resulting from analyses of the anthropometric dataset.

	Modified GEE		Modified QIF		GMM	EL
Variable	Proposed	Corr Test	Proposed	Corr Test	LS Test	
BMI	-0.052 (0.045)	-0.052 (0.045)	-0.049 (0.045)	-0.049 (0.044)	-0.033 (0.042)	-0.024 (0.070)
Age	-0.012 (0.004)	-0.012 (0.004)	-0.011 (0.004)	-0.011 (0.004)	-0.010 (0.003)	-0.014 (0.013)
Gender	0.146 (0.110)	0.146 (0.110)	0.125 (0.110)	0.125 (0.109)	0.110 (0.106)	0.191 (0.326)
SR 2	-0.279 (0.112)	-0.279 (0.112)	-0.270 (0.112)	-0.270 (0.112)	-0.303 (0.109)	-0.218 (0.198)
SR 3	0.024 (0.129)	0.024 (0.129)	0.045 (0.128)	0.045 (0.128)	-0.013 (0.125)	-0.034 (0.287)
Type	I	I	I	I	I	

GEE - generalized estimating equations; QIF - quadratic inference function;
GMM - generalized method of moments;
EL - empirical likelihood approach of Leung *et al.* [11];
Corr Test - hypothesis testing approach of Lalonde *et al.* [2] using correlations;
LS Test - GMM-based hypothesis testing approach of Lai and Small [1];
Type - working covariate type for BMI.

Chapter 4 Marginal Quantile Regression for Longitudinal Data Analysis in the Presence of Time-Dependent Covariates

4.1 Introduction

Generalized estimating equations (GEE) [3] are well-known for their use in the marginal analysis of data from longitudinal studies in which measurements contributed from the same subject are correlated over time. As long as a correct mean structure is given, the regression parameters are consistently estimated even when the working correlation structure is misspecified. However, when certain types of time-dependent covariates are presented, the estimating equations, and thus estimates, can be biased unless an independence working correlation structure is employed [4]. Unfortunately, the resulting regression parameter estimation can be inefficient because not all valid moment conditions are utilized [5, 6]. Therefore, multiple approaches have been proposed to use all valid moments [1, 9, 36]. Most recently, the modified GEE approach proposed by Chen and Westgate [36] has been shown to perform best in terms of improving estimation efficiency.

Methods for the marginal analysis of longitudinal data in the presence of time-dependent covariates have only been developed for the modeling of the mean. An example carried out in this literature focuses on anthropometric screening data from Bouis and Haddad [32], in which the outcome of interest is morbidity index and time-dependent covariates include BMI, among others. Unfortunately, modeling the conditional mean of morbidity index may not be ideal because the response distribution is severely right skewed (Figure 4.1). Therefore, we desire the use of marginal quantile regression and are highly interested in how the distribution of the longitudinally measured morbidity index is associated with the time-dependent covariates.

Quantile regression for independent outcomes, introduced by Koenker and Bassett [42], has advantages relative to mean regression in that it is robust to outliers and it does not require any specified error distribution. In addition, quantile regression can provide a thorough description on the entire conditional distribution of a response variable. However, when correlated outcomes are present, modeling the within-subject correlation structure can be difficult. A safe approach, which ensures unbiased regression parameter estimates, proposed in the literature is to simply use an independence working correlation structure [43, 44, 45], although this may result in less efficient regression parameter estimation when data are highly correlated [46, 47, 48, 49].

Therefore, multiple approaches have recently been proposed for improving regression parameter estimation in marginal quantile regression for longitudinal data [46, 50]. However, the specification of a correlation structure is required for the quasi-score method of Jung [50], and regression parameter estimation from the use of quadratic inference function (QIF) approach of Tang and Leng [46] is not guaranteed to work well even if the correlation structure is correctly specified [13, 18]. Therefore, Fu and Wang [47] suggested a combination of the between- and within- weighted estimating equations under the working exchangeable structure, which was firstly introduced by Stoner and Leroux [51]. Additionally, Fu and Wang [47] extended their approach to allow any type of working correlation structure [13]. As a result, not only does this approach improve estimation performance, but it is robust to different error distributions. Nevertheless, in a longitudinal study some of the covariates may change over time and cause feed-back effects from the response variable, yet this issue has not been explored in the marginal quantile literature.

In this chapter, we therefore first propose an approach for marginal quantile regression in the presence of time-dependent covariates. This proposed method combines the estimating equations approach of Fu *et al.* [13] with the modified GEE approach of Chen and Westgate [36]. In consequence, the proposed approach can achieve notable gains in efficiency when compared with estimating equations under an independence correlation structure. Second, we propose a strategy to select a working type of time-dependency because in practice it may not be the case that the researcher knows the type of time-dependent covariate. In the marginal analysis literature with time-dependency, criteria such as the mean squared error (MSE), taking into account the influences moment conditions have on both the efficiency and bias of regression parameter estimation, can be used to select a working correlation structure [18, 29] or a classification type of time-dependent covariate [52]. In this chapter, we extend the use of the MSE to choose a working classification type such that consistent regression parameter estimation is a result.

This chapter is organized as follows. Section 4.2 introduces a marginal quantile regression and types of time-dependent covariates for longitudinal data. In Section 4.3, we propose the modified estimating equations for quantile regression in the presence of time-dependent covariates. Furthermore, we introduce the approach to selecting a working classification type for time-dependent covariates. In Section 4.4, we carry out a simulation study to compare the estimation performance and assess the utility of the proposed selection criterion relative to estimating equations with an independence working structure, and Section 4.5 demonstrates the proposed method in application to the motivating anthropometric screening data [32, 33]. Finally, we give concluding remarks in Section 4.6.

4.2 Quantile Regression and Time-Dependent Covariates

4.2.1 Notation and Quantile Regression

For ease of illustration, suppose a longitudinal study in which N independent subjects are repeatedly measured over T distinct time points. However, in general, the number of repeated measurements is allowed to vary across subjects. Let $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iT}]^T$ denote the observed outcome vector for the ith subject, and assume that the 100τth quantile of Y_{ij}, $j = 1, \dots, T$; $i = 1, \dots, N$ for $\tau \in (0, 1)$ is denoted by $Q(Y_{ij}|\mathbf{x}_{ij}, \tau)$ $=\mathbf{x}_{ij}^T \boldsymbol{\beta}^\tau$, where $\mathbf{x}_{ij} = [1, x_{1ij}, \dots, x_{pij}]^T$ is a vector observed at time point j for subject i, and $\boldsymbol{\beta}^\tau = [\beta_0^\tau, \beta_1^\tau, \dots, \beta_p^\tau]^T$ is an unknown vector corresponding to the regression coefficients at the 100τth quantile. Let $S_{ij}^\tau = \tau - I[Y_{ij} \le \mathbf{x}_{ij}^T \boldsymbol{\beta}^\tau]$ and $\boldsymbol{S}_i^\tau = [S_{i1}^\tau, \dots, S_{iT}^\tau]^T$, where $I(.)$ is an indicator function. The corresponding covariance matrix for $\boldsymbol{S}_i^\tau$ is given by $\boldsymbol{V}_i^\tau = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i^\tau (\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i = diag[\tau(1-\tau), \dots, \tau(1-\tau)]$ is a diagonal matrix representing the marginal variances, and $\boldsymbol{R}_i^\tau (\boldsymbol{\alpha})$ is a symmetric positive definite correlation matrix with 1 along the diagonal and one or more unknown correlation parameters given by $\boldsymbol{\alpha}$.

To find the estimate of the regression parameters, $\hat{\boldsymbol{\beta}}^\tau$, we consider the following optimal estimating equations [47, 48, 49, 50]

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{\Lambda}_i \boldsymbol{A}_i^{-1/2} \boldsymbol{R}_i^{\tau^{-1}} (\boldsymbol{\alpha}) \boldsymbol{A}_i^{-1/2} \boldsymbol{S}_i^\tau = \mathbf{0}, \tag{4.1}$$

in which $\boldsymbol{\Lambda}_i = diag[f_{i1}(0), \dots, f_{iT}(0)]$ with $f_{ij}(0)$ assumed to be a constant can be further eliminated [47]. The score function for the mth component corresponding to $\boldsymbol{\alpha}$, as well as the first partial derivative of the working Gaussian log-likelihood function for $(\boldsymbol{S}_1^\tau, \dots, \boldsymbol{S}_N^\tau)$ with respect to the mth component of $\boldsymbol{\alpha}$, can be expressed as [13]

$$\sum_{i=1}^{N} tr\left[\frac{\partial \boldsymbol{R}_i^{\tau^{-1}} (\boldsymbol{\alpha})}{\partial \alpha_m} (\boldsymbol{A}_i^{-1/2} \boldsymbol{S}_i^\tau \boldsymbol{S}_i^{\tau^T} \boldsymbol{A}_i^{-1/2} - \boldsymbol{R}_i^\tau) \right].$$

The correlation parameter α_m and its corresponding working correlation structure then can be estimated and constructed by optimizing this score function. We note that the asymptotic estimator for $Cov(\hat{\boldsymbol{\beta}}^\tau)$ is hardly obtained due to the involvement of unknown density functions of the errors. As a result, an induced smoothing technique [53, 54] has been commonly used to the marginal quantile regression models [47, 48, 49, 55]

In Equation (4.1), the $(k + 1)$th row corresponds to the estimating equation for

β_k^τ and is given by

$$\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{j=1}^{T} x_{kis}\upsilon_i^{sj}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau]) = 0,$$

where υ_i^{sj}, $i = 1, ..., N$ and $s, j = 1, ..., T$, is the (s,t)th element of $\boldsymbol{V}_i^{\tau^{-1}}$. If β_k^τ corresponds to certain types of time-dependent covariates, as will be specified in the following subsection, then we may not have $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] = 0 \; \forall \; s, j$.

4.2.2 Types of Time-Dependent Covariates

Four existing types of time-dependent covariates have been introduced in the marginal analysis literature for longitudinal data [1, 2]. In the manner of quantile regression modeling, the kth covariate is classified as a Type I time-dependent covariate if $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] = 0 \; \forall \; s, j$; $s, j = 1, \ldots, T$, at a given quantile level τ, a Type II if $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] = 0$ for $s \geqslant j$, a Type III if $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] \neq 0$ for some $s > j$, and a Type IV, which is the opposite of a Type II, if $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] = 0$ for $s \leqslant j$.

If β_k^τ corresponds to a time-dependent covariates which is classified as Type II, III, or IV, then $E\big[x_{kis}(\tau - I[Y_{ij} \leq x_{kij}\beta_k^\tau])\big] \neq 0$ for some s, j, will result in invalid moments. Pepe and Anderson [4] supported the use of GEE with an independence working correlation structure for marginal mean regression, then the only moment conditions utilized are the ones such that $s = j$ which are always valid regardless of the covariate type. Unfortunately, this safe approach can cause a great efficiency loss if the covariate is not of Type III because additional valid moment conditions are not used [1, 5]. Therefore, approaches allowing the use of all valid moment conditions have been proposed to achieve more efficient parameter estimation [1, 9, 36]. However, these methods only focus on mean regression and have not been extended to quantile regression when time-dependent covariates exist. We therefore propose approaches to improve estimation efficiency and select a working type of time-dependency which is often unknown in practice.

4.3 Proposed Methods

4.3.1 Improving Efficiency: Modified Estimating Equations for Quantile Regression

We first propose a modified estimating equations approach for improved efficiency by combining the estimating equations approach of Fu *et al.* [13] with the modified GEE approach of Chen and Westgate [36], which practically takes advantage of GEE's popularity. We replace elements with 0 in the inverse of the correlation matrix and the replacement is executed for each individual biased estimating equation, depending on the covariate type. Specifically, our proposed estimating equations for β_k^τ, $k = 0, 1, \ldots, p$, are given by

$$\sum_{i=1}^{N} \boldsymbol{X}_i^{k+1} \boldsymbol{A}_i^{-1/2} \boldsymbol{R}_i^{\tau *^{-1}}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{-1/2} \boldsymbol{S}_i^\tau = \mathbf{0}, \tag{4.2}$$

where $\boldsymbol{X}_i^{k+1}$ is the $(k+1)$th row of $\boldsymbol{X}^T$, and the elements of $\boldsymbol{R}_{ik}^{\tau *^{-1}}(\boldsymbol{\alpha})$, $k = 0, 1, \ldots, p$, are restricted to a certain type of covariate at a given quantile level τ. The modified approach then puts together these estimating equations and estimates regression parameter, correlation parameter, and standard error (SE) in the same nature as with the approach used in marginal quantile regression [13].

We propose to create $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ given in Equation (4.2) by modifying the inverse of a working correlation structure in general, $\boldsymbol{R}_i^{\tau^{-1}}$, employed in Equation (4.1) based on the specific type of time-dependent covariate. If parameter k is classified as a Type I time-dependent or time-independent covariate, then the information from all T^2 valid moment conditions is incorporated, and therefore $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ is equal to $\boldsymbol{R}_i^{\tau^{-1}}$, indicating that the estimating equations from Equations (4.1) and (4.2) are identical. When the estimating equation of a parameter corresponds to a Type II time-dependent covariate, $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ is constrained to be a lower triangular matrix such that the $T(T+1)/2$ moment conditions for $s \geqslant j$, $s, j = 1, \ldots, T$, are valid. In other wards, $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ is obtained by making all upper non-diagonal elements equal to 0. With respect to a Type IV time-dependent covariate, a contrast of a Type II, $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ can be obtained by taking $\boldsymbol{R}_{i}^{\tau^{-1}}$ and making all lower non-diagonal elements equal to 0. Finally, when the parameter corresponds to a Type III time-dependent covariate, $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ is considered to be diagonal matrices in the estimating equation.

4.3.2 Selection of Working Classification Type for Time-Dependency

Use of the approach just proposed requires data analysts know the covariate's type of time-dependency, although this is likely unknown in practice. Therefore, we now propose an approach to select a working type of time-dependency with the goal of producing the least variable regression parameter estimate possible. We note that although more than one type of time-dependent covariate can be chosen at any given quantile level τ, for simplicity of notation we assume there is only one time-dependent covariate of unknown type.

To choose a working type for this covariate, we first consider an estimated MSE given by

$$\widehat{MSE}(\hat{\boldsymbol{\beta}}_c^{\tau}) = \widehat{Cov}(\hat{\boldsymbol{\beta}}_c^{\tau}) + \big(\hat{\boldsymbol{\beta}}_c^{\tau} - \hat{\boldsymbol{\beta}}_{III}^{\tau}\big)\big(\hat{\boldsymbol{\beta}}_c^{\tau} - \hat{\boldsymbol{\beta}}_{III}^{\tau}\big)^T, \tag{4.3}$$

where $\hat{\boldsymbol{\beta}}_c^{\tau}$ is the vector of regression parameter estimates in which the time-dependent covariate is assumed to be Type c, $c = I, II, III$, or IV, and $\widehat{Cov}(\hat{\boldsymbol{\beta}}_c^{\tau})$ denotes an empirically estimated covariance matrix of $\hat{\boldsymbol{\beta}}_c^{\tau}$. We note that $\widehat{Cov}(\hat{\boldsymbol{\beta}}_z)$ can be obtained by using the induced smoothing method [53]. In Equation (4.3), we replace the unknown $\boldsymbol{\beta}^{\tau}$ with $\hat{\boldsymbol{\beta}}_{III}^{\tau}$ because $\hat{\boldsymbol{\beta}}_{III}^{\tau} \xrightarrow{p} \boldsymbol{\beta}^{\tau}$, thus providing a consistent bias estimate, which is $(\boldsymbol{\beta}_c^{\tau} - \boldsymbol{\beta}^{\tau})$. Here, the estimate of bias is followed by the defined $\boldsymbol{\beta}_c^{\tau}$ such that $\hat{\boldsymbol{\beta}}_c^{\tau} \xrightarrow{p} \boldsymbol{\beta}_c^{\tau}$. As $N \to \infty$, $\widehat{Cov}(\hat{\boldsymbol{\beta}}_c^{\tau}) \to 0$ and $\widehat{MSE}(\hat{\boldsymbol{\beta}}_c^{\tau}) \to (\boldsymbol{\beta}_c^{\tau} - \boldsymbol{\beta}^{\tau})(\boldsymbol{\beta}_c^{\tau} - \boldsymbol{\beta}^{\tau})^T$. Therefore, if a given working covariate type yields bias, then asymptotically this type will not be chosen when using the selection approach. Specifically, if the truth is of Type I, then any working type produces consistent regression parameter estimation and can be chosen through this approach. If the true type is II (IV), then this approach method will choose either II (IV) or III. Moreover, asymptotically our method will choose Type III if this is the true type.

In order to utilize this estimated MSE to select a working classification type, we propose choosing the type that occurs with the smallest value for the trace of an empirical covariance matrix, $tr\left[\widehat{MSE}(\hat{\boldsymbol{\beta}}_c^{\tau})\right]$. We note that this criterion has been proven to perform well for the selection of a working covariate type [52]. In addition, the true variance of a corresponding regression parameter estimate relies upon the complex probabilities of each type being chosen, and therefore $\widehat{Cov}(\hat{\boldsymbol{\beta}}_c^{\tau})$ can result in a biased estimate of the variance. In consequence, cluster bootstrapped SEs should be adopted for statistical inference [40, 41, 52]. Note that the empirical coverage probabilities of 95% confidence intervals using bootstrapped SEs resulted in near-nominal coverage, although the results are not shown in the simulation study.

4.4 Simulation Study

4.4.1 Study Description

We now compare the performances of our proposed selection approach for covariate type of time-dependency to the use of an independence working correlation structure, which treats unknown types of time-dependency as Type III, in the marginal quantile analysis. The selection approach is demonstrated with the modified estimating equations method using a first-order autoregressive (AR-1) working correlation structure, as AR-1 may be preferred over other structures such as exchangeable in a longitudinal study [31].

Three scenarios are carried out in the simulation study, corresponding to true Type I, II, and III time-dependent covariates, with results presented in Tables 4.1-4.3, respectively. Each scenario has the same marginal model given by $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$, $i = 1, \ldots, N$; $j = 1, \ldots, T$. The data generation depending on the covariate type are described in the following paragraph. The setting with $N = 500$ and $T = 5$ is conducted through 1,000 simulations using R version 3.1.2 [30]. Furthermore, models are based on previous marginal mean regression literature for time-dependent covariates [1, 11, 31] and marginal quantile regression literature [13, 47, 49]. Although marginal quantile models including multiple types of time-dependent covariates were also studied, results were similar and therefore are not presented.

When the time-dependent covariate is either Type I, II, or III, data are generated from $Y_{ij} = \tilde{\beta}_0 + \tilde{\beta}_1 x_{ij} + \tilde{\beta}_2 x_{i,j-1} + \gamma_i + \epsilon_{ij}$ and $x_{ij} = \kappa x_{i,j-1} + \theta\gamma_i + \delta_{ij}$, $i = 1, \ldots, 500$; $j = 1, \ldots, 5$, where $\tilde{\boldsymbol{\beta}} = [0, 1, 1]^T$, and random effects, γ_i and δ_{ij}, are mutually independent and normally distributed with mean 0 and variance 1 [1, 31]. Note that $Var(\gamma_{ij}) = \sigma_\gamma^2$ and $Var(\delta_{ij}) = \sigma_\delta^2$. The covariate is of Type I if $\tilde{\beta}_2 = \theta = 0$, while the covariate is of Type II if $\theta = 0$. Additionally, x_{i0} follows a normal distribution with mean 0 and variance $(\theta^2\sigma_\gamma^2 + \sigma_\delta^2)/(1 - \kappa^2)$ because the time process for x_{ij} is stationary. Here let $\kappa = 0.5$ and $\theta = 1.5$. The marginal mean given by $E[Y_{ij}|x_{ij}] = \tilde{\beta}_0 + \{\tilde{\beta}_1 + \kappa\tilde{\beta}_2 + [(\theta^2\sigma_\gamma^2)(1+\kappa)/\theta(\theta^2\sigma_\gamma^2 + \sigma_\delta^2)]\}x_{ij}$ gives true values of $\tilde{\beta}_0 = 0$ for the marginal intercept, and $\tilde{\beta}_1 = 1$, $\tilde{\beta}_1 + \kappa\tilde{\beta}_2 = 1.5$, and $\tilde{\beta}_1 + \kappa\tilde{\beta}_2 + [(\theta^2\sigma_\gamma^2)(1+\kappa)/\theta(\theta^2\sigma_\gamma^2 + \sigma_\delta^2)] = 2.19$ for the marginal parameters corresponding to the Type I, II, and III covariates, respectively. Furthermore, let $\epsilon_{ij} = q + e_{ij}$ and the use of q is to guarantee $p(\epsilon_{ij} \leqslant 0) = \tau$, the quantile level. Four cases are accounted for $\boldsymbol{e}_i = [e_{i1}, \ldots, e_{i5}]^T$: cases (1)-(3) assume that $\boldsymbol{e}_i$ follows multivariate normal distribution, multivariate Student's t-distribution with three degrees of freedom, and multivariate log-normal distribution, respectively, incorporating combinations of either an exchangeable or AR-1 working structure with a correlation parameter 0.3 or 0.7; in order to create correlated heteroscedastic errors, cases (4) assumes $e_{ij} = 0.25(1+|x_{ij}|)\zeta_{ij}$, where $\boldsymbol{\zeta}_i = [\zeta_{i1}, \ldots, \zeta_{i5}]^T$ follows multivariate normal distribution with the same combinations as cases (1)-(3).

In order to examine differences in estimation performances, in Tables 4.1-4.3 we present empirical biases corresponding to either the reference approach with an independence working structure or our proposed approach, and ratios of empirical MSEs of estimates for β_1, which we refer to as relative efficiencies (REs). For any given RE, the numerator is the MSE resulting from the use of reference approach, and the denominator is the MSE resulting from the use of our approach. Furthermore, we present the number of times a working covariate type is selected out of the 1,000 simulations. Note that we do not use Type IV for selection, as in practice this type may be rare because it assumes that outcomes have an impact on covariate values in the future but these covariate values cannot influence future outcomes.

4.4.2 Results

Results corresponding to either a true Type I, II, or III time-dependent covariate (Tables 4.1, 4.2, and 4.3, respectively) show that the proposed selection approach used with the modified estimating equations method is more efficient than the approach incorporating an independence working correlation structure, i.e., use of working Type III, in the presence of within-subject correlation (cases 1-4). The REs ranged from 1.09 to 1.30, 1.04 to 1.10, and 1.00 to 1.06, respectively, over scenarios 1-3. When correlated heteroscedastic errors were accounted for (case 4), the results, in terms of REs and selection frequencies, were similar to those with errors following correlated parametric distributions (cases 1-3). The reason for these efficiency gains is because the modified approach technically employs working correlation matrices with zero elements in order to ensure only valid moment conditions are implemented.

Additionally, the proposed approach worked well in terms of REs and selection frequencies for any given quantile level relative to the independence estimating equations approach. The RE results corresponding to cases 1-4 and three quantile levels under the three scenario settings also demonstrate that, given a higher within-subject correlation, the proposed selection method, in general, resulted in the greater efficiency and chose most often the desired type of covariate. The results with respect to REs and selection frequencies were comparable regardless of the given correlation structure. The selection approach had efficiency gains when the true Type I or II was under consideration and, as desired, can ensure the chosen Type III covariate was the actual type of time-dependent covariate. Specifically, because of none selection contributed to Type I, which can cause bias, under the true Type II, based on theoretical expectations, negligible biases of regression parameter estimation were found when the truth were Type I and II, and therefore the REs were dominated by the efficiency of regression parameter estimation (Tables 4.1 and 4.2).

4.5 Application

We adopt the anthropometric screening data from the children study in the Philippines [32, 33] to examine the association between anthropometric factors and morbidity index over time. The data were originally obtained from 448 households from 1984 to 1985 [32]. Then, a subset of data containing 370 children ($\leqslant 14$ years) was used as the final data [1, 33], in which each child had measurements at three time points with four months between each subsequent measurement. Children with incomplete measurements were excluded, and only one child per household was selected for eliminating statistical correlation resulted from household clustering [33].

We use the marginal model suggested in the existing literature [1, 9, 11, 36], but employ marginal quantile regression at three quantile levels, $\tau = 0.25$, 0.50, and 0.75, given by

$$Y_{ij} = \beta_0 + \beta_1 BMI_{ij} + \beta_2 Age_{ij} + \beta_3 Female_i + \beta_4 SR2_{ij} + \beta_5 SR3_{ij} + \epsilon_{ij}; \quad j = 1, 2, 3,$$

where Y_{ij}, as presented below, is the ith child's morbidity index during the jth four-month interval, and the morbidity index was conducted through the logistic transformation [1, 33].

$$Y_{ij} = \log\left(\frac{\text{days child was sick in last 2 weeks prior to time } j + 0.5}{14.5 - \text{days child was sick in last 2 weeks prior to time } j}\right).$$

Three covariates, including age in months and two indicators for survey rounds 2 and 3, are categorized as the known Type I time-dependent covariates, whereas BMI's classification type of time-dependency is unknown and is the main focus of this analysis.

As in the simulation study, we analyze this data using the independence estimating equations method and our modified method with an AR-1 correlation structure, and select a working type for BMI through the use of our selection approach under three given quantiles. Table 4.4 gives the estimates of regression parameters and corresponding cluster bootstrapped SEs using 2,000 cluster bootstrap samples, as well as the working type for BMI selected by our method.

The proposed approach assigns a working Type III classification for BMI at the first quartile (25th quantile) and median (50th quantile), whereas a working Type I classification is chosen at the third quartile (75th quantile) based on the smallest criterion value. In addition, at the 25th and 50th quantile levels both approaches produce similar results in terms of regression parameter and SE estimates for BMI due to the choice of Type III. Furthermore, our proposed approach produces smaller SE estimates than the reference approach at the 75th quantile, thus revealing our proposed method's potential for efficiency improvement. For the other time-dependent covariates of known type, smaller SE estimates are obtained using the proposed method. The use of a marginal quantile analysis provides a complete description of the morbidity index distribution to model the BMI, rather than the marginal mean analysis which gives support for the use of a working Type I specification [52].

4.6 Concluding Remarks

Covariates or predictors in a longitudinal study may change over time. Marginal mean regression analyses for longitudinal data have been widely introduced when time-dependent covariates are presented. However, for some real-world data the use of mean regression models may be sensitive to skewness and outliers in the data. In such cases, the use of marginal quantile analysis for modeling the conditional quantiles of the response variable is recommended. Therefore, we first proposed a modified approach for marginal quantile regression to utilize all valid moment conditions in order to improve regression parameter estimation, compared to the approach incorporating an independence working structure, while still attaining valid inference. Furthermore, as a data analyst, to decide which type of time-dependent covariate being used for the analysis of any given dataset can be challenging. As a result, we proposed an approach to determine the working type of covariate, and through a simulation study we presented that our method is preferable to the approach with an independence structure. The proposed selection approach is superior under scenarios of true Types I and II, and is as efficient as the reference approach when the true covariate is of Type III.

Although we only considered independence and AR-1 working correlation structures in the chapter, other structures are available as well, including exchangeable and Toeplitz correlation matrices. We note that with our modified approach, the working structure is technically not an actual correlation structure because some non-zero elements of $\boldsymbol{R}_i^{\tau^{-1}}$ corresponding to invalid moment conditions are replaced with zeros, and therefore $\boldsymbol{R}_{ik}^{\tau *^{-1}}$ will not be the inverse of a true correlation matrix when β_k^τ corresponds to a Type II or IV.

Our simulation study and application example were analyzed via marginal quantile regression models with balanced repeated measurements. Nonetheless, the proposed estimation approach and selection approach in this chapter are applicable to subjects with varying repeated measurements. Future study can be extended to improve efficiency of estimation performance of composite marginal quantile regression [55], which has been proposed when multiple quantiles share common characteristics, in the presence of time-varying covariates. Furthermore, approaches using a general stationary autocorrelation structure [49] and a selection technique, via the use of a Gaussian pseudolikelihood in substitution for a parametric likelihood [13], to decide the most adequate working correlation structure have been suggested to prevent the specification of any specific working correlation structures. Simultaneously selecting a working correlation structure and deciding a covariate type of time-dependency can be further developed. Additionally, because of the increasingly complex data generation in regards to time-dependent covariates, future work accounting for other marginal quantile models is needed on the simulation.

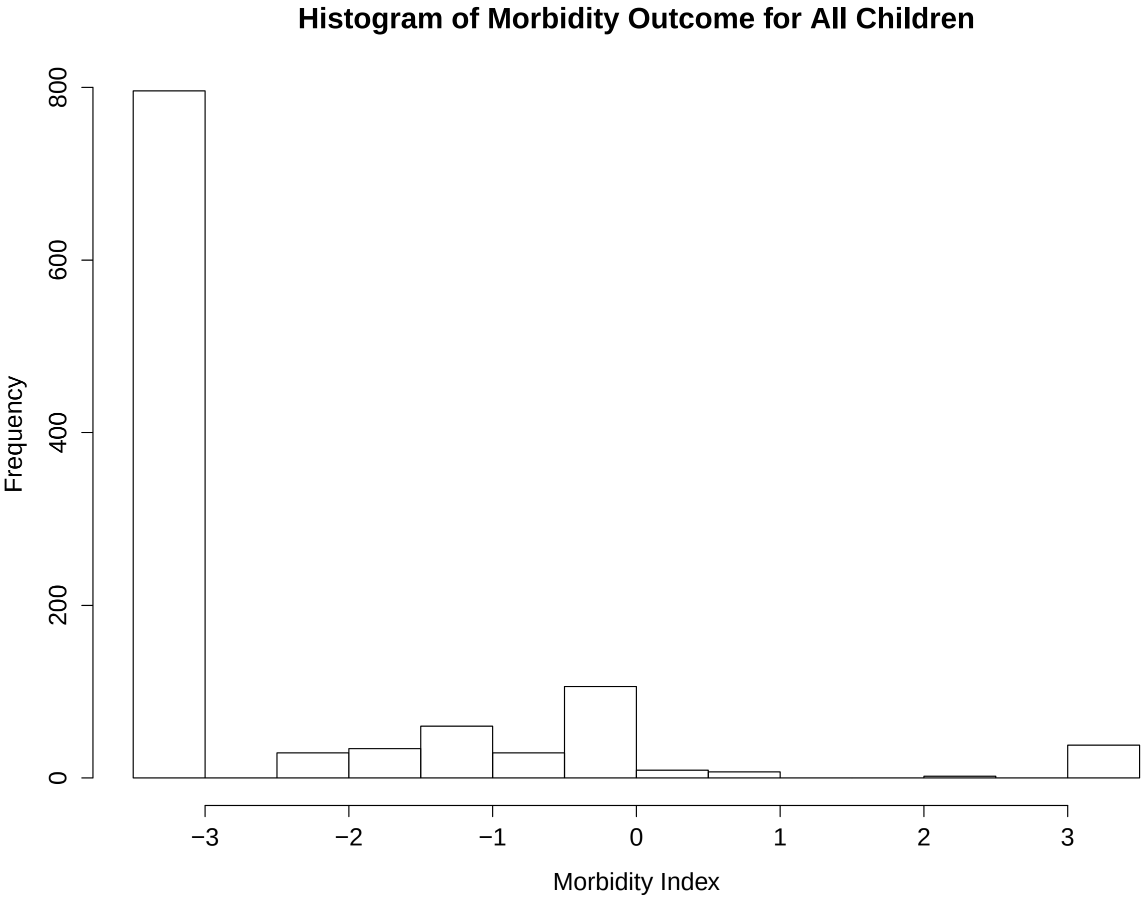

Figure 4.1: Histogram of morbidity index for all 370 children.

Table 4.1: Results for all Cases 1-4 in which one Type I time-dependent covariate is incorporated.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0003	-.0002	.0009	.0032	-.0001	-.0001	.0002	.0036	.0005	.0007	.0016	.0018
	$Bias_P$	-.0001	-.0006	.0004	.0028	.0000	-.0000	.0000	.0034	.0007	.0008	.0015	.0020
Case	RE	1.137	1.256	1.126	1.229	1.146	1.230	1.095	1.221	1.164	1.222	1.131	1.221
(1)	Type I	598	563	635	587	562	545	579	547	615	562	640	567
	Type II	285	340	259	309	300	366	283	339	273	339	237	324
	Type III	117	97	106	104	138	89	138	114	112	99	123	109
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0016	.0004	-.0018	.0005	-.0012	-.0001	-.0021	.0004	-.0016	-.0005	-.0021	-.0009
	$Bias_P$	-.0014	.0003	-.0020	.0007	-.0013	.0005	-.0015	.0007	-.0015	-.0007	-.0017	-.0008
Case	RE	1.139	1.217	1.120	1.197	1.096	1.202	1.115	1.183	1.128	1.232	1.148	1.236
(2)	Type I	617	566	623	584	555	528	565	541	644	569	590	607
	Type II	276	339	278	295	316	360	303	339	253	339	292	289
	Type III	107	95	99	121	129	112	132	120	103	92	118	104

Table 4.1: Continued.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0002	-.0006	-.0009	.0022	-.0008	-.0006	-.0008	.0032	-.0007	.0002	-.0005	.0023
	$Bias_P$	-.0003	-.0005	-.0006	.0020	-.0009	-.0003	-.0007	.0032	-.0005	-.0007	-.0007	.0023
Case	RE	1.181	1.287	1.161	1.276	1.139	1.240	1.123	1.228	1.108	1.242	1.106	1.204
(3)	Type I	565	538	595	591	552	488	584	563	662	560	650	591
	Type II	313	363	287	309	327	394	286	334	236	325	251	290
	Type III	122	99	118	100	121	118	130	103	102	115	99	119
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0008	.0011	-.0002	-.0020	-.0004	.0014	.0012	-.0019	-.0017	.0006	.0024	-.0021
	$Bias_P$	-.0009	.0007	.0000	-.0015	-.0005	.0015	.0006	-.0025	-.0019	.0007	.0022	-.0018
Case	RE	1.229	1.217	1.175	1.280	1.195	1.238	1.146	1.202	1.213	1.298	1.196	1.252
(4)	Type I	486	482	501	462	463	486	471	459	509	523	506	474
	Type II	389	403	382	431	419	411	399	422	390	388	361	415
	Type III	125	115	117	107	118	103	130	119	101	89	133	111

τ - quantile level; ρ - correlation parameter; Exch - exchangeable; AR-1 - first-order autoregressive;
$Bias_I$ - empirical bias of the approach with an independence structure in estimating the regression parameter;
$Bias_P$ - empirical bias of the proposed approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the estimation method with an independence structure to the MSE from the proposed method;
Types I-III - the number of times out of 1,000 simulations that the given covariate type was selected.

Table 4.2: Results for all Cases 1-4 in which one Type II time-dependent covariate is incorporated.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0011	.0016	.0009	.0035	-.0000	.0003	.0017	.0037	.0009	-.0009	.0014	.0005
	$Bias_P$	-.0010	.0013	.0008	.0033	-.0002	.0008	.0011	.0036	.0007	-.0010	.0012	.0006
Case	RE	1.054	1.088	1.059	1.058	1.053	1.066	1.052	1.063	1.064	1.080	1.060	1.075
(1)	Type I	0	0	0	0	0	0	0	0	0	0	0	0
	Type II	820	846	795	827	760	810	735	795	833	850	787	844
	Type III	180	154	205	173	240	190	265	205	167	150	213	156
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0008	.0000	-.0006	.0003	-.0014	-.0010	-.0026	-.0005	-.0011	-.0002	-.0036	-.0006
	$Bias_P$	-.0008	-.0002	-.0006	.0002	-.0012	-.0008	-.0028	-.0007	-.0014	.0002	-.0036	.0002
Case	RE	1.081	1.070	1.047	1.086	1.059	1.072	1.052	1.043	1.053	1.081	1.059	1.086
(2)	Type I	0	0	0	0	0	0	0	0	0	0	0	0
	Type II	806	845	762	821	771	832	727	783	826	857	791	835
	Type III	194	155	238	179	229	168	273	217	174	143	209	165

Table 4.2: Continued.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0017	-.0021	.0017	.0007	-.0015	-.0007	.0006	.0009	-.0010	.0009	-.0002	.0010
	$Bias_P$	-.0015	-.0020	.0016	.0006	-.0008	-.0004	.0006	.0004	-.0010	.0012	-.0003	.0008
Case	RE	1.058	1.097	1.054	1.048	1.036	1.088	1.046	1.072	1.084	1.078	1.060	1.092
(3)	Type I	0	0	0	0	0	0	0	0	0	0	0	0
	Type II	788	829	774	797	753	787	763	804	815	841	789	815
	Type III	212	171	226	203	247	213	237	196	185	159	211	185
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	-.0025	-.0024	-.0001	-.0041	-.0018	-.0017	.0011	-.0023	-.0025	.0008	.0024	-.0030
	$Bias_P$	-.0024	-.0023	.0002	-.0034	-.0020	-.0016	.0017	-.0021	-.0021	.0006	.0023	-.0031
Case	RE	1.088	1.100	1.060	1.073	1.071	1.077	1.048	1.058	1.087	1.091	1.085	1.090
(4)	Type I	0	0	0	0	0	0	0	0	0	0	0	0
	Type II	841	857	811	859	795	789	752	764	851	859	825	841
	Type III	159	143	189	141	205	211	248	236	149	141	175	159

τ - quantile level; ρ - correlation parameter; Exch - exchangeable; AR-1 - first-order autoregressive;
$Bias_I$ - empirical bias of the approach with an independence structure in estimating the regression parameter;
$Bias_P$ - empirical bias of the proposed approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the estimation method with an independence structure to the MSE from the proposed method;
Types I-III - the number of times out of 1,000 simulations that the given covariate type was selected.

Table 4.3: Results for all Cases 1-4 in which one Type III time-dependent covariate is incorporated.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
Case (1)	$Bias_I$	.0163	.0161	.0166	.0158	.0172	.0175	.0176	.0169	.0154	.0156	.0167	.0151
	$Bias_P$	.0159	.0155	.0161	.0151	.0170	.0171	.0174	.0164	.0150	.0151	.0161	.0144
	RE	1.036	1.052	1.048	1.053	1.022	1.036	1.020	1.047	1.025	1.040	1.050	1.060
	Type I	46	10	62	14	126	7	135	7	44	10	44	16
	Type II	227	64	172	63	214	41	210	51	214	52	215	59
	Type III	727	926	766	923	660	952	655	942	742	938	741	925
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
Case (2)	$Bias_I$	.0149	.0156	.0155	.0154	.0176	.0184	.0179	.0169	.0155	.0166	.0156	.0147
	$Bias_P$	.0142	.0151	.0150	.0151	.0171	.0179	.0174	.0164	.0150	.0161	.0150	.0142
	RE	1.047	1.040	1.037	1.025	1.041	1.029	1.035	1.041	1.039	1.022	1.052	1.032
	Type I	23	3	19	3	59	2	52	5	13	1	20	2
	Type II	200	35	174	23	190	24	191	17	180	40	189	31
	Type III	777	962	807	974	751	974	757	978	807	959	791	967

Table 4.3: Continued.

		τ=0.25				τ=0.50				τ=0.75			
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	.0223	.0222	.0225	.0221	.0196	.0195	.0199	.0194	.0115	.0110	.0111	.0116
	$Bias_P$	.0222	.0217	.0223	.0215	.0193	.0189	.0195	.0187	.0109	.0104	.0104	.0109
Case	RE	1.017	1.043	1.022	1.053	1.031	1.059	1.034	1.059	1.048	1.025	1.061	1.032
(3)	Type I	213	18	187	21	95	15	64	13	11	0	21	2
	Type II	191	72	175	84	195	31	198	36	199	49	192	42
	Type III	596	910	638	895	710	954	738	951	790	951	787	956
		Exch		AR-1		Exch		AR-1		Exch		AR-1	
	ρ	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
	$Bias_I$	.0158	.0157	.0152	.0173	.0171	.0159	.0165	.0159	.0154	.0154	.0157	.0148
	$Bias_P$	.0152	.0152	.0146	.0169	.0172	.0153	.0166	.0154	.0149	.0150	.0152	.0142
Case	RE	1.032	1.032	1.054	1.030	1.006	1.039	1.001	1.029	1.042	1.028	1.049	1.043
(4)	Type I	54	1	52	3	169	4	175	7	52	0	51	6
	Type II	218	73	222	82	245	114	230	117	215	96	207	82
	Type III	728	926	726	915	586	882	595	876	733	904	742	912

τ - quantile level; ρ - correlation parameter; Exch - exchangeable; AR-1 - first-order autoregressive;
$Bias_I$ - empirical bias of the approach with an independence structure in estimating the regression parameter;
$Bias_P$ - empirical bias of the proposed approach in estimating the regression parameter;
RE - relative efficiency or ratio of the empirical mean squared error (MSE) from the estimation method with an independence structure to the MSE from the proposed method;
Types I-III - the number of times out of 1,000 simulations that the given covariate type was selected.

Table 4.4: Parameter estimates, empirical and cluster bootstrapped standard error estimates (in parentheses), and working types of covariate for BMI resulting from analyses of the anthropometric dataset.

	Independence			Proposed*		
Variable	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
BMI	-0.20 (0.002)	-0.18 (0.003)	-0.05 (0.024)	-0.20 (0.002)	-0.18 (0.003)	-0.05 (0.019)
Age	-0.01 (0.001)	-0.01 (0.001)	-0.03 (0.005)	-0.01 (0.001)	-0.01 (0.001)	-0.03 (0.005)
Gender	-0.02 (0.021)	-0.02 (0.030)	0.41 (0.267)	-0.02 (0.026)	-0.01 (0.036)	0.42 (0.221)
SR 2	-0.08 (0.025)	-0.08 (0.036)	-0.69 (0.328)	-0.07 (0.021)	-0.06 (0.033)	-0.66 (0.248)
SR 3	0.002 (0.027)	0.05 (0.035)	0.42 (0.374)	0.01 (0.024)	0.06 (0.034)	0.47 (0.281)
Type				III	III	I

τ - quantile level; SR - survey round; Type - working covariate type for BMI.
*Note that the standard error estimates are obtained using the cluster bootstrapped method.

Chapter 5 Summary

5.1 Findings and Future Work

This dissertation researched the existing approaches that use all valid moment conditions in order to improve efficiency relative to GEE with an independence working correlation structure when certain types of time-dependent covariates are included in a marginal model, and proposed a modified GEE to improve their performances. The other topic of interest was to select a combination of estimation approach and working structure, resulting in the smallest variances of regression parameter estimates, that is generally unknown to the analyst. Additionally, previous literature assumed the researcher knows the type of time-dependent covariate, which realistically may not be the case. Therefore, another concern was given to choose a unknown type of time-dependent covariate. Finally, for some real-world datasets the use of marginal mean regression models may be sensitive to skewness and outliers in the data, and thus we studied marginal quantile analysis for longitudinal data so as to model conditional quantiles of the response variable.

GMM and modified QIF approaches that utilize all valid moment conditions have been proposed to improve efficiency for the marginal analysis of longitudinal data in the presence of time-dependent covariates. However, we found that these approaches may result in invalid inference. To improve upon the validity of inference with the GMM approach, we developed a modified, non-singular weighting matrix to ensure nominal coverage probabilities. Unfortunately, this modified GMM did not work well in terms of regression parameter estimation, and therefore we do not support its use in practice. The proposed modified GEE often outperformed all other methods that have been proposed. Nonetheless, the modified QIF did perform best, in terms of estimating the regression parameter corresponding to a Type II time-dependent covariate, in some large-sample settings in our simulation study due to its theoretical efficiency advantage. Furthermore, the CIC worked well in terms of selecting the best method and structure combination and thus regression parameter estimation.

To select a working covariate type of time-dependency, we proposed a selection method to utilize an estimated MSE and allow for the concurrence of both the efficiency that results from the use of the moment conditions corresponding to Type z, $z = I, II, III$, or IV as well as the bias that may arise, and via a simulation study we showed that our proposed method is preferable to existing methods, including the use of hypothesis testing and the EL approach. Although the proposed approach is conservative relative to the hypothesis testing methods when the true covariate is of Type I, it is superior under settings of true Type II or III as the hypothesis testing techniques can perform poorly as they favor a Type I specification, thus resulting in biased regression parameter estimation. In Chapters 2 and 3, the simulation studies and application example analyzed marginal models with continuous outcomes. However, the estimation approach and selection criterion are applicable to marginal generalized linear models in general, regardless of the outcome type, and subjects with varying repeated measurements are allowable. Future work is need to simultaneously select a working correlation structure, incorporated in the modified GEE, and time-dependent covariate type in order to improve regression parameter estimation. In small-sample settings, adjustments to covariance estimators from the GMM, modified QIF, and modified GEE approaches may further be considered to correct for negative bias.

To improve regression parameter estimation in marginal quantile regression for longitudinal data, we first proposed a modified approach to account for all valid moment conditions. Compared to the approach incorporating an independence working correlation structure, the proposed approach was more efficient. We then extended the selection method from Chapter 3 to determine the working type of covariate, which likely is unknown to the data analyst, and through a simulation study we presented that our method was preferable to the approach with an independence structure. In the application example, the use of a marginal quantile analysis, along with our proposed approach, provided a complete description of the morbidity index distribution to model the response variable, rather than the marginal mean analysis which advocated for the use of a working Type I specification. Future work can be done by simultaneously selecting a working correlation structure, incorporated in the modified estimating equations approach, and deciding a covariate type of time-dependency.

Although we only considered specific working correlation structures in the simulation studies and application example, other structures with less parsimonious forms are available as well. In addition, our simulation studies and application example were analyzed via marginal analysis with balanced repeated measurements. Nonetheless, all the proposed approaches in this dissertation are applicable to subjects with varying repeated measurements.

Bibliography

[1] Lai TL, Small D. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B* 2007; **69**:79–99.

[2] Lalonde TL, Wilson JR, Yin J. Gmm logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Statistics in Medicine* 2014; **33**:4756–4769.

[3] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.

[4] Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation* 1994; **23**:939–951.

[5] Fitzmaurice GM. A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics* 1995; **51**:309–317.

[6] Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika* 2003; **90**:29–41.

[7] Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica* 1982; **50**:1029–1054.

[8] Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**:823–836.

[9] Zhou Y, Lefante J, Rice J, Chen S. Using modified approaches on marginal regression analysis of longitudinal data with time-dependent covariates. *Statistics in Medicine* 2014; **33**:3354–3364.

[10] Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 2009; **28**:642–658.

[11] Leung DHY, Small DS, Qin J, Zhu M. Shrinkage empirical likelihood estimator in longitudinal analysis with time-dependent covariates–application to modeling the health of filipino children. *Biometrics* 2013; **69**:624–632.

[12] Owen A. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988; **75**:237–249.

[13] Fu L, Wang YG, Zhu M. A gaussian pseudolikelihood approach for quantile regression with repeated measurements. *Computational Statistics and data Analysis* 2015; **84**:41–53.

[14] Windmeijer F. A finite sample correction for the variance of linear two-step gmm estimators. *Working Paper Series No. W00/19*, Institute for Fiscal Studies, London 2000.

[15] Windmeijer F. A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of Econometrics* 2005; **126**:25–51.

[16] Westgate PM, Braun TM. The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions. *Statistics in Medicine* 2012; **31**:2209–2222.

[17] Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. *Statistics in Medicine* 2012; **31**:4003–4022.

[18] Westgate PM. Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach. *Biometrical Journal* 2014; **56**:461–476.

[19] Han P, Song PXK. A note on improving quadratic inference functions using a linear shrinkage approach. *Statistics and Probability Letters* 2011; **81**:438–445.

[20] Small CG, McLeish DL. *Hilbert Space Methods in Probability and Statistical Inference*. New York: Wiley, 1994.

[21] Qu A, Lindsay BG. Building adaptive estimating equations when inverse of covariance estimation is difficult. *Journal of the Royal Statistical Society: Series B* 2003; **65**:127–142.

[22] Cho GY, Dashnyam O. Upgraded quadratic inference functions for longitudinal data with type ii time-dependent covariates. *Journal of the Korean Data and Information Science Society* 2014; **25**:211–218.

[23] Westgate PM. A comparison of utilized and theoretical covariance weighting matrices on the estimation performance of quadratic inference functions. *Communications in Statistics - Simulation and Computation* 2014; **43**:2432–2443.

[24] Mancl LA, DeRouen TA. A covariance estimator for gee with improved small-sample properties. *Biometrics* 2001; **57**:126–134.

[25] Moore EH. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society* 1920; **26**:394–395.

[26] Penrose R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society* 1955; **51**:406–413.

[27] Westgate PM. A covariance correction that accounts for correlation estimation to improve finite-sample inference with generalized estimating equations: a study

on its applicability with structured correlation matrices. *Journal of Statistical Computation and Simulation* 2016; **86**:1891–1900.

[28] Westgate PM. A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. *Statistics in Medicine* 2013; **32**:2850–2858.

[29] Westgate PM. Improving the correlation structure selection approach for generalized estimating equations and balanced longitudinal data. *Statistics in Medicine* 2014; **33**:2222–2237.

[30] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011. URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

[31] Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. *The Analysis of Longitudinal Data*. 2nd edn., New York: Oxford University Press, 2002.

[32] Bouis HE, Haddad LJ. Effects of agricultural commercialization on land tenure, household resource allocation, and nutrition in the philippines. *Research Report 79*, International Food Policy Research Institute, Washington DC 1990.

[33] Bhargava A. Modelling the health of filipino children. *Journal of the Royal Statistical Society: Series A* 1994; **157**:417–432.

[34] Westgate PM. A bias-corrected covariance estimator for improved inference when using an unstructured correlation with quadratic inference functions. *Statistics and Probability Letters* 2013; **83**:1553–1558.

[35] Newey WK, Smith RJ. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica* 2004; **72**:219–255.

[36] Chen IC, Westgate PM. Improved methods for the marginal analysis of longitudinal data in the presence of time-dependent covariates. *Statistics in Medicine* 2017; **36**:2533–2546.

[37] Newey WK. Higher order properties of gmm and generalized empirical likelihood estimators. *Journal of Econometrics* 1985; **29**:229–256.

[38] Hall AR. Introduction to the generalized method of moments estimation. *Generalized Method of Moments Estimation*, Mátyás L (ed.). New York: Cambridge University Press, 1999.

[39] Conneely KN, Boehnke M. So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *American Journal of Human Genetics* 2007; **81**:1158–1168.

[40] Moulton LH, Zeger SL. Analyzing repeated measures on generalized linear models via the bootstrap. *Biometrics* 1989; **45**:381–394.

[41] Sherman M, le Cessie S. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics-Simulation and Computation* 1997; **26**:901–925.

[42] Koenker R, Bassett G. Regression quantiles. *Econometrica* 1978; **46**:33–50.

[43] Chen L, Wei LJ, Parzen MI. Quantile regression for correlated observations, in: Proceedings of the second seattle symposium in biostatistics: Analysis of correlated data. 2004; **179**:51–70.

[44] Yin G, Cai J. Quantile regression models with multivariate failure time data. *Biometrics* 2005; **61**:151–161.

[45] Wang HJ, Zhu Z. Empirical likelihood for quantile regression model with longitudinal data. *Journal of Statistical Planning and Inference* 2011; **141**:1603–1615.

[46] Tang CY, Leng C. Empirical likelihood and quantile regression in longitudinal data analysis. *Biomerika* 2011; **98**:1001–1006.

[47] Fu L, Wang YG. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and data Analysis* 2012; **56**:2526–2538.

[48] Leng C, Zhang W. Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing* 2014; **24**:123–136.

[49] Lu X, Fan Z. Weighted quantile regression for longitudinal data. *Computational Statistics* 2015; **30**:569–592.

[50] Jung SH. Quasi-likelihood for median regression models. *Journals of American Statistical Association* 1996; **91**:251–257.

[51] Stoner JA, Leroux BG. Analysis of clustered data: a combined estimating equations approach. *Biometrika* 2002; **89**:567–578.

[52] Chen IC, Westgate PM. A novel approach to selecting classification types for time-dependent covariates for the marginal analysis of longitudinal data (submitted).

[53] Brown BM, Wang YG. Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* 2005; **92**:149–158.

[54] Pang L, Lu W, Wang HJ. Variance estimation in censored quantile regression via induced smoothing. *Computational Statistics and Data Analysis* 2010; **56**:785–796.

[55] Yang CC, Chen YH, Chang HY. Composite marginal quantile regression analysis for longitudinal adolescent body mass index data. *Statistics in Medicine* 2017; **36**:3380–3397.

Vita

Education

M.S., Statistics, 2013

Colorado State University, Fort Collins, Colorado

M.S., Public Health with a concentration in Biostatistics, 2007

National Cheng Kung University, Tainan, Taiwan

B.B.A., Statistics, 2005

National Cheng Kung University, Tainan, Taiwan

Professional Experience

Graduate Teaching Assistant and Research Assistant

August, 2014 - Present

Department of Biostatistics, University of Kentucky, Lexington, Kentucky

Graduate Teaching Assistant and Statistical Consultant

September, 2012 - December, 2013

Department of Statistics, Colorado State University, Fort Collins, Colorado

Research Assistant

November, 2008 - July, 2011

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Graduate Teaching Assistant and Research Assistant

September, 2005 - June, 2007

Department of Public Health, National Cheng Kung University, Tainan, Taiwan

Awards

Highest Score on the 2016 Ph.D. Comprehensive Examination in Epidemiology and Biostatistics, College of Public Health, University of Kentucky

Graduate School Student Travel Funding, University of Kentucky (Summer, 2016)

Student Travel Funding, Department of Biostatistics, University of Kentucky (Summer, 2017)

Manuscripts

Chen I-C, Westgate PM. Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates. *Statistics in Medicine* 2017; **36**: 2533–2546. `http://onlinelibrary.wiley.com/doi/10.1002/sim.7307/full`

Yang H-C, **Chen I-C**, Tsay Y-C, Li Z-R, Chen C-h, Hwu H-G, Chen C-H. Using an Event-History with Risk-Free Model to Study the Genetics of Alcoholism. *Scientific Reports* 2017; **7**: 1975. `https://www.nature.com/articles/s41598-017-01791-4`

Fiorillo CE, Hughes AL, **Chen I-C**, Westgate PM, Gal TJ, Bush ML, Comer BT. Factors Associated with Patient No-Show Rates in an Academic Otolaryngology Practice. *The Laryngoscope* 2018; **128**: 626–631. `http://onlinelibrary.wiley.com/doi/10.1002/lary.26816/full`

Strickland JC, **Chen I-C**, Wang C, Fardo DW. Longitudinal Data Methods for Evaluating Genome by Epigenome Interactions in Families. *BMC Genetics* (accepted).

Chen I-C, Westgate PM. A Novel Approach to Selecting Classification Types for Time-Dependent Covariates for the Marginal Analysis of Longitudinal Data (submitted).

Chen I-C, Westgate PM. Marginal Quantile Regression for Longitudinal Data Analysis in the Presence of Time-Dependent Covariates (in preparation).

Bunn TL, Singleton MD, **Chen I-C**. Concordance of Crash, Fatality Analysis Reporting System, and Death Certificate Datasets in Identification of Drugs in Fatally Injured Motor Vehicle Drivers (in preparation).

Presentations

Oral Presentations

Chen I-C, Westgate PM. "Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates." The Student Research Symposium in the Kentucky Chapter – American Statistical Association (KY–ASA) Meeting. Spring 2016. Lexington, KY.

Chen I-C, Westgate PM. "Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates." The 2016 Joint Statistical Meetings (JSM) Annual Conference. Summer 2016. Chicago, IL.

Chen I-C, Westgate PM. "Selecting Classification Types for Time-Dependent Covariates to Improve the Marginal Analysis of Longitudinal Data." The Student Research Symposium in the Kentucky Chapter – American Statistical Association (KY–ASA) Meeting. Spring 2017. Louisville, KY.

Chen I-C, Westgate PM. "Selecting Classification Types for Time-Dependent Covariates to Improve the Marginal Analysis of Longitudinal Data." The 2017 Joint Statistical Meetings (JSM) Annual Conference. Summer 2017. Baltimore, MD.

Poster Presentations

Chen I-C, Yeh K-C, Liu I-T. "HIV/AIDS Knowledge, Attitude and Practices (KAP) Survey among New Recruiting Military Candidates in Certain Country of Southern Africa." The 40th Asia–Pacific Academic Consortium for Public Health (APACPH) Annual Conference. Spring 2008. Kuala Lumpur, Malaysia.

Chen I-C, Westgate PM. "Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates." The College of Public Health Research Day in Conjunction with the Center for Clinical and Translational Science (CCTS) Conference. Spring 2016. Lexington, KY.

Chen I-C, Westgate PM. "Improved Methods for the Marginal Analysis of Longitudinal Data in the Presence of Time-Dependent Covariates." The 2016 Joint Statistical Meetings (JSM) Annual Conference. Summer 2016. Chicago, IL.

Chen I-C, Westgate PM. "Selection Approaches of Time-Dependent Covariates for the Marginal Analysis of Longitudinal Data." The College of Public Health Research Day in Conjunction with the Center for Clinical and Translational Science (CCTS) Conference. Spring 2017. Lexington, KY.

Chen I-C, Westgate PM. "Selecting Classification Types for Time-Dependent Covariates to Improve the Marginal Analysis of Longitudinal Data." The 2017 Joint Statistical Meetings (JSM) Annual Conference. Summer 2017. Baltimore, MD.