Marginal Regression Analysis of Longitudinal Data With Time-Dependent Covariates: A Generalised Method of Moments Approach

Tze Leung Lai   Dylan Small

Stanford University   University of Pennsylvania

August 3, 2006

Summary. We develop a new approach to using estimating equations to estimate marginal regression models for longitudinal data with time-dependent covariates. Our approach classifies time-dependent covariates into three types – Types I, II and III. The type of covariate determines what estimating equations can be used involving the covariate. We use the generalised method of moments to make optimal use of the estimating equations made available by the covariates. Previous literature has suggested using generalised estimating equations (GEE) with the independent working correlation when there are time-dependent covariates. We conduct a simulation study that shows that our approach can provide substantial efficiency gains over GEE with the independent working correlation when a time-dependent covariate is of Types I or II while our approach remains consistent and comparable in efficiency to GEE with the independent working correlation when a time-dependent covariate is of Type III. We apply our approach to analyze the relationship between body mass index and future morbidity among children in the Philippines.

1. INTRODUCTION

Marginal regression models for longitudinal data seek to characterize the expectation of a subject's response y at time t as a function of the subject's covariates at time t. Marginal models differ from transition models that seek to characterize the expectation of a subject's response y at time t as a function of the subject's covariates at times $1, \ldots, t$ and the subject's past responses at times $1, \ldots, t-1$. Marginal models are appropriate when inferences about the population average are the primary focus (Diggle *et al.*, 2002) or when future applications of the results require the expectation of the response as a function of the current covariates (Pepe and Anderson, 1994). When there are time-dependent covariates, previous literature (e.g., Pepe and Anderson, 1994) has suggested that marginal models be estimated by generalised estimating equations (GEE) with the independent working correlation. In this paper, we develop a more efficient estimation approach for marginal regression models with time-dependent covariates. We classify time-dependent covariates into three types – Types I, II and III. We show that when a time-dependent covariate is of Type I or II, GEE with the independent working correlation does not exploit all of the available estimating equations involving that covariate. We develop an approach that makes efficient use of all the estimating equations made available by time-dependent covariates.

In longitudinal data sets, there is typically correlation among a subject's repeated measurements. In a marginal model, this correlation is not of primary interest but it must be taken into account to make proper inferences. For analyzing marginal models, the seminal paper of Liang and Zeger (1986) developed the generalised estimating equation (GEE) approach. In the GEE approach, a "working" correlation structure for the correlation among a subject's repeated measurements is postulated. A valuable feature of GEE with time-independent covariates is that it produces efficient estimates if the working correlation structure is correctly specified but remains consistent and provides correct standard errors if the working correlation structure is incorrectly specified. However, when there are time-dependent covariates, Hu (1993) and Pepe and Anderson (1994) have pointed out that the consistency of GEE is not assured with arbitrary working correlation structures unless a key assumption is satisfied. Consistency is, however, assured regardless of the validity of the key assumption when a working correlation that a subject's repeated measurements are independent (the independent working correlation) is employed. Consequently, Pepe and Anderson suggest using the independent working correlation when using GEE with time-dependent

covariates as a "safe" analysis choice.

In this paper we develop a new approach to marginal regression analysis with time-dependent covariates that provides more efficient estimates than GEE with the independent working correlation under certain conditions but maintains GEE with time-independent covariates' attractive feature of being consistent under all correlation structures for subjects' repeated measurements. GEE can be viewed as a method for combining certain estimating equations. When there are time-dependent covariates, some of the estimating equations combined by GEE with an arbitrary working correlation structure are not valid, explaining its potential inconsistency, whereas all of the estimating equations combined by GEE with the independent working correlation are valid, explaining its consistency (Pepe and Anderson, 1994; Pan *et al.*, 2000).

We distinguish between three types of time-dependent covariates (Types I, II and II) and show that when a covariate is of Type I or II, there are valid estimating equations available that are not exploited by GEE with the independent working correlation. Our approach makes use of these additional estimating equations not exploited by GEE with the independent working correlation. Type I and Type II time-dependent covariates are covariates for which there is no "feedback" from the response process to the covariate process. For Type I and Type II covariates, the additional valid estimating equations not used by GEE with the independent working correlation involve future values of the covariate being uncorrelated with current residuals from the marginal regression model. Type I time-dependent covariates have the additional feature that past values of the covariate are uncorrelated with current residuals. In order to make optimal use of the valid estimating equations, we use the generalised method of moments (Hansen, 1982). We also provide a test for whether a time-dependent covariate is of a certain type. In a simulation study, we show that our approach provides substantial efficiency gains over GEE with the independent working correlation when a time-dependent covariate is of Types I or II, while our approach remains consistent and comparable in efficiency to GEE with the independent working correlation when a time-dependent covariate is of Type III. We apply our approach to analyze the association between body mass index and future morbidity among children in the rural Philippines.

Our paper is organized as follows. In Section 2, we describe the model for longitudinal data that we consider, review the GEE approach of Liang and Zeger (1986) and review the issues that arise in using GEE when there are time-dependent covariates. In Section 3, we

3

provide our new approach to estimation of marginal regression models for longitudinal data with time-dependent covariates. In Section 4, we conduct a simulation study to demonstrate the value of our approach. In Section 5, we apply our approach to predicting morbidity among children in the Bukidnon region in the Philippines based on body mass index. Section 6 provides discussion.

2. MARGINAL REGRESSION MODELS FOR LONGITUDINAL DATA

The data structure we consider is a longitudinal data set comprised of an outcome variable y_{it} and a $p \times 1$ vector of covariates $\mathbf{x}_{it}$, observed at times $t = 1, \ldots, T$ for subjects $i = 1, \ldots, N$. We do not require that all subjects be observed at all time points, but we do require that the number of time points T at which subjects are observed is small relative to N. Such longitudinal data sets in which measurements are collected at regular times are common in public health and social science research. We assume that if there is missing data, whether a subject's data is missing at a given time point t is conditionally independent, given the subject's covariates at time point t ($\mathbf{x}_{it}$), of the subject's missing outcomes, past outcomes, future outcomes and covariates at past or future time points (this is a special case of the missing completely at random assumption described in Little and Rubin, 2002). For simplicity of notation, we will assume that each subject is observed at each time point and note later how our procedures are easily adapted when some subjects are not observed at all time points.

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ be the $T \times 1$ vector of outcome values associated with $p \times 1$ covariate vectors $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ for the ith subject and let $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$. For $i \neq j$, we assume $\mathbf{y}_i$ and $\mathbf{y}_j$ are independent, but generally the components of $\mathbf{y}_i$ are correlated. The marginal density of y_{it} is assumed to follow a generalised linear model (McCullaugh and Nelder, 1989) of the form

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}/\phi],$$

where $\theta_{it} = h(\mathbf{x}_{it}^T \boldsymbol{\beta})$ for some monotone link function h and

$$E(y_{it} \mid \mathbf{x}_{it}) \equiv \mu_{it} = a'(\theta_{it}) = a'(h(\mathbf{x}_{it}^T \boldsymbol{\beta})), \tag{1}$$

$$\text{var}(y_{it}) = \ddot{a}(\theta_{it})\phi. \tag{2}$$

Let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})'$; μ_{it} is the marginal mean of y_{it} conditional on $\mathbf{x}_{it}$.

4

2.1 *Generalised estimating equations*

The GEE method, introduced by Liang & Zeger (1986) for estimating the parameter vector $\boldsymbol{\beta}$ of the marginal regression model (1)-(2), allows the user to specify any "working" correlation structure for the correlation matrix of a subject's outcomes $\mathbf{y}_i$. The working correlation structure can depend on an unknown $s \times 1$ parameter vector $\boldsymbol{\alpha}$. The observation times and correlation matrix can differ from subject to subject, but the correlation matrix $R_i(\boldsymbol{\alpha})$ of the ith subject is fully specified by $\boldsymbol{\alpha}$. The working covariance matrix of $\mathbf{y}_i$ is given by $V_i(\boldsymbol{\alpha}) = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2} \phi$, where $A_i = \text{diag}\{\ddot{a}(\theta_{it})\}$. The generalised estimating equations are $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)' V_i^{-1}(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})))(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})). \tag{3}$$

The form of the estimating equations (3) is motivated by the fact that when $R_i(\boldsymbol{\alpha}) = I$, (3) are the score equations from a likelihood analysis that assumes that the repeated observations from a subject are independent of one another. Liang & Zeger (1986) established the following properties of the estimator $\hat{\boldsymbol{\beta}}$ that satisfies $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ under the assumption that the estimating equation is asymptotically unbiased in the sense that

$$\lim_{N \to \infty} E_{\boldsymbol{\beta}_0}[U(\boldsymbol{\beta}_0)] = \mathbf{0}, \tag{4}$$

and suitable regularity conditions: (i) $\hat{\boldsymbol{\beta}}$ is consistent regardless of whether the actual correlation matrix of $\mathbf{y}_i$ is $R_i(\boldsymbol{\alpha})$; and (ii) $\text{cov}(\hat{\boldsymbol{\beta}})$ can be consistently estimated regardless of whether the actual correlation matrix of $\mathbf{y}_i$ is $R_i(\boldsymbol{\alpha})$. Although correct specification of the working correlation structure does not affect consistency, correct specification enhances efficiency. Note that in asymptotic statements throughout the paper such as (4), the number of subjects N is assumed to increase to infinity but the number of time points at which subjects are observed T is assumed to remain fixed.

2.2 *Problems posed by time-dependent covariates*

For time-independent covariates, the validity of assumption (1) about the marginal mean implies that (4) holds as we shall show below and consequently the GEE estimate of $\boldsymbol{\beta}$ is consistent regardless of the choice of working correlation structure. However, for time-dependent covariates, assumption (4) might not hold for arbitrary working correlation structures and consequently the GEE estimate of $\boldsymbol{\beta}$ is not necessarily consistent (Hu, 1993; Pepe and Anderson, 1994; for further discussion, see Emond *et al.*, 1997, Pan *et al.*, 2000 and Diggle *et*

al., 2002, Section 12.3). To review this result, assume that $\lim_{N\to\infty} V_i^{-1}(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \hat{\phi}(\boldsymbol{\beta}_0))) = H_i$ so that

$$\lim_{N\to\infty} E_{\boldsymbol{\beta}_0}[U_j(\boldsymbol{\beta}_0)] = E_{\boldsymbol{\beta}_0}\left[\sum_{t=1}^{T}\sum_{s=1}^{T} H_{i[st]}\frac{\partial\mu_{is}(\boldsymbol{\beta}_0)}{\partial\beta_j}(y_{it}-\mu_{it}(\boldsymbol{\beta}_0))\right],$$

where $U_j(\boldsymbol{\beta})$ denotes the jth equation in (3). Assumption (1) implies that

$$E_{\boldsymbol{\beta}_0}[H_{i[tt]}\frac{\partial\mu_{it}(\boldsymbol{\beta}_0)}{\partial\beta_j}(y_{it}-\mu_{it}(\beta_0))] = 0 \tag{5}$$

for $t = 1,\ldots,T$, but assumption (1) does not guarantee that

$$E_{\boldsymbol{\beta}_0}[H_{i[st]}\frac{\partial\mu_{is}(\boldsymbol{\beta}_0)}{\partial\beta_j}(y_{it}-\mu_{it}(\beta_0))] = 0 \tag{6}$$

for all s, t and consequently (4) does not necessarily hold. Pepe and Anderson provide the following example of a model with time-dependent covariates in which (6) does not hold for all s, t: $y_{it} = \alpha y_{i,t-1} + \beta x_{it} + u_{it}$, $(x_{i1},\ldots,x_{iT_i},u_{i1},\ldots,u_{iT_i}) \sim N(\mathbf{0}, I)$, and $y_{i0} = 0$. Under this model, $E(y_{it} \mid x_{it}) = \beta x_{it}$ but $E(x_{is}(y_{it}-\beta x_{it})) = \alpha^{t-s}\beta$ for $s \le t - 1$.

A sufficient condition for (6) to hold for all s, t is

$$E(y_{it} \mid \mathbf{x}_{it}) = E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) \text{ for } t = 1,\ldots T, \tag{7}$$

because under (1) and (7),

$$
\begin{aligned}
E\left[\frac{\partial\mu_{is}(\boldsymbol{\beta})}{\partial\beta_j}(y_{it}-\mu_{it}(\boldsymbol{\beta}))\right] &= E\left[E\left[\frac{\partial\mu_{is}(\boldsymbol{\beta})}{\partial\beta_j}(y_{it}-\mu_{it}(\boldsymbol{\beta}))|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}\right]\right] \\
&= E\left[\frac{\partial\mu_{is}(\boldsymbol{\beta})}{\partial\beta_j}E\left[y_{it}-\mu_{it}(\boldsymbol{\beta})|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}\right]\right] \\
&= E\left[\frac{\partial\mu_{is}(\boldsymbol{\beta})}{\partial\beta_j}E\left[y_{it}-\mu_{it}(\boldsymbol{\beta})|\mathbf{x}_{it}\right]\right] \\
&= 0,
\end{aligned}
$$

where the third equality follows from (7) and the fourth equality from (1). Note that if all covariates are time-independent, then (7) holds and consequently the GEE estimator of $\boldsymbol{\beta}$ is consistent regardless of the working correlation structure.

Although the GEE estimator is not necessarily consistent with an arbitrary working correlation structure when there are time-dependent covariates, the GEE estimator is consistent with the independent working correlation $(R = I)$. This is because with the independent working correlation, H_i is a diagonal matrix and therefore (see (5))

$$\lim_{N\to\infty} E_{\boldsymbol{\beta}_0}[U_j(\boldsymbol{\beta}_0)] = E_{\boldsymbol{\beta}_0}\left[\sum_{t=1}^{T} H_{i[tt]}\frac{\partial\mu_{it}(\boldsymbol{\beta}_0)}{\partial\beta_j}(y_{it}-\mu_{it}(\boldsymbol{\beta}_0))\right] = 0.$$

6

3. A NEW APPROACH TO TIME-DEPENDENT COVARIATES

Because the GEE estimator of $\boldsymbol{\beta}$ with the independent working correlation is always consistent, Pepe and Anderson (1994) recommend using the independent working correlation as a "safe" analysis choice. The independent working correlation often has high efficiency for estimation of coefficients associated with time-independent covariates (Fitzmaurice, 1995). However, for time-dependent covariates, Fitzmaurice (1995) shows that the independent working correlation can result in a substantial loss of efficiency for estimation of the coefficients associated with the time-dependent covariates and provides an example in which using the independent working correlation is only 60% efficient relative to the true correlation structure. This creates a dilemma for time-dependent covariates: use of the independent working correlation may be very inefficient but use of a non-independent working correlation structure may produce inconsistent estimates. To resolve this dilemma, Pan and Connett (2002) propose to use resampling-based methods to choose the estimator which best predicts y_{it} based on the covariates $\mathbf{x}_{it}$ among GEE estimators based on different working correlation structures. Although their method is capable of choosing the best estimator among a class of GEE estimators, for certain types of time-dependent covariates, there are available valid estimating equations that are not exploited by the usual GEE estimators. In this section, we develop an estimator that takes advantage of these additional estimating equations when appropriate to produce consistent and more efficient estimates than the class of usual GEE estimators. A key component in our approach is a classification of time-dependent covariates into three types.

3.1 *Classification of Time-Dependent Covariates*

Let x^j, $j = 1, \ldots, p$, denote the jth covariate. We classify time-dependent covariates into three types.

Type I: We classify a time-dependent covariate x^j as being of Type I if it satisfies

$$E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] = 0 \ \text{ for all } s, t, \ s = 1, \ldots, T, t = 1, \ldots, T. \tag{8}$$

A sufficient condition for all covariates to be of Type I is that (7) holds; see the argument below (7). For a linear model, a sufficient condition for a covariate x^j to be of Type I is that

$$f(x_{i1}^j, \ldots, x_{iT}^j | y_{it}, \mathbf{x}_{it}) = f(x_{i1}^j, \ldots, x_{iT}^j | \mathbf{x}_{it}). \tag{9}$$

Variables that plausibly satisfy (9) include age, time variables and the treatment assignment for subject i at time t in a randomized crossover trial.

7

Type II: We classify a time dependent covariate x^j as being of Type II if it satisfies

$$E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] = 0 \ \text{ for all } s \geq t, \ t = 1, \ldots, T. \tag{10}$$

Note that the class of Type I covariates is a subset of the class of Type II covariates. A sufficient condition for all covariates to be of Type II is that

$$f((\mathbf{x}_{i,t+1}, \ldots, \mathbf{x}_{iT})|y_{it}, \mathbf{x}_{it}) = f(\mathbf{x}_{i,t+1}, \ldots, \mathbf{x}_{iT}|\mathbf{x}_{it}). \tag{11}$$

Condition (11) says that the time-dependent covariate process $\mathbf{x}_{i,t+1}, \ldots, \mathbf{x}_{iT}$ is not affected by the response y_{it} at time t conditional on $\mathbf{x}_{it}$, i.e., it rules out "feedback" from the response process to the covariate process. The reason that (11) implies (10) is that under condition (11), for $s \geq t$,

$$
\begin{aligned}
E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] &= E_{\boldsymbol{\beta}_0}\left[E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))|\mathbf{x}_{it}\right]\right] \\
&= E_{\boldsymbol{\beta}_0}\left[E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}|\mathbf{x}_{it}\right] E_{\boldsymbol{\beta}_0}\left[y_{it} - \mu_{it}(\boldsymbol{\beta}_0)|\mathbf{x}_{it}\right]\right] \\
&= 0,
\end{aligned}
$$

where the second equality follows from condition (11) and the third equality follows from (1). For a linear model, a sufficient condition for x^j to be of Type II is that

$$f((x^j_{i,t+1}, \ldots, x^j_{iT})|y_{it}, \mathbf{x}_{it}) = f((x^j_{i,t+1}, \ldots, x^j_{iT})|\mathbf{x}_{it}). \tag{12}$$

Condition (11) is related to but not the same as the condition of a covariate process being exogenous with respect to the outcome process that is defined by Diggle *et al.* (2002, Section 12.1); condition (11) differs from Diggle *et al.*'s definition in that the left hand side of (11) conditions on y_{it} rather than $y_{i1}, \ldots, y_{it}$. Diggle *et al.*'s definition of exogeneity of a covariate process is equivalent to Chamberlain's (1982) and Engle, Hendry and Richards' (1983) extensions of Granger's (1969) definition that y does not Granger-cause $\mathbf{x}$. Chamberlain (1982) shows that if $\mathbf{x}$ is exogenous with respect to the outcome process in the definition of Diggle *et al.*, then the covariate process satisfies (11).

Type III: We classify a time-dependent covariate x^j to be of Type III if it is not of Type II, i.e.,

$$E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] \neq 0 \ \text{ for some } s > t. \tag{13}$$

As an example of the distinction between types of time-dependent covariates, consider the study presented by Zeger and Liang (1991) of infectious diseases and vitamin A deficiency in Indonesian children. For the outcome of diarrheal disease and the time-dependent covariate xerophthalmia (an ocular condition due to vitamin A deficiency), Zeger and Liang find evidence that xerophthalmia is a Type III covariate – there is a feedback cycle in which xerophthalmia increases the risk of diarrheal disease which further increases the risk of future xerophthalmia. In contrast, for the outcome of respiratory disease and the same covariate of xerophthalmia, there is no evidence of a feedback cycle and hence it is plausible that xerophthalmia is a Type II covariate (Diggle *et al.*, 2002).

3.2 *Generalised Method of Moments*

Consider using estimating equations to estimate $\boldsymbol{\beta}$. The assumption (1) about the marginal mean function provides the basis for Tp estimating equations for the p dimensional parameter $\boldsymbol{\beta}$:

$$E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{it}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] = 0, \quad t = 1,\ldots,T, j = 1,\ldots,p. \qquad (14)$$

Because there are more estimating equations (Tp) than parameters (p), these estimating equations must be combined in some way to form p estimating equations. GEE is one method for combining estimating equations. For an arbitrary working correlation structure that allows nonzero correlations between any two time points, GEE uses (3) to combine the T^2p estimating equations (8) for $j = 1,\ldots,p$. Because some of these estimating equations are not valid for Type II and Type III covariates, use of an arbitrary working correlation structure may produce inconsistent estimates. However, for the independent working correlation, GEE combines only the estimating equations (14) and is hence consistent as long as (1) is correct. This phenomenon is what prompts the suggestion of Pepe and Anderson (1994) to use the independent working correlation as a safe analysis choice when working with time-dependent covariates.

For Type I covariates, Pan and Connett's resampling method safely enhances efficiency by providing an approach to decide when a covariate is of Type I versus Type II or III and using a more efficient working correlation structure when there is evidence that a covariate is of Type I. When a time-dependent covariate is of Type II, the only working correlation structure that is guaranteed to produce consistent estimates is the independent working correlation so that Pan and Connett's method will tend to use GEE with the independent

9

working correlation. However, the following estimating equations are valid for a Type II covariate, but are not used by GEE with the independent working correlation:

$$E_{\boldsymbol{\beta}_0}\left[\frac{\partial \mu_{is}(\boldsymbol{\beta}_0)}{\partial \beta_j}(y_{it} - \mu_{it}(\boldsymbol{\beta}_0))\right] = 0, \quad s > t, \ t = 1, \ldots, T, j = 1, \ldots, p. \tag{15}$$

Here we propose the use of the generalised method of moments (GMM) as a method to optimally combine all available valid estimating equations.

GMM was introduced by Hansen (1982) and has had a large influence in econometrics; Mátyás (1999) provides an overview of GMM methods and their applications. GMM has roots in the minimum χ^2 method (Neyman, 1949; Ferguson, 1958); Lindsay and Qu (2003) provide good discussion of the connection between GMM and other ideas in statistics. Qu, Lindsay and Li (2000) apply GMM to marginal regression analysis of longitudinal data to improve usual GEE estimates for Type I covariates whereas here our focus is on improving estimates for Type II covariates.

The use of GMM to estimate the p-dimensional parameter $\boldsymbol{\beta}$ requires an $r \geq p$ vector $\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$ of "valid" moment conditions where valid means

$$E_{\boldsymbol{\beta}_0}[\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}_0)] = \mathbf{0}. \tag{16}$$

For a Type I time-dependent covariate, we have the T^2 valid moment conditions (8); for a Type II time-dependent covariate, we have the $T(T+1)/2$ valid moment conditions (10); and for a Type III time-dependent covariate or a time-independent covariate, we have the T valid moment conditions (14). The valid moment conditions for each of the p covariates are combined into the vector $\mathbf{g}$. The sample version of (16) is

$$G_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}). \tag{17}$$

The usual method of moments would estimate $\boldsymbol{\beta}$ by solving $G_N(\boldsymbol{\beta}) = \mathbf{0}$, but when $r > p$, this system of equations is overdetermined. To overcome this difficulty, Hansen introduces a positive definite weight matrix W_N and minimizes a quadratic form in deviations between the sample moments and the population moment conditions. The GMM estimator is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}), \quad Q_N(\boldsymbol{\beta}) = G_N(\boldsymbol{\beta})^T W_N G_N(\boldsymbol{\beta}).$$

Equivalently, the GMM estimator solves the estimating equation

$$\left(\frac{\partial G_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T W_N G_N(\boldsymbol{\beta}) = \mathbf{0}.$$

Thus, the GMM estimator can be thought of as a method of moments estimator for a p-dimensional set of moment conditions which is formed by taking linear combinations of the $r > p$ available moment conditions. The relative importance of each of the r original moment conditions in the GMM estimator is determined by its informativeness about $\boldsymbol{\beta}$, measured by $\frac{\partial G_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, and the weight matrix W_N.

Hansen (1982) shows that the asymptotically optimal choice of the weight matrix W_N is the inverse of the covariance matrix of the moment conditions $\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta_0})$, denoted by $V^{-1} = \{Cov(\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}_0))\}^{-1}$ (where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$). The intuition is that this weight matrix gives less weight to those sample moment conditions with large variances (Qu, Lindsay and Li, 2000). Furthermore, Hansen shows that the same asymptotic efficiency as using the optimal W_N is obtained by the two step procedure of using an initial consistent estimator $\tilde{\boldsymbol{\beta}}$ to obtain a consistent estimate of V^{-1}, $\hat{V}_N^{-1}$, and then estimating $\boldsymbol{\beta}$ by GMM with the weight matrix $\hat{V}_N^{-1}$. We denote this two step estimator by $\hat{\boldsymbol{\beta}}_{GMM}$. In our case, we use GEE with the independent working correlation as the initial estimator $\tilde{\boldsymbol{\beta}}$ and estimate V^{-1} by

$$\hat{V}_N^{-1} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \tilde{\boldsymbol{\beta}}) \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \tilde{\boldsymbol{\beta}})^T \right\}^{-1}. \tag{18}$$

Under suitable regularity conditions, the estimator $\hat{\boldsymbol{\beta}}_{GMM}$ is asymptotically normal (as $N \to \infty$) with asymptotic variance

$$\left\{ E\left(\frac{\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T V^{-1} E\left(\frac{\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \right\}^{-1},$$

where $\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The asymptotic variance can be consistently estimated by

$$\left\{ \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \hat{V}_N^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \right\}^{-1} \tag{19}$$

where $\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{GMM}$ (Hansen, 1982). Under regularity conditions, the two step GMM estimator is semiparametrically efficient in the sense of Bickel, Klaasen, Ritov and Wellner (1993) for the family of distributions satisfying the moment conditions (16) (Chamberlain, 1987).

When not all subjects are observed at each time point, then we modify the above procedure as follows. For a moment condition $g_j(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$, let I_j denote the set of subjects

11

whose observation times allow $g_j(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$ to be computed, e.g., for a moment condition of the form (10), I_j is the set of subjects observed at both times s and t. Instead of (17), the sample moment conditions can be expressed as

$$G_N(\boldsymbol{\beta}) = \left(\frac{1}{|I_1|} \sum_{i \in I_1} g_1(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}), \ldots, \frac{1}{|I_r|} \sum_{i \in I_r} g_r(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}) \right)',$$

where $|I_j|$ denotes the number of subjects in I_j. Instead of (18), our estimate of V^{-1} is $\hat{V}_N^{-1} = \left\{ \frac{1}{|I_{1,\ldots,r}|} \sum_{i \in I_{1,\ldots,r}} \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \tilde{\boldsymbol{\beta}}) \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \tilde{\boldsymbol{\beta}})^T \right\}^{-1}$, where $I_{1,\ldots,r}$ is the set of subjects whose observation times allow $g_j(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$ to be computed for all $j = 1, \ldots, r$.

Comparing GMM to GEE, when both methods are combining the same estimating equations, GMM has the same asymptotic efficiency as GEE if the working correlation structure in GEE is correctly specified but GMM is more asymptotically efficient when the working correlation in GEE is misspecified (Qu, Lindsay and Li, 2000). An additional advantage of the GMM approach is that it facilitates combining a set of estimating equations such as (10) (which are valid for a Type II covariate). No GEE working correlation structure can combine the estimating equations (10) in such a way that at least one of the estimating equations with $s \neq t$ in (10) has nonzero weight (for such a working correlation structure, $R_i(\boldsymbol{\alpha})$ would have to be upper triangular with at least one nonzero element off the diagonal, which would mean that $R_i(\boldsymbol{\alpha})$ is not a correlation matrix – a contradiction). Although GMM provides the flexibility to combine the estimating equations in (10), if some of the estimating equations in (10) are invalid so that the covariate is in fact of Type III, then the GMM estimator that combines the estimating equations in (10) will be inconsistent. Consequently, it is important to be able to test whether a covariate is of Type II vs. Type III (or Type I vs. Type II or III).

3.3 *Testing a Covariate's Type*

One approach to testing is to test the sufficient condition (11) (or (12) for a linear model). Another approach to testing works directly in the GMM framework. Consider testing

$$H_0 : E_{\boldsymbol{\beta}_0} [\mathbf{g}_a(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0} \text{ and } E_{\boldsymbol{\beta}_0} [\mathbf{g}_b(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0} \tag{20}$$

vs.

$$H_a : E_{\boldsymbol{\beta}_0} [\mathbf{g}_a(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0} \text{ and } E_{\boldsymbol{\beta}_0} [\mathbf{g}_b(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})] \neq \mathbf{0}, \tag{21}$$

where $(\mathbf{g}'_a, \mathbf{g}'_b)' = \mathbf{g}$ and $\mathbf{g}_a$ has dimension $q \geq p$. A test of H_0 vs. H_a can be based on the statistic

$$C_N = N\{Q_{ab,N}(\hat{\boldsymbol{\beta}}_{GMM,ab}) - Q_{a,N}(\hat{\boldsymbol{\beta}}_{GMM,a})\}, \tag{22}$$

where $Q_{ab,N}(\hat{\boldsymbol{\beta}}_{GMM,ab})$ is the GMM minimand using the full set of moment conditions in H_0 and $Q_{a,N}(\hat{\boldsymbol{\beta}}_{GMM,a})$ is the GMM minimand using only the moment conditions $\mathbf{g}_a$ (where $\hat{\boldsymbol{\beta}}_{GMM,a}$ is estimated using only the moment conditions $\mathbf{g}_a$) (Eichenbaum, Hansen and Singleton, 1988). C_N has an asymptotic χ^2_{r-q} distribution under H_0, where $r - q$ is the dimension of $\mathbf{g}_b$. The statistic C_N is similar in spirit to the likelihood ratio statistic from maximum likelihood theory (Hall, 1999). See Newey (1985) and Hall (1999) for further discussion of testing the validity of moment conditions in the GMM framework. In Appendix A, we present a set of regularity conditions under which the test of (20) vs. (21) based on (22) is consistent.

An alternative approach to assessing a covariate's type besides the test based on (22) is to examine the predictive performance of different estimators that are based on different assumptions about a covariate's type. Specifically, we can choose among the GMM estimators that make different assumptions about a covariate's type (which lead to different moment conditions) the estimator that minimizes a resampling based estimate of the predictive mean squared error for predicting y_{it} based on $\mathbf{x}_{it}$. Pan and Connett (2002) developed this predictive mean squared error approach for choosing among the class of usual GEE estimators when there are time-dependent covariates; see their paper for implementation details. The approach to determining a covariate's type by using the test based on (22) is useful when a researcher has a strong prior belief that the moment conditions in H_0 are all valid and is using the test to see if there is any evidence in the data against this belief. When a researcher has more uncertainty about the validity of some of the moment conditions in H_0, the predictive mean square error approach is useful. The approach of using the test based on (22) has the advantage of being computationally simpler. We focus hereafter on the approach using the test based on (22).

3.4 Overall Approach

Our overall GMM approach to marginal regression analysis with time-dependent covariates is as follows. Unless there are substantive reasons to think that a time-dependent covariate is of Type I or Type II, we assume it is of Type III and use the moment conditions

13

(14) for it. If there are substantive reasons to think that a covariate is of Type II (or Type I), then we test the null hypothesis that it is of Type II (or Type I) versus the alternative that it is of Type III and if the test is not rejected, we use the moment conditions (10) (or (8)) in our GMM estimator. We call the GMM estimator that chooses which moment conditions to use based on the test (22) the GMM Moment Selection estimator. The GMM Moment Selection estimator gains efficiency for Type I and Type II covariates compared to GEE with the independent working correlation when our hypothesis that a covariate is of Type I or Type II is correct. The GMM Moment Selection estimator remains consistent when our hypothesis that a covariate is of Type I or Type II is wrong as long as the test based on (22) is consistent (see Appendix A for regularity conditions under which the test is consistent). R functions that implement our approach are available from the authors.

4. SIMULATION STUDY

To examine the performance of our proposed GMM approach, we performed a small simulation study for two settings, the first setting has a Type II time-dependent covariate and the second setting has a Type III time-dependent covariate.

Setting I: Type II time-dependent covariate.

This setting was considered by Diggle *et al.* (2002). We simulated data under the following mechanism:

$$y_{it} = \gamma_0 + \gamma_1 x_{it} + \gamma_2 x_{i,t-1} + b_i + e_{it}$$

$$x_{it} = \rho x_{i,t-1} + \epsilon_{it}$$

$(b_i, e_{it}, \epsilon_{it}) \sim$ mutually independent normal mean zero with variances 4,1 and 1 respectively; the x_{it} process is stationary, i.e., $x_{i0} \sim N(0, \sigma_\epsilon^2/(1-\rho^2))$.

This model represents the plausible scenario that a time-dependent covariate has an autoregressive structure and a response variable depends on both current and lagged values of the covariate. The model yields the marginal mean

$$E(y_{it}|x_{it}) = \beta_0 + \beta_1 x_{it}, \tag{23}$$

where $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1 + \gamma_2\rho$. The mean that conditions on all past, present and future covariates, $E(y_{it}|x_{i1}, \ldots, x_{iT}) = \gamma_0 + \gamma_1 x_{it} + \gamma_2 x_{i,t-1}$, does not equal the marginal mean (23) so that (7) does not hold. Furthermore,

$$E\left[x_{is}(y_{it} - E(y_{it}|x_{it}))\right] = \begin{cases} 0 & \text{for } s \geq t \\ (\gamma_2\rho^{t-s-1} - \gamma_2\rho^{t-s+1})\frac{\sigma_\epsilon^2}{1-\rho^2} & \text{for } s < t \end{cases}$$

14

Thus, x_{it} is a Type II time-dependent covariate.

We consider six estimators: 1) GMM using the moment conditions (10) for x_{it}, labeled GMM Type II; 2) GMM using the moment conditions (14) for x_{it}, labeled GMM Type III; 3) the GMM Moment Selection estimator that uses the .05 level test based on (22) to decide between the GMM Type II and GMM Type III estimators; 4) GEE using the independent working correlation, labeled GEE Independence; 5) GEE using the exchangeable working correlation structure (see Liang and Zeger, 1986, pp. 18, Example 3), labeled GEE Exchangeable; and 6) GEE using the AR-1 working correlation structure (see Liang and Zeger, 1986, pp. 18, Example 4), labeled GEE AR-1. We simulated 2000 data sets, each of which contained $N = 500$ subjects observed at five time points with $\gamma_0 = 0$, $\gamma_1 = 1$, $\gamma_2 = 1$ and $\rho = 0.5$. Table 1 shows the bias, root mean square error and the efficiency of each estimator relative to the GEE Independence estimator (the efficiency is the ratio of the mean square error of GEE Independence to that of the estimator). The GEE Exchangeable and GEE AR-1 estimators exhibit substantial bias and perform poorly. The GEE Independence and GMM Type III estimators are unbiased and perform similarly. The GMM Type II estimator is unbiased and is substantially more efficient than GEE Independence (almost twice as efficient). The GMM Moment Selection estimator performs almost as well as the GMM Type II estimator and is more than 75% more efficient than GEE Independence. The nominal .05 level test (22) of the null hypothesis that x_{it} is a Type II time-dependent covariate (as opposed to Type III) rejected in .0665 of the simulations, showing that the asymptotic level of the test is reasonably reliable. Furthermore, a 95% confidence interval for β_1, formed using the asymptotic variance (19), covered the true β_1 in 93.2% of the simulations.

<center>INSERT TABLE 1 ABOUT HERE</center>

Setting II: Type III time-dependent covariate.

We consider the following model:

$$y_{it} = \beta x_{it} + \kappa y_{i,t-1} + u_{it} \tag{24}$$

$$x_{it} = \gamma y_{i,t-1} + v_{it}, \tag{25}$$

where $(u_{i1}, \ldots, u_{iT}, v_{i1}, \ldots, v_{iT})$ are mutually independent mean zero normal random variables with variances σ_u^2 for $(u_{i1}, \ldots, u_{iT})$ and σ_v^2 for $(v_{i1}, \ldots, v_{iT})$, and the (x_{it}, y_{it}) process is

<center>15</center>

stationary (i.e., $y_{i0} \sim N(0, \frac{\beta^2}{1-(\beta\gamma+\kappa)^2}\sigma_v^2 + \frac{1}{1-(\beta\gamma+\kappa)^2}\sigma_u^2)$. The marginal mean model is

$$E(y_{it}|x_{it}) = \left[\beta + (\kappa\gamma) \left(\frac{\beta^2\sigma_v^2 + \sigma_u^2}{\gamma^2\sigma_u^2 + \sigma_v^2 - 2\sigma_v^2\beta\kappa\gamma - \sigma_v^2\kappa^2} \right) \right] x_{it} \equiv \theta x_{it}. \tag{26}$$

The response process y_{it} has a feedback effect on the covariate process x_{it} and consequently x_{it} is a Type III time-dependent covariate.

We simulated 2000 data sets, each of which contained $N = 500$ subjects observed at five time points, with $\beta = 0.5$, $\kappa = 0.3$, $\gamma = 0.5$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 1$. Table 2 shows the bias, the root mean squared error and the efficiency relative to GEE Independence for the same six estimators considered in Setting I. As in Setting I, the GEE Exchangeable and GEE AR-1 estimators are substantially biased and perform poorly. The GMM Type II estimator that uses the invalid moment conditions (15) is also substantially biased and performs poorly in this setting. The GMM Type III and GEE Independence estimators are unbiased and perform similarly. The .05 level test based on (22) of the null hypothesis that x_{it} is a Type II time-dependent covariate was rejected in all 2000 simulations. Consequently, the GMM Moment Selection estimator always equaled the GMM Type III estimator, and performed similarly to GEE Independence. A nominal 95% confidence interval for β_1 based on (19) for the GMM Type III estimator covered the true β_1 in 93.7% of the simulations, showing that the asymptotic confidence interval has reasonably reliable coverage for this setting.

INSERT TABLE 2 ABOUT HERE

The simulations of settings I and II illustrate that 1) when there is a Type II time-dependent covariate, the GMM estimator that uses the additional moment conditions (15) that are only valid for Type II time-dependent covariates can be substantially more efficient than GEE with the independent working correlation (almost twice as efficient in Setting I); 2) when there is a Type III time-dependent covariate, the test based on (22) has power to detect that the covariate is of Type III. Thus, the simulations illustrate that our GMM approach that uses the additional moment conditions (15) only when there is substantive reason to believe that the covariate is of Type II and the test (22) that the covariate is of Type II is not rejected can considerably enhance efficiency when there is a Type II time-dependent covariate while remaining consistent if our hypothesis that the time-dependent covariate is of Type II is wrong. We now study further the power of the test (22) for testing the type of a time-dependent covariate.

16

Setting I continued: We expand the model in Setting I to allow for x_{it} to depend on b_i:

$$
\begin{aligned}
y_{it} &= \gamma_0 + \gamma_1 x_{it} + \gamma_2 x_{i,t-1} + b_i + e_{it} \\
x_{it} &= \rho x_{i,t-1} + \kappa b_i + \epsilon_{it},
\end{aligned}
\tag{27}
$$

where all parameter settings are the same as in Setting I except for κb_i being added to the x_{it} equation. Setting I corresponds to $\kappa = 0$. When $\kappa \neq 0$, the covariate x_{it} is no longer of Type II and is instead of Type III. Figures 1(a) and 1(b) show the power of the level 0.05 test based on (22) for different values of κ (a grid of κ from $[0, 0.2]$ is considered, with each value of κ in the grid 0.001 apart) for sample sizes $N = 500$ and $N = 2000$ respectively. 2000 simulations were done for each combination of κ and sample size. Figure 1 shows that for both sample sizes $N = 500$ and $N = 2000$, the power of the test (22) approaches 1 quickly as κ moves away from zero.

INSERT FIGURE 1 ABOUT HERE

Figure 1 shows that for model (27) with a sample size of at least 500, the test (22) has reasonable finite sample power. With regards to asymptotic power, if the time-dependent covariate is of Type III rather than Type I or II, then, under the conditions in Appendix A, as the number of subjects increases to infinity, the probability that the test (22) rejects converges to 1. Consequently, when the time-dependent covariate is of Type III, the GMM Moment Selection estimator is asymptotically as efficient as the GMM Type III estimator. However, for a given sample size, there will be some region of parameter values for which the time-dependent covariate is of Type III, but the test (22) will have low power for rejecting the null hypothesis that the time-dependent covariate is of Type II (or I), and consequently the GMM Moment Selection estimator will tend to use the Type II moment conditions. To study further the consequences of this for the efficiency of the GMM Moment Selection estimator, we simulated from model (27) with different values of κ. Table 3 shows the root mean squared error for the GMM Type II estimator, the GMM Type III estimator and the GMM Moment Selection estimator for sample sizes $N = 500$ and $N = 2000$ for different values of κ. 2000 simulations were done for each combination of sample size and κ. We see that the GMM Moment Selection estimator performs better than the GMM Type III estimator but not as well as the GMM Type II estimator for κ near 0 ($\kappa < 0.01$ for both sample sizes), performs better than the GMM Type II estimator but not as well as the

GMM Type III estimator for κ close but not very close to zero (κ between 0.01 and 0.06 for $N = 500$ and κ between 0.01 and 0.03 for $N = 2000$) and performs comparably to the GMM Type III estimator for κ further from zero (κ greater than 0.06 for $N = 500$ and κ greater than 0.03 for $N = 2000$).

As indicated in Section 3.4, we recommend using the GMM Moment Selection estimator only when there are substantive reasons to think that a time-dependent covariate is of Type II (or I). Furthermore, it is advisable to conduct simulation studies for the given sample size to determine whether the tradeoff in efficiency between the GMM Moment Selection estimator and the GMM Type III estimator for different parameter values is favorable given one's prior expectations about the parameter values. In some cases, the tradeoff can be made more favorable by choosing a level for the test (22) other than 0.05. A decision-theoretic study of how to choose the level of the test (22) is a valuable topic for future research. For a test that decides among different estimators of the normal linear regression model for cross-sectional data, Sawa and Hiromatsu (1973) and Brook (1976) studied how to set the level of the test optimally according to a minimax regret criterion.

<center>INSERT TABLE 3 ABOUT HERE</center>

<center>5. APPLICATION TO ANTHROPOMETRIC SCREENING</center>

Because of the ease of obtaining anthropometric measurements, anthropometric examination is useful in developing countries for identifying individual children at risk of illness or death and for prioritizing areas to be targeted for government feeding or health programs (Martorell and Habicht, 1986; Zerfas *et al.*, 1986; World Health Organization, 1995). For developing appropriate screening criteria to best allocate scarce resources on the basis of anthropometric examinations, it is important to quantify the association between anthropometric measurements and future morbidity. In developing countries, a record of previous anthropometric examinations for children visiting a health clinic is often unavailable so that it is useful to quantify the association between the anthropometric measurements of a child at a given time point and the outcome of interest (e.g., future morbidity or mortality). Examples of studies that have examined the prospective relationship between anthropometrics at a given time point and morbidity at a future time point include Chen *et al.* (1981), Tonglet *et al.* (1999) and Kossman *et al.* (2000).

<center>18</center>

To estimate the association between anthropometric measurements and morbidity k periods after the anthropometric measurements based on a sample of children, we only need one observation of the anthropometric measurements and one observation of the morbidity outcome k periods after the anthropometric measurements for each child. However, there are several high quality longitudinal data sets containing repeated observations for each child. In particular, the International Food Policy Research Institute (IFPRI) has collected several high quality longitudinal data sets that contain information on child morbidity and anthropometrics (these are publicly available, see www.ifpri.org). It is desirable to make efficient use of these longitudinal data sets. We consider here a data set collected by the IFPRI in the Bukidnon Province in the Philippines and focus on quantifying the association between body mass index (BMI), which equals weight (in kg) divided by height (in cm) squared, and morbidity four months into the future.

The data was collected in 1984-1985 by surveying 448 households living within a 20-mile radius. Data was collected at four time points, separated by four month intervals. For more details on the data, see Bouis and Haddad (1990) and Bhargava (1994). Similar to Bhargava (1994), we focus on the youngest child (1-14 years) in each household and only consider those children who have complete data at all time points, resulting in 370 children with three observations of (BMI at time t, morbidity at time $t+$ 4 months) each. For the morbidity outcome, we follow Bhargava (1994) and use the empirical logistic transformation (Cox, 1970) of the proportion of time over the two weeks prior to the interview that the child was sick,

$$y_{it} = \log \left(\frac{\text{days over last two weeks prior to time } t \text{ child was sick } + 0.5}{14.5 \text{ - days over last two weeks prior to time } t \text{ child was sick}} \right) \quad (28)$$

Besides BMI, we use as predictors gender, age (in months) and dummy variables for the round of the survey (to account for seasonality in morbidity). For illustrative purposes, we focus on a linear model.

Gender is a time-independent covariate. The time-dependent covariates age and survey round dummy are clearly Type I time-dependent covariates in a linear model because (9) holds for these variables (for age, (x_{i1}, x_{i2}, x_{i3}) is a function of x_{it} for each of $t = 1, 2, 3$ and for survey round dummy, (x_{i1}, x_{i2}, x_{i3}) is the same for all children). BMI is plausibly a Type II time-dependent covariate. In order for BMI to be a Type II time-dependent covariate, there cannot be any feedback effect from a child's sickness to the child's BMI. Two reasons

19

that there might be a feedback effect are (a) if a child is sick, the child may not eat much and this could affect the child's weight in the future and (b) infections have generalised effects on nutrient metabolism and utilization (Martorell and Ho, 1984). However, both reasons (a) and (b) for a potential feedback effect are most relevant for diarrheal infections (Martorell and Ho, 1984) and the proportion of children in our study who were sick (over a two week period) who had diarrheal infection was only 9%. Furthermore, because the rounds of the survey were four months apart, the effect of a child's sickness in one round of the survey on the child's weight at the next round is likely to be small. To empirically test the null hypothesis of BMI being a Type II time-dependent covariate, we use the test (22), which gives an insignificant p-value of 0.21.

Table 4 shows the estimated coefficients and standard errors (in parentheses) for the GMM Type II estimator and the GEE Independence estimator. The bottom row of Table 3 shows the average squared prediction error (ASPE) of the two estimators,

$$\text{ASPE} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{y}_{it}^{(-i)} - y_{it} \right)^2 / (NT), \tag{29}$$

where $\hat{y}_{it}^{(-i)}$ is the predicted value of y_{it} based on $\mathbf{x}_{it}$ using the model fitted by leaving out the ith child, i.e., only using the observations from the other $N - 1$ children. The GMM Type II estimator does not show an efficiency gain over GEE Independence for this data set. This is because the correlation among a child's repeated morbidity outcomes is small. The largest correlation between residuals at any two time points using the GEE Independence residuals was 0.21. Our simulation study from Section 4 suggests that for outcomes that are more highly correlated within a child, the GMM approach can provide large efficiency gains.

INSERT TABLE 4 ABOUT HERE

An interesting finding from our analysis is that BMI is not a strong predictor of morbidity. This is consistent with the findings of Chen *et al.* (1981) for a study in Bangladesh and Tonglet *et al.* (1999) for a study in Congo but not with Kossman *et al.* (2000), who found a significant effect of anthropometrics on future morbidity in Sudan. The value of anthropometrics for predicting future morbidity appears to depend considerably on local conditions.

6. DISCUSSION

We have developed a new approach to marginal regression analysis for time-dependent covariates. Our approach classifies time-dependent covariates into three types. The type of covariate determines what estimating equations can be used involving the covariate. We use the generalised method of moments to make optimal use of the estimating equations made available by the covariates. We carried out a simulation study that showed that our approach can provide substantial efficiency gains over GEE with the independent working correlation when the time-dependent covariates are of a certain type while our approach remains consistent even when our assumption about a covariate's type is wrong.

We have focused on marginal regression analysis. Our approach is also useful for obtaining more efficient estimates in the partly conditional model of Pepe and Couper (1997). This model conditions on part of the past history of covariates and outcomes. The partly conditional model is intermediate between the marginal model that conditions only on the covariates at time t and the transition (fully conditional) model that conditions on the full history of covariates and outcomes at time t.

Our focus in using marginal regression analysis for time-dependent covariates is on prediction. For estimating causal relationships involving time-dependent covariates, see Robins *et al.* (1999) and Diggle *et al.* (2002, Section 12.5) for discussion of appropriate methods.

For the simulation studies considered, the asymptotically justified inferences for GMM based on (19) were reasonably reliable. However, when many moment conditions are used relative to the sample size, the asymptotically justified inferences for GMM can be unreliable (Newey and Smith, 2003). For a Type II time-dependent covariate, the number of moment conditions increases quadratically as the number of time periods T increases; specifically there are $T(T+1)/2$ valid moment conditions. For Setting I in Section 4, as T is increased, we found the substantial gains in efficiency of the GMM Type II estimator compared to that of GEE with the independent working correlation estimator persist but the coverage of the nominal 95% confidence intervals for β_1 based on (19) drops. Specifically, for 2000 simulations with a sample size of $N = 500$, the coverage rate of the nominal 95% confidence interval was 93.2% for $T = 5$, 88.1% for $T = 10$, 81.0% for $T = 15$ and 65.5% for $T = 20$. When the sample size was increased to $N = 2000$, the performance of the approximate confidence interval based on (19) improved substantially. In 2000 simulations, the coverage rate of the nominal 95% confidence interval was 95.15% for $T = 5$, 93.60% for $T = 10$, 92.25% for $T = 15$ and 90.10% for $T = 20$.

For our empirical study, the GMM Type II estimator involves 21 moment conditions for six parameters. We studied the coverage of the nominal 95% confidence interval for our empirical study in the following way. We calculated the empirical likelihood estimate (Owen, 2001) of the population distribution under the assumption that the GMM Type II moment conditions hold and that the GMM Type II estimates are the true population parameters. We then simulated data from a population with probabilities equal to the empirical likelihood estimates and checked the coverage rate of the nominal 95% confidence interval based on (19) for the coefficient of BMI. Note that under a population with probabilities equal to the empirical likelihood estimates discussed above, the GMM Type II moment conditions hold and the coefficient on BMI is the GMM Type II estimate of the coefficient on BMI. In 2000 simulations under this setup, the nominal 95% confidence interval based on (19) covered the true coefficient on BMI 86.2% of the time. The above simulations indicate that the nominal coverage of the approximate confidence interval based on (19) can be unreliable when there are many moment conditions relative to the sample size. Similarly, we found that the asymptotic distribution of the test statistic (22) can be unreliable when there are many moment conditions relative to the sample size, with the actual type I error rate exceeding the nominal value when the asymptotic distribution is used to set the critical value of the test.

One approach to dealing with the situation of having many moment conditions relative to the sample size is to use a different estimator than usual GMM. Qu and Lindsay (2003) develop a dimension reduction approach to combining moment conditions that has better properties than usual GMM when there are many moment conditions. Another approach to combining moment conditions is empirical likelihood (Qin and Lawless, 1994; Owen, 2001). GMM and empirical likelihood have the same first-order asymptotic properties but empirical likelihood has better higher-order asymptotic properties when there are many moment conditions (Newey and Smith, 2003). A valuable topic for future research is to compare Qu and Lindsay's approach to GMM as well as empirical likelihood to our use of two-step GMM for marginal regression analysis with time-dependent covariates when there are many moment conditions.

22

APPENDIX A: Consistency of Test for Covariate's Type

We present a set of conditions for the test of (20) vs. (21) based on (22) to be consistent. Let $G_0(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}_0}[\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})]$, $G_0^a(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}_0}[\mathbf{g}_a(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})]$ and $G_0^b(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}_0}[\mathbf{g}_b(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})]$. Also let $G_N^a(\boldsymbol{\beta}) = \frac{1}{N}\mathbf{g}_a(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$ and $G_N^b(\boldsymbol{\beta}) = \frac{1}{N}\mathbf{g}_b(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$, let $(W_N)^{[a]}$ be the submatrix of W_N that corresponds to the moment conditions in $\mathbf{g}_a$ for a weight matrix W_N, let $\|\cdot\|$ denote the Euclidean norm and let $\mathcal{B}$ denote the parameter space of $\boldsymbol{\beta}$. Our conditions for consistency for a $\mathbf{g}_a$, $\mathbf{g}_b$ satisfying the alternative hypothesis (21) are the following:

(C1) (i) $G_0(\boldsymbol{\beta})$ exists and is finite for all $\beta \in \mathcal{B}$; (ii) $G_0^a(\boldsymbol{\beta}) = \mathbf{0}$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C2) There exists a neighborhood $\mathbf{A}$ of β_0 and constants $c_1, c_2 > 0$ such that if $\beta \in \mathbf{A}$, then $\|G_0^b(\boldsymbol{\beta})\| > c_1$ and if $\beta \notin \mathbf{A}$, then $\|G_0^a(\boldsymbol{\beta})\| > c_2$.

(C3) $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \|G_N(\boldsymbol{\beta}) - G_0(\boldsymbol{\beta})\| \xrightarrow{p} 0$.

(C4) $\hat{V}_N^{-1} \xrightarrow{p} W$ for some positive definite matrix W.

Proposition 1. *Let $\mathbf{g}_a, \mathbf{g}_b$ satisfy the alternative hypothesis (21) and conditions (C1)-(C4). The probability that the test based on (22) rejects H_0 for a fixed α level converges to 1 as $N \to \infty$.*

Proof. Let $Q_0(\boldsymbol{\beta}) = G_0(\boldsymbol{\beta})^T W G_0(\boldsymbol{\beta})$ and $Q_0^a(\boldsymbol{\beta}) = G_0^a(\boldsymbol{\beta})^T (W)^{[a]} G_0^a(\boldsymbol{\beta})$. Under conditions (C3) and (C4), we have

$$Q_{a,N}(\hat{\boldsymbol{\beta}}_{GMM,a}) \xrightarrow{p} \inf_{\boldsymbol{\beta} \in \mathcal{B}} Q_0^a(\boldsymbol{\beta}) \tag{30}$$

$$Q_{ab,N}(\hat{\boldsymbol{\beta}}_{GMM,ab}) \xrightarrow{p} \inf_{\boldsymbol{\beta} \in \mathcal{B}} Q_0^{ab}(\boldsymbol{\beta}). \tag{31}$$

Under condition (C1), $\inf_{\boldsymbol{\beta} \in \mathcal{B}} Q_0^a(\boldsymbol{\beta}) = G_0^a(\boldsymbol{\beta}_0)^T (W)^{[a]} G_0^a(\boldsymbol{\beta}_0) = 0$. Under condition (C2) and the positive definiteness of W from (C4), there exists $c_3 > 0$ such that $\inf_{\boldsymbol{\beta} \in \mathcal{B}} Q_0^{ab}(\boldsymbol{\beta}) > c_3$. Combining the results in the above two sentences with (30) and (31), we conclude that for any fixed α,

$$P\{N(Q_{ab,N}(\hat{\boldsymbol{\beta}}_{GMM,ab}) - Q_{a,N}(\hat{\boldsymbol{\beta}}_{GMM,a})) \geq \chi_{r-q}^2(1-\alpha)\} \xrightarrow{p} 1,$$

where $\chi_{r-q}^2(1-\alpha)$ denotes the $1-\alpha$ quantile of the χ_{r-q}^2 distribution. This proves the proposition. □

23

Harris and Mátyás (1999) discuss more primitive conditions that are sufficient for (C3) and (C4) to hold.

REFERENCES

Bhargava, A. (1994). Modelling the health of Filipino children. *Journal of the Royal Statistical Society, Series A* **157**: 417-432.

Bickel, P., Klaasen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: John Hopkins University Press.

Bouis, H.E. and Haddad, L.J. (1990). *Effects of Agricultural Commercialization on Land Tenure, Household Resource Allocation, and Nutrition in the Philippines.* Research Report 79. International Food Policy Research Institute, Washington.

Brook, R.J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association* **71**: 126-131.

Chamberlain, G. (1982). The general equivalence of Granger and Sims causality. *Econometrica* **50**: 569-582.

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

Chen, L.C., Huq, E. and Huffman, S. (1981). A prospective study of the risk of diarrheal diseases according to the nutritional status of children. *American Journal of Epidemiology* **114**: 284-292.

Cox, D.R. (1970). *Analysis of Binary Data.* London: Chapman and Hall.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y. and Zeger, S.L. (2002). *The Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

Eichenbaum, M., Hansen, L.P and Singelton, K.S. (1988). A time series analysis of representative agent models of consumption and leisure under uncertainty. *Quarterly Journal of Economics* **103**: 51-78.

Emond, M.J., Ritz, J. and Oakes, D. (1997). Bias in GEE estimates from misspecified models for longitudinal data. *Communications in Statistics* **26**: 15-32.

Engle, R.F., Hendry, D.F. and Richard, J.-F. (1983). Exogeneity. *Econometrica* **51**: 277-304.

Ferguson, T.S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Annals of Mathematical Statistics* **29**: 1046-1062.

Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics* **51**: 309-317.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**: 424-438.

Hall, A.R. (1999). Hypothesis testing in models estimated by GMM. In *Generalized Method of Moments Estimation*, ed. L. Mátyás. Cambridge University Press: New York.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**: 1029-1054.

Harris, D. and Mátyás, L. (1999). Introduction to the generalized method of moments estimation. In *Generalized Method of Moments Estimation*, ed. L. Mátyás. New York: Cambridge University Press.

Hu, F.-C. (1993). A statistical methodology for analyzing the causal health effect of a time-dependent exposure from longitudinal data. Unpublished Sc.D. dissertation, Harvard School of Public Health, Department of Biostatistics, Boston, MA.

Kossmann, J., Nestel, P., Herrera, M.G., El Amin, A. and Fawzi, W.W. (2000). Undernutrition in relation to childhood infections: a prospective study in the Sudan. *European Journal of Clinical Nutrition* **54**: 463-472.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13-22.

Lindsay, B.G. and Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science* **18**: 394-410.

Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley and Sons.

Martorell, R. and Habicht, J.-P. (1986). Growth in early childhood in developing countries. In *Human Growth: A Comprehensive Treatise*, eds. F. Falkner and J.M. Tanner. Plenum Press: New York.

Martorell, R. and Ho, T.J. (1984). Malnutrition, morbidity and mortality. *Population and Development Review* **10**: 49-68.

Mátyás, L., ed. (1999). *Generalized Method of Moments Estimation*. New York: Cambridge University Press.

McCullaugh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Newey, W.K. (1985). Generalized method of moments specification testing. *Journal of Econometrics* **29**: 229-256.

Newey, W.J. and Smith, R.J. (2003). Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica*. Forthcoming.

Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proc. Berkeley Symp. Math. Statist. Probab.*. University of California Press: 239-273.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman& Hall, Boca Raton.

Pan, W. and Connett, J.E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica* **12**: 475-490.

Pan, W., Louis, T.A., and Connett, J.E. (2000). A note on marginal linear regression with correlated response data. *American Statistician* **54**: 191-195.

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

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

Qu, A. and Lindsay, B.G. (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. *Journal of the Royal Statistical Society, Series B* **65**: 127-142.

Qu, A., Lindsay, B.G. and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions. *Biometrika* **87**: 823-836.

Robins, J.M., Greenland, S. and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repated binary outcome (with Discussion). *Journal of the American Statistical Association* **94**: 687-712.

Sawa, T. and Hiromatsu, T. (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica* **41**: 1093-1101.

Tonglet, R., Lembo, E.M., Zihindula, P.M., Wodon, A., Dramaix, M. and Hennart, P. (1999). How useful are anthropometric, clinical and dietary measurements of nutritional status as predictors of morbidity of young children in central Africa. *Tropical Medicine and International Health* **4**: 120-130.

World Health Organization (1995). Physical status: the use and interpretation of anthropometry. WHO technical report series 854. World Health Organization: Geneva.

Zeger, S.L. and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series. *Statistica Sinica* **1**: 51-64.

Zerfas, A., Jelliffe, D.B and Jelliffe, E.F.P (1986). Epidemiology and Nutrition. In *Human Growth: A Comprehensive Treatise*, eds. F. Falkner and J.M. Tanner. Plenum Press: New York.

Estimator	Bias	RMSE	Efficiency
GMM Type II	0.00	0.0407	1.99
GMM Type III	0.00	0.0574	1.00
GMM Moment Selection	0.00	0.0430	1.78
GEE Independence	0.00	0.0574	1.00
GEE Exchangeable	-0.28	0.2861	0.04
GEE AR-1	-0.61	0.6087	0.01

Table 1: Results of simulation study for Setting I, which has a Type II time-dependent covariate. For the parameter β_1 in (23), the table shows the bias, root mean squared error (RMSE) and ratio of mean squared error of GEE Independence to mean squared error of estimator (Efficiency) for 2000 simulations

| Estimator	Bias	RMSE	Efficiency
GMM Type II	-0.07	0.0739	0.07
GMM Type III	0.00	0.0196	0.99
GMM Moment Selection	0.00	0.0196	0.99
GEE Independence	0.00	0.0195	1.00
GEE Exchangeable	-0.08	0.0870	0.05
GEE AR-1	-0.15	0.1475	0.02

Table 2: Results of simulation study for Setting II, which has a Type III time-dependent covariate. For the parameter θ in (26), the table shows the bias, root mean squared error (RMSE) and ratio of mean squared error of GEE Independence to mean squared error of estimator (Efficiency) for 2000 simulations.

|	$N = 500$			$N = 2000$		
κ	GMM Type II	GMM Type III	GMM Moment Selection	GMM Type II	GMM Type III	GMM Moment Selection
0	.041	.057	.043	.021	.029	.023
.005	.046	.057	.048	.027	.029	.028
.01	.057	.058	.059	.043	.030	.042
.02	.088	.057	.084	.079	.027	.048
.03	.124	.060	.102	.118	.029	.031
.04	.159	.059	.097	.156	.030	.030
.05	.196	.081	.059	.193	.030	.030
.06	.231	.060	.064	.227	.030	.030
.07	.262	.059	.061	.260	.030	.030
.10	.335	.060	.060	.336	.030	.030

Table 3: Root mean squared error of GMM Type II, GMM Type III and GMM Moment Selection estimators for 2000 simulations of model (27) for various value of κ and N.

Variable	GMM Type II	GEE Independence
BMI	-0.049	-0.062
	(0.042)	(0.037)
Age	-0.010	-0.012
	(0.003)	(0.003)
Gender	-0.091	0.145
	(0.161)	(0.098)
Survey Round 2	-0.280	-0.280
Dummy	(0.107)	(0.120)
Survey Round 3	0.004	0.024
Dummy	(0.119)	(0.121)
Avs. Sq. Pred. Err.	2.68	2.67

Table 4: Estimated coefficients and standard errors (in parentheses) for the GMM Type II and GEE Independence estimators. The response variable is the child morbidity index (28). The bottom row is the average squared prediction error (29).
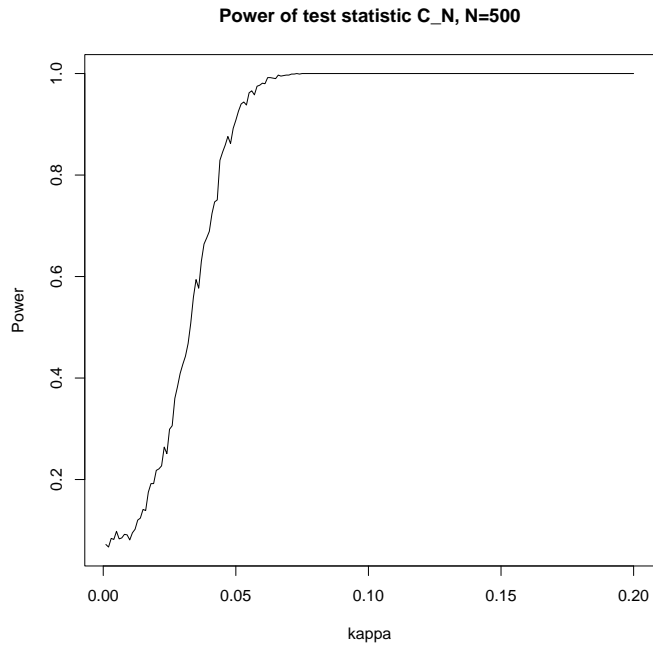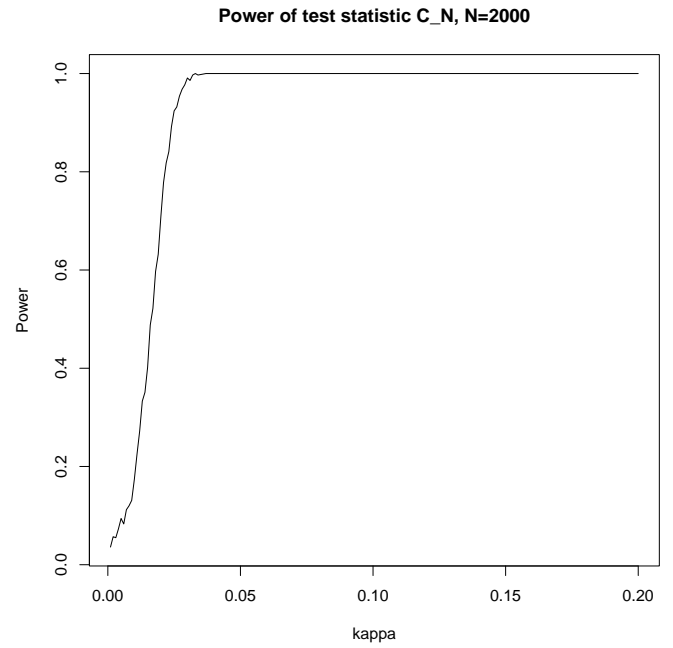
Figure 1: Plots of the power of the level 0.05 test based on (22) for different values of κ for model (27). Figure 1(a) shows the results for a sample size of $N = 500$ and Figure 1(b) shows the results for a sample size of $N = 2000$. 2000 simulations were performed for each value of κ.