

Adjusting a subject-specific time of event in longitudinal studies

Hyunkeun Ryan Cho,¹ Seonjin Kim² and Myung Hee Lee³

Statistical Methods in Medical Research
2020, Vol. 29(7) 1787–1798

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0962280219876957

journals.sagepub.com/home/smm



Abstract

Biomedical studies often involve an event that occurs to individuals at different times and has a significant influence on individual trajectories of response variables over time. We propose a statistical model to capture the mean trajectory alteration caused by not only the occurrence of the event but also the subject-specific time of the event. The proposed model provides a post-event mean trajectory smoothly connected with the pre-event mean trajectory by allowing the model parameters associated with the post-event mean trajectory to vary over time of the event. A goodness-of-fit test is considered to investigate how well the proposed model is fit to the data. Hypothesis tests are also developed to assess the influence of the subject-specific time of event on the mean trajectory. Theoretical and simulation studies confirm that the proposed tests choose the correctly specified model consistently and examine the effect of the subject-specific time of event successfully. The proposed model and tests are also illustrated by the analysis of two real-life data from a biomarker study for HIV patients along with their own time of treatment initiation and a body fatness study in girls with different age of menarche.

Keywords

Longitudinal trajectory, piecewise regression model, quadratic inference function, spline approximation, varying coefficient

1 Introduction

In longitudinal studies, subjects are repeatedly measured in an effort to understand the mean trend of response variables over time. The trend can often be influenced by a significant event. For example, the MIT Growth and Development Study^{1,2} was conducted to explore the mean trend of body fat in girls over an adolescent period. One of the key components in the prospective study is that menarche generally has a strong influence on changes in body fat accretion to such an extent that menarche is often regarded as a critical event in the development of obesity. Since age of menarche differs among individuals, it is very essential to accommodate not only the occurrence of the event but also a subject-specific age of the event when the mean trend of body fat over time is modeled.

Another example is a clinical trial of antiretroviral therapy (ART) in individuals infected with human immunodeficiency virus (HIV) conducted in Haiti.³ In the HIV study, each patient had a personalized ART timeline determined at the physician's discretion. As a result, time of ART initiation varied ranging from a few weeks to several years from the patient's enrollment. Preliminary data analysis indicates that ART is likely to be effective at reducing an inflammation biomarker, yet the rate of decrease is more likely to vary with time of ART initiation. This suggests that the elapsed time from enrollment to initiation of ART is crucial in understanding the inflammation biomarker mean trajectory over time.

The primary goal of this article is two-fold: (1) to provide interpretable models for fitting the dynamic mean trend of a response variable over time and (2) to assess how the subject-specific time of event has an influence on

¹Department of Biostatistics, University of Iowa, Iowa City, IA, USA

²Department of Statistics, Miami University, Oxford, OH, USA

³Center for Global Health, Department of Medicine, Weill Cornell Medicine, New York, NY, USA

Corresponding author:

Seonjin Kim, Miami University, 311 Upham Hall, Oxford, OH 45056, USA.

Email: kims20@miamioh.edu

changes in the longitudinal mean trend. If time of event remains the same across individuals, we can adopt the piecewise polynomial model

$$Y_i(t_{ij}) = \begin{cases} \alpha_0 + \alpha_1 t_{ij} + \cdots + \alpha_p t_{ij}^p + \epsilon_{ij} & t_{ij} \leq c \\ \beta_0 + \beta_1 t_{ij} + \cdots + \beta_p t_{ij}^p + \epsilon_{ij} & t_{ij} > c \end{cases} \quad (1)$$

for $i = 1, \dots, n$ and $j = 1, \dots, n_i$, where n_1, \dots, n_n are the number of measurements from the n subjects, $Y_i(t_{ij})$ is the response variable measured at time t_{ij} for subject i at the j th visit, α_k and β_k , $k = 0, \dots, p$, are unknown constant coefficients, ϵ_{ij} is a random error, and c is the common time of event for all individuals. In practice, t_{ij} is generally defined by the study objective, such as an age in the analysis of the MIT study or visit time relative to enrollment in the HIV study above.

By imposing the restrictions on model (1) that $E\{Y_i(t_{ij})\}$ and its first $p - 1$ derivatives are continuous in time, Gallant and Fuller⁴ connect two segments in equation (1) smoothly and develop the following model

$$Y_i(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \cdots + \alpha_p t_{ij}^p + \gamma(t_{ij} - c)_+^p + \epsilon_{ij} \quad (2)$$

where γ is a unknown coefficient and $(t_{ij} - c)_+^p = (t_{ij} - c)^p I(t_{ij} \geq c)$ is a p -degree truncated polynomial term with a fixed knot at the common time of event c . Although the trajectory change due to the event can be reflected on one last term in model (2), it is not applicable for the aforementioned studies in which time of event differs between individuals.

In order to accommodate the subject-specific time of event, we propose to formulate a time-of-event-dependent regression model with a p -degree truncated polynomial term with a varying knot at an individual time of event as

$$Y_i(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \cdots + \alpha_p t_{ij}^p + \gamma(c_i)(t_{ij} - c_i)_+^p + \epsilon_{ij} \quad (3)$$

where $\gamma(c_i)$ is a smooth function of c_i and c_i is the i th individual time of event, such as age of menarche in the MIT study or time of ART initiation in the HIV study for subject i . The proposed model allows time of the event to vary across subjects. In addition, this modeling is intuitively appealing to researchers in that the varying knot resets the origin of time of the event and provides an interpretable mean trajectory shift since the event occurs. In other words, the first $(p + 1)$ terms are served as one common pre-event longitudinal mean trajectory and the post-event mean trajectory change is reflected on the truncated polynomial term by incorporating the subject-specific time of event. Therefore, the effect of the time of the event on the mean trajectory alteration is illustrated by evaluating the varying coefficient $\gamma(c_i)$ in model (3). The model can also be separated into two segments as

$$Y_i(t_{ij}) = \begin{cases} \alpha_0 + \alpha_1 t_{ij} + \cdots + \alpha_p t_{ij}^p + \epsilon_{ij} & t_{ij} \leq c_i \\ \beta_0(c_i) + \beta_1(c_i)t_{ij} + \cdots + \beta_p(c_i)t_{ij}^p + \epsilon_{ij} & t_{ij} > c_i \end{cases} \quad (4)$$

where $\beta_k(c_i) = \alpha_k + \binom{p}{k}(-c_i)^{(p-k)}\gamma(c_i)$ for $k = 0, \dots, p$. Contrary to model (1), the longitudinal mean trajectory of outcomes after the event depends on time of the event. Therefore, the proposed model enables us to explore the dynamic mean change of the response variable by taking into account pre- and post-events simultaneously while accommodating the subject-specific time of event.

In order to fit the proposed model to data, we approximate the varying coefficient by a spline basis function expansion and employ quadratic inference functions (QIFs).⁵ QIF not only improves estimation efficiency of regression parameters in model (3) by incorporating the within-subject correlation but also provides an inference function for model diagnostic tests and goodness-of-fit tests. In model (3), the choice of a proper polynomial degree plays an important role in determining the overall mean trend of outcomes over time. The goodness-of-fit test based on QIF is able to select a proper polynomial degree of p consistently and consequently the correctly specified model. After an optimal value of p is chosen, a hypothesis test is further proposed to evaluate whether the longitudinal mean trajectory is influenced by the subject-specific time of event or not.

The notion of event is used to refer a presence of occurrence influencing longitudinal trend of a response variable and should not be confused with the notion indicating study outcome in survival analysis. Duration of time till event in our paper is not of interest rather understood as covariates in the longitudinal modeling. The remainder of the paper proceeds as follows: Section 2 provides estimation and inference about the parameters

in model (3). In Section 3, the finite sample performance of the proposed procedure is evaluated in two scenarios, where $\gamma(c_i)$ is varying or constant over a value of c_i . Simulation studies suggest that the proposed test successfully identifies whether or not time of the event affects the overall mean trajectory of response variables. In Section 4, we fit the proposed model to real data sets from the two aforementioned studies and conclude that the time of the event has an influence on changes in the longitudinal mean trajectory only in the clinical trial of HIV-infected patients study not in the MIT Growth and Development study. We conclude with remarks in Section 5 and place regularity conditions and theoretical proofs in Appendix 1.

2 Methodologies

2.1 Estimation of regression parameters

One of the key features of model (3) is that the unspecified nonparametric coefficient $\gamma(c_i)$ adds flexibility in modeling dynamic changes of the mean trend of response variables after a significant event occurs. To fit model (3) to data, we approximate the nonparametric coefficient by a basis function expansion as $\gamma(c_i) \approx \sum_{j=0}^h \gamma_j B_j(c_i)$, where γ_j 's are unknown regression coefficients and $\{B_j(\cdot), j = 0, \dots, h\}$ is a set of basis functions. Although the type of basis functions is not restricted in our procedure, we mainly focus our attention on the q -degree truncated power spline basis function approximation with a set of u knots $\{k_m, m = 1, \dots, u\}$, i.e., $\gamma(c_i)$ is modeled as

$$\gamma(c_i) \approx \gamma_0 + \gamma_1 c_i + \dots + \gamma_q c_i^q + \sum_{m=1}^u \gamma_{q+m} (c_i - k_m)_+^q \tag{5}$$

Note that the truncated power spline basis offers practical convenience in conducting statistical inference about the effect of time of the event, since the polynomial function is nested within equation (5). With basis function approximation, model (3) can be rewritten as

$$Y_i(t_{ij}) \approx \alpha_0 + \alpha_1 t_{ij} + \dots + \alpha_p t_{ij}^p + \left\{ \gamma_0 + \gamma_1 c_i + \dots + \gamma_q c_i^q + \sum_{m=1}^u \gamma_{q+m} (c_i - k_m)_+^q \right\} (t_{ij} - c_i)_+^p + \epsilon_{ij} = X_i(t_{ij})^\top \theta + \epsilon_{ij}$$

where $\theta = (\alpha_0, \dots, \alpha_p, \gamma_0, \dots, \gamma_{q+u})^\top$, $X_i(t_{ij}) = (1, t_{ij}, \dots, t_{ij}^p, (t_{ij} - c_i)_+^p, \dots, c_i^q (t_{ij} - c_i)_+^p, (c_i - k_1)_+^q (t_{ij} - c_i)_+^p, \dots, (c_i - k_u)_+^q (t_{ij} - c_i)_+^p)^\top$, and ϵ_{ij} satisfies $E(\epsilon_{ij}) = 0$.

For the estimation of θ , generalized estimating equations (GEE)⁶ can be considered under the marginal regression framework as

$$\sum_{i=1}^n X_i A_i^{-1/2} R_i(\rho)^{-1} A_i^{-1/2} (Y_i - X_i^\top \theta) = 0 \tag{6}$$

where $X_i = (X_i(t_{i1}), \dots, X_i(t_{im_i}))$, $Y_i = (Y_i(t_{i1}), \dots, Y_i(t_{im_i}))^\top$, A_i is a diagonal variance matrix of Y_i , and $R_i(\rho)$ is a working correlation matrix with a nuisance parameter vector of ρ . GEE can yield a consistent estimator of θ by solving equation (6) and improve estimation efficiency by incorporating the correlation among n_i measurements. However, an additional estimation of ρ in $R_i(\rho)$ is required to obtain the efficient estimator of θ . Moreover, GEE does not provide a proper inference function for a goodness-of-fit test, which is essential to select the optimal degree of a polynomial in model (3).

Alternatively, we adopt QIF.⁵ QIF models the inverse of $R_i(\rho)^{-1}$ in equation (6) as $R_i(\rho)^{-1} = \sum_{k=1}^d b_k M_{ik}$, where M_{i1}, \dots, M_{id} are known basis matrices representing a working correlation matrix of Y_i and b_1, \dots, b_d are unknown constant coefficients. The choice of basis matrices depends on the type of $R_i(\rho)$. For instance, if the working correlation matrix is assumed to be a compound symmetry structure, its inverse can be represented with two basis matrices, M_{i1} and M_{i2} , where M_{i1} is an identity matrix and M_{i2} is a symmetric matrix with 0 on the diagonal and 1 elsewhere. If the working correlation matrix corresponds to an AR(1) structure, two basis matrices, M_{i1} and M_{i2} , are needed to approximate the inverse matrix; M_{i1} is an identity matrix and M_{i2} is a symmetric matrix with 1 on the subdiagonal and 0 elsewhere. More details can be found in Qu et al.⁵ and Qu and Lindsay.⁷

By substituting M_{i1}, \dots, M_{id} for $R_i(\rho)^{-1}$ in equation (6), the estimator of θ is obtained by minimizing QIF

$$Q_{p,\ell}(\theta) = ng(\theta)^\top V(\theta)^{-1} g(\theta) \tag{7}$$

where $\ell = 1 + q + u$ is the number of basis functions, $g(\theta) = \sum_{i=1}^n g_i(\theta)/n$, and $V(\theta) = \sum_{i=1}^n g_i(\theta) g_i(\theta)^\top / n$ with

$$g_i(\theta) = \begin{pmatrix} X_i A_i^{-1/2} M_{i1} A_i^{-1/2} (Y_i - X_i^\top \theta) \\ \vdots \\ X_i A_i^{-1/2} M_{id} A_i^{-1/2} (Y_i - X_i^\top \theta) \end{pmatrix} \quad (8)$$

QIF can yield a consistent and more efficient estimator than the one assuming the working independence without estimating additional nuisance parameters associated with the working correlation structure.⁵ In addition, the resultant estimator is most efficient among estimators obtained from the same set of estimating equations in equation (8), since QIF optimally combines the estimating equations. We illustrate the asymptotic distribution of the proposed estimator in the case of fixed-knot asymptotics, where the number of knots is assumed to be fixed as the number of subjects goes to infinity. This is a very useful and practical condition^{8,9} in developing statistical inference about model (3).

Theorem 1. *Under regularity conditions 1–3 in Appendix 1, there exists a minimizer of $Q_{p,\ell}(\theta)$, denoted by $\hat{\theta}$, such that $\sqrt{n}(\hat{\theta} - \theta_0)$ follows an asymptotic normal distribution with mean 0 and variance covariance matrix $(\Phi^\top \Sigma^{-1} \Phi)^{-1}$, where $\Phi = E\{\partial g_i(\theta_0)/\partial \theta\}$, $\Sigma = E\{g_i(\theta_0) g_i(\theta_0)^\top\}$, and θ_0 is a true value of θ .*

The root n consistency and asymptotic normality of the resultant estimator still hold even when the assumed working correlation structure is misspecified under the regularity conditions. We remark that when regression models with multiple varying coefficients are fitted, representing all varying coefficients by a number of basis functions could be problematic due to overfitting the data. This could result in degrading the efficiency of the parameter estimates and poor performance of statistical inferences; see Ruppert¹⁰ and Tian et al.¹¹ An alternative is to employ the penalized approaches based on QIF^{9,11}. However, we are not concerned with overfitting in our study, since basis function approximation is used only for one varying coefficient, as shown in model (3).

2.2 Goodness-of-fit test and hypothesis

The choice of an optimal polynomial degree in model (3) is essential, since two segments of the longitudinal mean trajectory are modeled as polynomials in time and the pattern of the segments depends on the polynomial degree. Since QIF plays a similar role to the loglikelihood function, $Q_{p,\ell}(\theta)$ in equation (7) is an effective tool in measuring how well model (3) is fit to the data. We denote a chi-square distribution with r degrees of freedom as χ_r^2 . Recall that d , p , and ℓ denote the number of basis matrices for the working correlation matrix, the polynomial degree of the proposed model, and the number of basis functions for the varying coefficient $\gamma(c_i)$, respectively.

Theorem 2. *If regularity conditions 1–3 hold and model (3) is correctly specified, the asymptotic distribution of $Q_{p,\ell}(\hat{\theta})$ is $\chi_{(d-1)(p+\ell)}^2$.*

Theorem 2 ensures that $Q_{p,\ell}(\hat{\theta})$ can be regarded as a goodness-of-fit test statistic that indicates whether the proposed model is fit to the data sufficiently well. Under the assumption that a predetermined sufficiently large number of basis functions, say ℓ^* , guarantees a consistent estimate of $\gamma(c_i)$, a proper polynomial degree, say p^* , is obtained based on the criterion $Q_{p^*,\ell^*}(\hat{\theta}) < \kappa_\alpha(\chi_{(d-1)(p^*+\ell^*)}^2)$ at a significance level of α , where $\kappa_\alpha(\chi_{(d-1)(p^*+\ell^*)}^2)$ is the $(1 - \alpha)$ th quantile of $\chi_{(d-1)(p^*+\ell^*)}^2$. We remark that model (3) with order p^* is not nested within model (3) with a higher order polynomial due to the restrictions that the expected value of the response variable in model (3) and its first $p - 1$ derivatives are continuous in time. Our simulation studies also confirm that when model (3) with order p^* is true, the goodness-of-fit test always rejects all models except the one with the true order of p^* .

After the proper polynomial degree is chosen, it is of particular interest to evaluate whether a subject-specific time of event has an influence on longitudinal mean trajectory alteration or not. Since the truncated polynomial term reflects changes in the mean trend due to the event, $\gamma(c_i)$ illustrates the effect of time of event on longitudinal mean trajectory alteration. That is, this is equivalent to testing whether $\gamma(c_i)$ is constant over c_i or not

$$H_0 : \gamma(c_i) = \gamma_0 \quad \text{versus} \quad H_1 : \gamma(c_i) \neq \gamma_0 \quad (9)$$

We note that $\theta = (\alpha_0, \alpha_1, \dots, \alpha_p, \gamma_0, \gamma_1, \dots, \gamma_{q+u})^\top = (\theta_s^\top, \theta_\gamma^\top)^\top$ and $\theta_s = (\alpha_0, \alpha_1, \dots, \alpha_p, \gamma_0)^\top$. Since θ_γ is a zero vector under H_0 , an appropriate test statistic for testing H_0 against H_1 is

$$T = Q_{p^*, \ell^*}(\tilde{\theta}) - Q_{p^*, \ell^*}(\hat{\theta})$$

where $\tilde{\theta} = (\tilde{\theta}_s^\top, 0, \dots, 0)^\top$ and $\tilde{\theta}_s$ is a minimizer of $Q_{p^*, 1}(\theta_s) = n g_s(\theta_s)^\top V_s(\theta_s)^{-1} g_s(\theta_s)$ having $g_s(\theta_s) = \sum_{i=1}^n g_{is}(\theta_s)/n$, $V_s(\theta_s) = \sum_{i=1}^n g_{is}(\theta_s) g_{is}(\theta_s)^\top / n$, and $g_{is}(\theta_s)$ is a subset of estimating equations in $g_i(\theta)$ associated with θ_s only.

Theorem 3. *Under regularity conditions 1–3 and the null hypothesis in equation (9), the asymptotic distribution of the proposed test statistic T is $\chi_{\ell^*-1}^2$.*

Theorem 3 ensures that $\gamma(c_i)$ is varying with c_i if the test statistic is greater than the $(1 - \alpha)$ th quantile of $\chi_{\ell^*-1}^2$ at the nominal level of α . If the null hypothesis is not rejected, there is no substantial evidence that changes in the longitudinal mean trajectory are affected by the subject-specific event time. We remark that the hypothesis statements and corresponding test statistic can be readily modified upon the aim of the scientific interest. For instance, if the null hypothesis is not rejected, it is of particular interest to check if the pattern of the longitudinal mean trajectory over a range of time is altered after the event occurs. This can be readily examined by testing $H_0 : \gamma(c_i) = 0$ against $H_1 : \gamma(c_i) \neq 0$.

We remark that the aforementioned inferences are all conducted under the proposed model with a sufficiently large number of basis functions ℓ^* . Several references^{9–11} suggested that the upper limit of the degree of the truncated polynomial and the number of knots could be set as 3 and 10, respectively, where the knots are evenly distributed within the range of time in practice. Our extensive numerical studies have also suggested that the 3-degree truncated power spline basis with 10 equally distributed knots fits the true varying coefficient sufficiently well. Moreover, the proposed model is still effective even when $\gamma(c_i)$ is overfitted, i.e., the longitudinal mean trajectories are comparable in cases where the number of knots are 10 or above. The results suggest the penalized based approach would be unnecessary, although the number of basis functions can be reduced using QIF-based Bayesian information criterion proposed by Wang and Qu.¹²

3 Simulation studies

In this section, empirical studies are conducted to numerically examine the efficiency of the proposed parameter estimation method and the effectiveness of the hypothesis tests suggested in Section 2.2. Correlated continuous responses are generated as

$$Y_i(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 t_{ij}^2 + \gamma(c_i)(t_{ij} - c_i)_+^2 + \epsilon_{ij} \tag{10}$$

where $\alpha_0 = \alpha_1 = \alpha_2 = 1$, $t_{ij} = j + \text{Unif}(-1, 0)$ for $i \in \{1, \dots, 200\}$ and $j \in \{1, \dots, 6\}$, an individual value of c_i is generated independently from $\text{Unif}(1, 5)$, and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i6})^\top$ is generated independently from a multivariate normal distribution $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i6}) \sim N(0, \Sigma)$ with Σ being an AR(1) correlation structure with a correlation coefficient of 0.8. To generate unbalance data, we drop Y_{ij} from the simulated data set when $M_{ij} = 0$, where M_{ij} is an indicator randomly generated from $\text{Pr}(M_{ij} = 1) = 0.8$. This leads to a different number of observations measured from each subject at unequally spaced time points between 0 and 6. From 1000 simulated data sets, the performance of our proposed procedure is demonstrated in two scenarios, where $\gamma(c_i)$ is varying with c as $\gamma(c_i) = c_i + \cos(\pi c_i)$ or constant as $\gamma(c_i) = 1$.

3.1 Scenario I: $\gamma(\mathbf{c}_i) = \mathbf{c}_i + \cos(\pi \mathbf{c}_i)$

We first conduct a goodness-of-fit test to find the optimal degree of the polynomial using the proposed approach with the spline basis of $\gamma(c_i)$ having 10 equally distributed knots and 3-degree polynomial basis functions under the AR(1) and compound symmetry working correlation structures. The rejection rates of choosing the quadratic polynomial at a significance level of 0.05 are 0.035 and 0.031, respectively, while both linear and cubic regression models are always rejected in the 1000 simulation runs. Figure 1 provides quantile-quantile plots for testing whether or not model (3) with $p = 2$ is fit to each simulated data set sufficiently well. The top two plots confirm the effectiveness of the goodness-of-fit test regardless of the correlation structure. We further test $H_0 : \gamma(c_i) = \gamma_0$ and note that the proposed test rejects the null hypothesis every time.

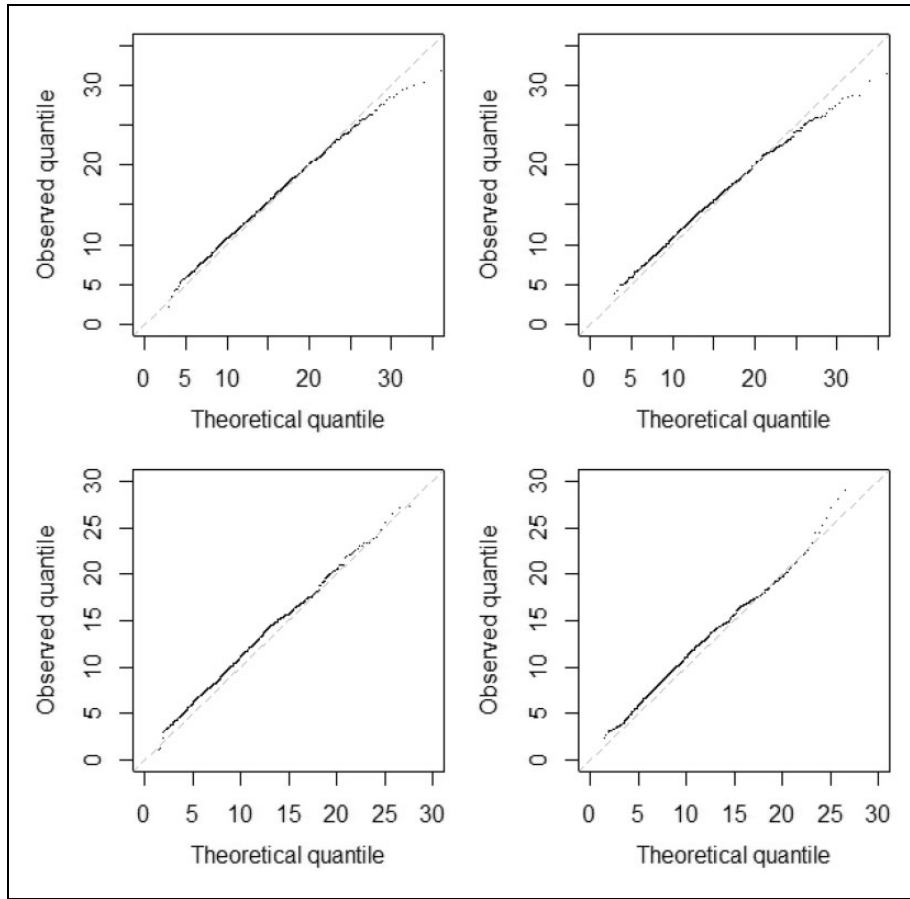


Figure 1. Top: Quantile-quantile plots for goodness-of-fit tests comparing the chi-square distribution with 17 degrees of freedom and quadratic inference function under AR(1) (left) and compound symmetry (right) in Scenario 1. Bottom: Quantile-quantile plots for hypothesis tests comparing the chi-square distribution with 13 degrees of freedom and the proposed test statistic under AR(1) (left) and compound symmetry (right) in Scenario 2.

In model (10), the estimator of α_k , $k=0, 1, 2$, and their standard error are evaluated under the AR(1), compound symmetry, and independent working correlation structures. Table 1 reports mean squared errors, coverage probabilities, and average lengths of 95% confidence intervals. Note that the confidence intervals are formulated based on the asymptotic result in Theorem 1 using the estimated limiting covariance matrix of θ

$$\left[\left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial g_i(\hat{\theta})}{\partial \theta} \right\}^\top \left\{ \frac{1}{n} \sum_{i=1}^n g_i(\hat{\theta}) g_i(\hat{\theta})^\top \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial g_i(\hat{\theta})}{\partial \theta} \right\} \right]^{-1}$$

The results show that the mean squared errors under the AR(1) structure are smaller than the ones ignoring the within-subject correlation. Even if the working correlation is misspecified with a compound symmetry structure, the proposed method still yields more efficient estimators compared to the one under the independent correlation structure. With regard to statistical inference about the regression parameter, coverage probabilities are between 0.93 and 0.95 when correlation information is accommodated, while those are above 0.96 under the independent correlation structure. This can be explained by the fact that the confidence intervals, assuming the working independence, are wider in all cases under consideration. These results suggest that the proposed procedure achieves estimation efficiency and effective inference by accommodating the within-subject correlation.

To evaluate how the proposed method estimates $\gamma(c_i)$, the mean integrated squared error for $\gamma(c_i)$ is defined as $MISE\{\hat{\gamma}(c)\} = \sum_{k=1}^{39} \{\hat{\gamma}(c_k) - \gamma(c_k)\}^2/39$, where $\hat{\gamma}(c_k)$ are estimates of $\gamma(c_k)$ from c_1, \dots, c_{39} and are evenly space time points on (1, 5). Figure 2 provides fitted varying coefficient curves corresponding to nine deciles of the mean

Table 1. Mean squared errors (MSE), coverage probabilities, and average lengths of 95% confidence interval under the AR(1), compound symmetry (CS), and independent (IN) working correlation structures in Scenario 1, $\gamma(c_i) = c_i + \cos(\pi c_i)$, and Scenario 2, $\gamma(c_i) = 1$, respectively.

Scenario		MSE \times 100			Coverage probability			Average length		
		α_0	α_1	α_2	α_0	α_1	α_2	α_0	α_1	α_2
1	AR(1)	0.815	0.477	0.024	0.932	0.931	0.940	0.164	0.120	0.026
	Varying	0.949	0.645	0.034	0.932	0.943	0.940	0.171	0.132	0.030
	$\gamma(c_i)$	0.987	0.730	0.039	0.969	0.981	0.965	0.211	0.197	0.042
2	AR(1)	0.779	0.414	0.021	0.935	0.937	0.940	0.163	0.119	0.026
	Varying	0.897	0.500	0.025	0.937	0.939	0.942	0.170	0.134	0.030
	$\gamma(c_i)$	0.970	0.696	0.037	0.965	0.979	0.963	0.203	0.190	0.040
Constant	AR(1)	0.741	0.289	0.010	0.946	0.953	0.948	0.164	0.100	0.018
	γ_0	0.815	0.317	0.011	0.951	0.949	0.950	0.170	0.105	0.019
	IN	0.901	0.408	0.015	0.956	0.984	0.971	0.190	0.150	0.026

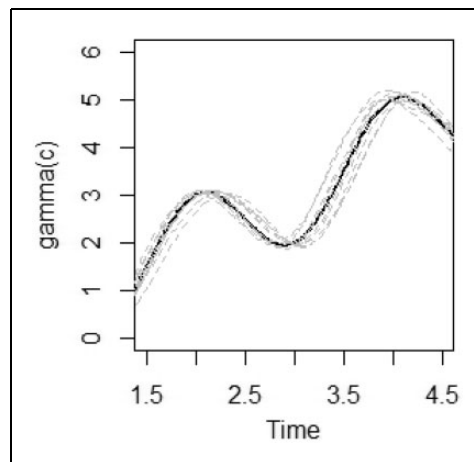


Figure 2. A true value of $\gamma(c_i)$ (black solid curve) and fitted varying coefficients (gray dashed curve) corresponding to nine deciles of mean integrated squared errors from 1000 simulations in Scenario 1.

integrated squared errors from 1000 simulations. This figure shows that the fitted curves successfully capture the true pattern of $\gamma(c_i)$.

3.2 Scenario 2: $\gamma(c_i) = 1$

In line with the aforementioned process in Section 3.1, the goodness-of-fit test is conducted to select the true model. The type-I errors are close to a level of 0.05: 0.041, and 0.036 under the AR(1) and compound symmetry in model (10), respectively. Quantile-quantile plots are also drawn and similar to the top two plots in Figure 1 and, thus, are omitted here. Next, we conduct a hypothesis test to determine whether $\gamma(c_i) = \gamma_0$ or not. The rejection rates at a nominal level of 0.05 are 0.05 and 0.046 under both AR(1) and compound symmetry structures, respectively. The bottom plots in Figure 1 also show that the empirical quantiles of the test statistic follow the theoretical chi-square quantile sufficiently well. We further test $H_0 : \gamma(c_i) = 0$ to check if a one-piece function of time is enough to fit data and the hypothesis test always reject H_0 .

Following the above test results, we set $\gamma(c_i)$ to be a nonzero constant value of γ_0 and evaluate the estimator of the regression parameters and their standard error in model (10) under the AR(1), compound symmetry, and independent structures. For the purpose of comparison, we also let $\gamma(c_i)$ vary with 10 knots and 3-degree polynomial basis functions and fit model (10) to the data. As shown in the second- and third-row blocks in Table 1, the results under the varying coefficient of $\gamma(c_i)$ are comparable with those reported in Scenario 1. However, the proposed procedure in the model with a constant value of $\gamma(c_i)$ yields more efficient estimators

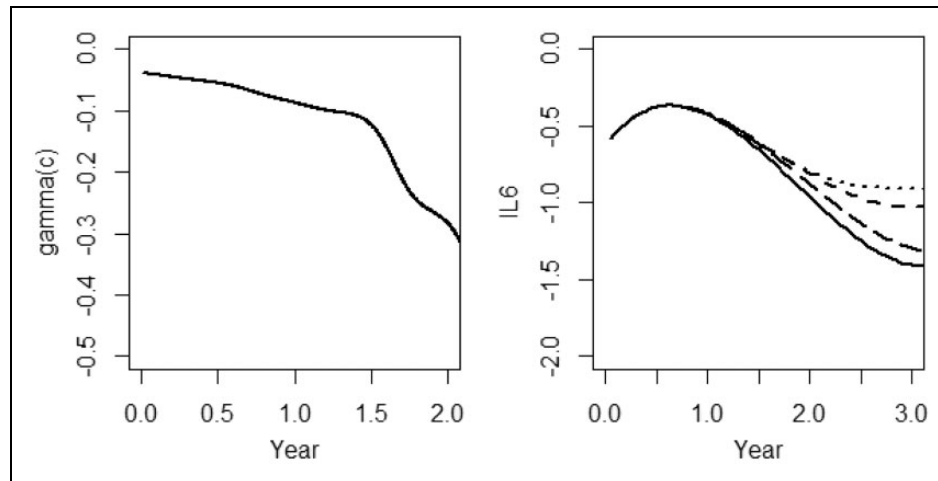


Figure 3. An estimated varying coefficient over 2 years (left) and four fitted IL-6 mean trajectories (right) at different times of treatment; year of 0.5 (solid), 1 (longdash), 1.5 (dashed), and 2 (dotted).

than the one in the model with a varying coefficient in terms of smaller mean squared errors. In addition, all the coverage probabilities in the former model are closer to the nominal level in the cases under consideration.

4 Real data analysis

In this section, we consider the proposed model for analyzing data from two different studies. The first illustration uses data on a biomarker on inflammation from a clinical trial of HIV-infected patients conducted in Haiti. The other uses data on body fat accretion from a prospective study of the development of obesity in a cohort of girls.

4.1 Influence of time of therapy on changes in the biomarker trend

This data set consists of longitudinal biomarker measurements from 408 HIV-infected adults as part of a clinical trial study conducted in Haiti. A description of the study and information on the whole data set can be found in Severe et al.³ Here, we present a brief summary. HIV-infected adults with CD4 counts between 200 and 350/mm³ at baseline were enrolled in the study and ART was given while patients were in care. In HIV-infected patients, inflammation is associated with other disease progression such as cardiovascular disease and chronic anemia. An inflammation biomarker, interleukin (IL)-6, was collected every year or less from enrollment before ART initiation and every 6 months thereafter. ART timeline was determined at the physician's discretion and ART initiation could be at any in-between visits other than scheduled visits. As a result, time of ART initiation varied across patients. Median pre-ART follow-up time was 1.3 years with first and third quartiles of 0.9 and 2.0 years, respectively. Median post-ART follow-up time was 2.7 years with first and third quartiles of 1.8 and 2.8 years, respectively.

The objective of the study for clinicians was to study the IL-6 mean trend over time while accounting for the subject-specific time of therapy and to assess whether the post-ART IL-6 mean trajectory shift depends on the elapsed time until treatment. Therefore, we define t_{ij} and c_i for patient i as time relative to enrollment at the j th visit and waiting time from enrollment to treatment, respectively, and explore the longitudinal mean trajectory of IL-6 by fitting the proposed model in equation (3) to the data. Through the goodness-of-fit test under the AR(1) correlation structure, the piecewise cubic polynomial model in which $\gamma(c_i)$ consists of 10 knots and 3-degree polynomial is chosen along with a P -value of 0.12 at a level of 0.05. Note that P -values for linear and quadratic polynomial models are 0.006 and 0.001, respectively.

We further conduct the hypothesis test of the constant value of $\gamma(c_i)$ over c and reject the null hypothesis with a P -value of 0.02. The estimated varying coefficient $\gamma(c_i)$ is drawn in Figure 3. The figure shows that the estimated coefficient is negative and becomes smaller over time. Table 2 reports estimated invariant coefficients along with standard errors, test statistics, and P -values. The results indicate that all coefficients are statistically significant at a level of 0.05. For ease of presentation, we further provide the fitted mean trend of outcomes at four different time

Table 2. Estimated coefficients, standard errors, and test statistics along with *P*-values in Sections 4.1 and 4.2.

Section	Coefficient	SE	Test statistic	<i>P</i> -value
4.1				
α_0	-0.617	0.069	-8.962	<0.001
α_1	0.851	0.158	5.382	<0.001
α_2	-0.837	0.123	-6.831	<0.001
α_3	0.183	0.028	6.530	<0.001
4.2				
λ_0	21.175	0.570	37.164	<0.001
λ_1	0.046	0.137	0.334	0.739
λ_2	2.164	0.220	9.824	<0.001
α_0	19.951	1.557	12.812	<0.001
α_1	0.103	0.133	0.773	0.439
γ_0	2.098	0.217	9.663	<0.001

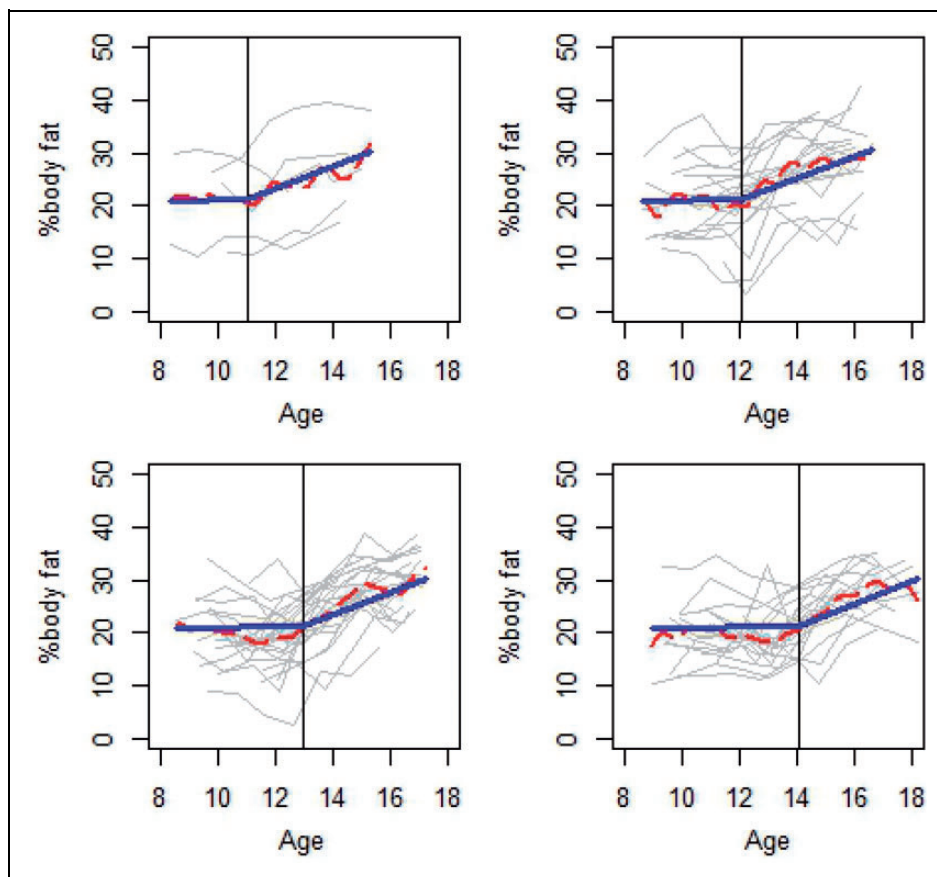


Figure 4. Blue curve is the fitted % body based on the proposed methods and red curve is fitted to a subset of data where age of menarche, displayed as a vertical line, is 11, 12, 13, and 14 using the kernel regression.

of treatment in Figure 3. This figure shows that although the average of IL-6 decreases after treatment regardless of treated times, the rate of decrease becomes smaller as time of ART initiation is delayed. This confirms that ART is effective at reducing an inflammation biomarker, but the effectiveness of ART decreases as time of ART initiation is delay.

4.2 Influence of menarche on changes in body fat accretion

An objective of the MIT Growth and Development Study^{1,2} is to explore the mean trend of body fat of girls over an adolescent period. As mentioned in Section 1, menarche has an influence on changes in body fat accretion. To be specific, increases in body fat in girls is likely to begin just before or around menarche. To examine the influence of menarche on changes in percent body fat, Naumova et al.¹³ analyzed 162 girls from a subset of data from the MIT Growth and Development Study. All girls were pre-menarcheal and nonobese at the start of the study and each subject was measured annually from approximately an age of 10 until 4 years after menarche. More details on these data can be found in Naumova et al.¹³

Naumova et al.¹³ defined time relative to menarche as $\tau_{ij} = t_{ij} - c_i$, where t_{ij} and c_i are the i th subject's age at the j th visit and age of menarche, respectively. This allows us to fit model (1) with $p = 1$ and the same time of event, i.e., $Y_i(t_{ij}) = \lambda_0 + \lambda_1\tau_{ij} + \lambda_2(\tau_{ij})_+ + \epsilon_{ij}$, where $\tau_{ij} = 0$ at menarche and Y_{ij} is the i th subject's percent body fat at the j th visit. Although the analysis can address the mean rates in change of percent body fat due to menarche as shown in Table 2, the mean percent body fat at a particular age of menarche is not feasible because time relative to menarche is used for modeling. As a result, the inability to investigate the mean trend of the percent of body fat with physical age could be interpreted as a weakness of this approach. Moreover, the linearity between outcomes and measurement times and the same mean rates change in percent body fat across age of menarche were assumed without proper evaluation beforehand.

To tackle these problems, we fit model (3) to the data under the AR(1) correlation structure. A goodness-of-fit test suggests that model (3) with the spline basis of $\gamma(c_i)$ having 10 knots and 3-degree polynomial basis functions fits the data sufficiently well only when the polynomial degree is one. Under the time-of-event-dependent piecewise linear model, we test $H_0 : \gamma(c_i) = \gamma_0$ and fail to reject H_0 with a P -value of 0.66. Hence, we fit the piecewise linear model with a constant value of γ_0 to the data and report the estimated coefficients, standard errors, and test statistics along with P -values in Table 2. The results show that an estimate of γ_0 is positive and statistically significant at a nominal level of 0.05, yet an estimated one of α_1 is not significantly different from zero with a P -value of 0.439. This suggests that the mean percent body fat remains the same before menarche, yet starts to increase at the same rate after menarche, regardless of the age of menarche. In Figure 4, we also provide the fitted piecewise linear mean trend of outcomes and the Nadaraya–Watson kernel regression curve obtained from a subset of data where the age of menarche is 11, 12, 13, and 14 years. The figures show that the proposed model successfully describes the pattern of percent body fat over time in all cases under consideration.

5 Concluding remarks

In biomedical studies, it is often of interest to evaluate the effect of an event on changes in an outcome. A randomized clinical trial can be a useful tool in determining the effect of the event while controlling for time of the event. In practice, however, many medical studies are not eligible for randomized clinical trials, which can lead to a subject-specific time of the event. Therefore, we have proposed a new statistical model to study repeatedly measured outcomes for longitudinal data in the presence of a subject-specific time of event. The proposed model enables us to determine not only whether the event has an impact on changes in the outcome but also whether the effectiveness is influenced by the time of the event.

In particular, we have modeled the mean response over time where the degree of the mean trajectory alteration is indexed by time of event. With the basis function approximation in equation (5), the proposed model can be readily fitted to the data by existing R package QIF or SAS Macro QIF. More details about this implementation can also be found in the Supplement. In addition to ease of the implementation, goodness-of-fit test based on QIF is readily available as a byproduct and is useful for researchers exploring various shapes of pattern and seeking an optimal degree of a polynomial. We take full advantage of QIF by employing the goodness-of-fit test to evaluate the effect of time of event to mean trend alteration as well. Our simulation studies have confirmed that QIF is an effective tool in both estimation and statistical inference. In addition, we have applied the proposed model to the biomarker study for HIV patients with different times of ART initiation and the body fat study in girls with their own age of menarche and confirmed different results on time of each event. That is, the body fatness turns to increase after menarche, yet the average rate of increase remains the same regardless of time of the event, whereas IL-6 is influenced by both ART and its initiated time in the HIV study.

Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Hyunkeun Ryan Cho  <https://orcid.org/0000-0003-2361-258X>

Seonjin Kim  <https://orcid.org/0000-0001-6058-0420>

Supplemental material

Supplemental material for this article is available online.

References

1. Bandini LG, Must A, Spadano JL, et al. Relation of body composition, parental overweight, pubertal stage, and race-ethnicity to energy expenditure among premenarcheal girls. *Am J Clin Nutr* 2002; **76**: 1040–1047.
2. Phillips SM, Bandini LG, Compton DV, et al. A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls. *J Nutr* 2003; **133**: 1419–1425.
3. Severe P, Juste MAJ, Ambroise A, et al. Early versus standard antiretroviral therapy for HIV-infected adults in Haiti. *New England J Med* 2010; **363**: 257–265.
4. Gallant AR and Fuller WA. Fitting segmented polynomial regression models whose join points have to be estimated. *J Am Stat Assoc* 1973; **68**: 144–147.
5. Qu A, Lindsay BG and Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**: 823–836.
6. Liang KY and Zeger SL. Longitudinal data analysis using generalised linear models. *Biometrika* 1986; **73**: 12–22.
7. Qu A and Lindsay BG. Building adaptive estimating equations when inverse of covariance estimation is difficult. *J R Stat Soc B* 2003; **65**: 127–142.
8. Yu Y and Ruppert D. Penalized spline estimation for partially linear single-index models. *J Am Stat Assoc* 2002; **97**: 1042–1054.
9. Qu A and Li R. Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* 2006; **62**: 379–391.
10. Ruppert D. Selecting the number of knots for penalized splines. *J Comput Graph Stat* 2002; **11**: 735–757.
11. Tian R, Xue L and Liu C. Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data. *J Multivariate Anal* 2014; **132**: 94–110.
12. Wang L and Qu A. Consistent model selection and data-driven tests for longitudinal data in the estimating equation approach. *J R Stat Soc B* 2009; **71**: 177–190.
13. Naumova EN, Must A and Laird NM. Tutorial in biostatistics: evaluating the impact of ‘critical periods’ in longitudinal studies of growth using piecewise mixed effects models. *Int J Epidemiol* 2001; **30**: 1332–1341.

Appendix I

The following standard conditions are imposed to study the asymptotic properties of $\hat{\theta}$ in the proposed marginal mean regression model:

- (1) There exists a $\theta_0 \in \Theta$, where Θ is the compact parameter space, such that $E\{g_i(\theta)\} = 0$ for $i = 1, \dots, n$ if and only if $\theta = \theta_0$.
- (2) The vector $g_i(\theta)$ is continuously differentiable with respect to θ and $\Phi = E\{\partial g_i(\theta_0)/\partial \theta\}$ is of full rank.
- (3) The matrix $\Sigma = E\{g_i(\theta_0)g_i(\theta_0)^\top\}$ is positive definite.

Proof of Theorem 1. By Taylor expansion, we have

$$g(\hat{\theta}) = g(\check{\theta}_0) + \dot{g}(\theta)(\hat{\theta} - \theta_0) \quad (11)$$

where $\dot{g}(\check{\theta}) = \partial g(\theta)/\partial \theta$ and $\check{\theta}$ lies between $\hat{\theta}$ and θ_0 . By multiplying the equation (11) by $\dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1}$, it immediately follows that

$$\dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} g(\hat{\theta}) = \dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} g(\theta_0) + \dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} \dot{g}(\check{\theta})(\hat{\theta} - \theta_0)$$

Note that the left hand side $\dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} g(\hat{\theta}) = 0$ because $\hat{\theta}$ is the minimizer of $g(\theta)^\top V(\theta)^{-1} g(\theta)$. Therefore, the equation can be rearranged as

$$(\hat{\theta} - \theta_0) = -\{\dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} \dot{g}(\check{\theta})\}^{-1} \dot{g}(\hat{\theta})^\top V(\hat{\theta})^{-1} g(\theta_0) \tag{12}$$

It follows from $\dot{g}(\hat{\theta}) \xrightarrow{p} \Phi$, $V(\hat{\theta}) \xrightarrow{p} \Sigma$, and $\sqrt{n}g(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ that we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (\Phi^\top \Sigma^{-1} \Phi)^{-1})$$

Proof of Theorem 2. By Taylor expansion and equation (12), we have

$$\begin{aligned} g(\hat{\theta}) &= g(\theta_0) + \dot{g}(\theta_0)(\hat{\theta} - \theta_0) + o_p(1) \\ &= \{I_{d(p+\ell)} - \Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1}\} g(\theta_0) + o_p(1) \end{aligned} \tag{13}$$

where $I_{d(p+\ell)}$ is the $d(p + \ell)$ -dimensional identity matrix. It follows from (13) and $V(\hat{\theta}) = \Sigma + o_p(1)$ that

$$\begin{aligned} Q_{p,\ell}(\hat{\theta}) &= ng(\hat{\theta})^\top V(\hat{\theta})^{-1} g(\hat{\theta}) \\ &= ng(\theta_0)^\top \{\Sigma^{-1} - \Sigma^{-1} \Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1}\} g(\theta_0) + o_p(1) \\ &= \{\sqrt{n}\Sigma^{-1/2}g(\theta_0)\}^\top \{I_{d(p+\ell)} - \Sigma^{-1/2} \Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1/2}\} \{\sqrt{n}\Sigma^{-1/2}g(\theta_0)\} + o_p(1) \end{aligned}$$

Since $\sqrt{n}\Sigma^{-1/2}g(\theta_0) \xrightarrow{d} \mathcal{N}(0, I_{d(p+\ell)})$, $\Sigma^{-1/2} \Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1/2}$ is an idempotent and symmetric matrix, and its trace is $p + \ell$, $Q_{p,\ell}(\hat{\theta})$ converges to a chi-squared distribution with $(d - 1)(p + \ell)$ degrees of freedom.

Proof of Theorem 3. The null hypothesis can be rewritten as $H_0 : \gamma_1 = \dots = \gamma_{q+u} = 0$, where $\theta = (\alpha_0, \alpha_1, \dots, \alpha_p, \gamma_0, \gamma_1, \dots, \gamma_{q+u})^\top = (\theta_s^\top, \theta_\gamma^\top)^\top$ and $\theta_s = (\alpha_0, \alpha_1, \dots, \alpha_p, \gamma_0)^\top$. Under H_0 , i.e., θ_γ is a zero vector, $\tilde{\theta}_s$ is obtained by minimizing $ng_s(\theta_s)^\top V_s(\theta_s)^{-1} g_s(\theta_s)$, where $V_s(\theta_s)$ is a consistent estimate of $E\{g_s(\theta_s)g_s(\theta_s)^\top\}$ and $g_s(\theta_s)$ is a subset of $g(\theta)$ associated with θ_s . We accordingly denote $\Phi = (\Phi_s, \Phi_\gamma)$, where $\Phi_s = E\{\partial g(\theta)/\partial \theta_s\}$. It consequently follows that under H_0 , we have

$$\begin{aligned} T &= Q_{p,\ell}(\tilde{\theta}) - Q_{p,\ell}(\hat{\theta}) \\ &= ng(\theta_0)^\top \Sigma^{-1} \{\Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top - \Phi_s(\Phi_s^\top \Sigma^{-1} \Phi_s)^{-1} \Phi_s^\top\} \Sigma^{-1} g(\theta_0) + o_p(1) \end{aligned}$$

where

$$\begin{aligned} \Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top &\geq \begin{pmatrix} \Phi_s & \Phi_\gamma \end{pmatrix} \begin{pmatrix} \{\Phi_s^\top \Sigma^{-1} \Phi_s\}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Phi_s^\top \\ \Phi_\gamma^\top \end{pmatrix} \\ &= \Phi_s \{\Phi_s^\top \Sigma^{-1} \Phi_s\}^{-1} \Phi_s^\top \end{aligned}$$

Since both $\Phi(\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top$ and $\Phi_s(\Phi_s^\top \Sigma^{-1} \Phi_s)^{-1} \Phi_s^\top$ are idempotent and symmetric matrices with trace equal to $p + \ell$ and $p + 1$, respectively, the test statistic converges to a chi-squared distribution with $\ell - 1$ degrees of freedom.