MODEL SELECTION CRITERION BASED ON THE PREDICTION MEAN SQUARED ERROR IN GENERALIZED ESTIMATING EQUATIONS

Yu Inatsu[*1], Shinpei Imori
Department of mathematics, Graduate School of Science, Hiroshima University

ABSTRACT

The present paper considers a model selection criterion in regression models using generalized estimating equation (GEE). Using the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called PMSEG that reflects the correlation between responses. A numerical study reveals that the PMSEG has better performance than previous other criteria for model selection.

Key Words: Generalized estimating equations, Longitudinal data, Prediction mean squared error, Model selection.

1. INTRODUCTION

In real data analysis, correlated data are often discussed in health sciences, medical sciences, economics and many other fields. Longitudinal data, defined from observations of subjects measured repeatedly over time, often arise in these fields as an important example. In general, observations from each subject in longitudinal data are correlated. Liang and Zeger (1986) introduced an extension of the generalized linear model (Nelder and Wedderburn, 1972) to the analysis of longitudinal data, known as the generalized estimating equation (GEE) method. Defining features of the GEE method are that we can use a working (but not necessarily correct) correlation matrix as the correlation matrix, and a full specification of the joint distribution is not required. Hence, the GEE method is widely used in many fields for analyzing longitudinal data.

In addition, the model selection problem in the GEE methodology is also an important. The goodness of fit of the model is commonly measured by a risk function, and the model selection is performed by obtaining a certain estimator of the risk function. For example, the risk function based on the expected Kullback-Leibler (KL) information (Kullback and Leibler, 1951) is often used, and the most famous estimator of the risk function is Akaike's information criterion (AIC) proposed by Akaike (1973, 1974). The AIC is obtained by using the likelihood, it can be simply

[*1] Corresponding author
E-mail address: m126391@hiroshima-u.ac.jp

defined as AIC $= -2 \times$ the maximum log likelihood $+ 2 \times$ the number of parameters. Furthermore, Nishii (1984), Rao (1988) proposed the GIC as a general extension of the AIC, which is widely applied for selecting the best model, and considered about various properties in many literatures.

However, in the GEE method, the maximum likelihood based model selection criteria such as the AIC or GIC, are not applicable directly because the GEE method is not likelihood based. Some model selection criteria like the AIC or GIC in the GEE method have been already proposed. For example, Pan (2001) proposed the QIC that is based on the quasi likelihood (Wedderburn, 1974). Cantoni *et al.* (2005) proposed the GC_p as a general extension of the Mallows's C_p (Mallows, 1973). Hin and Wang (2009), Gosho *et al.* (2011) proposed the CIC to select the correlation structure. Unfortunately, above model selection criteria are not derived by taking account into the correlation between responses. For example, the risk function of the QIC is based on the independent quasi likelihood. In this respect, these criteria are not reflective of the significant feature in longitudinal data.

The principal aim of the present paper is to obtain a new model selection criterion that reflects the correlation between responses. In this study, we have focused on deciding the best subset of variables. By using the risk function based on the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called the PMSEG (the prediction mean squared error in the GEE).

The remainder of the present paper is organized as follows: In Section 2, we consider a stochastic expansion of a GEE estimator. In Section 3, we propose the PMSEG. In Section 4, we verify that the proposed PMSEG has good properties by conducting numerical experiments. In Section 5, we conclude our discussion. Technical details are provided in the Appendix.

2. STOCHASTIC EXPANSION OF THE GEE ESTIMATOR

Let y_{ij} be a scalar response variable, and let $\boldsymbol{x}_{*,ij}$ be a l-dimensional nonstochastic vector consists of all possible explanatory variables for the ith subject at the jth occasion, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Assume that response variables from different subjects are independent and response variables from the same subject are correlated. For $i = 1, \ldots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$, $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \ldots, \boldsymbol{x}_{*,im})'$, and let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im})'$ be a $m \times p$ submatrix of the matrix $\boldsymbol{X}_{*,i}$. Liang and Zeger (1986) used the GLM to model the marginal density of y_{ij},

$$f(y_{ij}; \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \phi) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(y_{ij}, \phi)], \tag{2.1}$$

where, $a(\cdot)$, $b(\cdot)$ are known functions, θ_{ij} is an unknown location parameter and ϕ is a scale parameter. In the GLM framework, the location parameter $\theta_{ij} = u(\eta_{ij}) = \theta_{ij}(\boldsymbol{\beta})$, where $u(\cdot)$ is a known function, $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is a p-dimensional unknown parameter. In the present paper, we assume that ϕ is known, and we also assume Θ to be the *natural parameter space* (see, Xie and Yang, 2003) of the exponential family of distributions presented in (2.1), and the interior of Θ is

denoted as Θ^0. Then Θ is convex, and in the Θ^0, all derivatives of $a(\cdot)$ and all moments of y_{ij} exist. Under such model conditions, the first two moments of y_{ij} are given by

$$\mu_{ij}(\boldsymbol{\beta}) = \mathrm{E}[y_{ij}] = \dot{a}(\theta_{ij}),\ \sigma^2_{ij}(\boldsymbol{\beta}) = \mathrm{Cov}[y_{ij}] = \ddot{a}(\theta_{ij})\phi \equiv \nu(\mu_{ij}(\boldsymbol{\beta}))\ \text{(say)}.$$

In this situation, the expectation of the response y_{ij} is modeled by a link function $g(t) = (\dot{a} \circ u)^{-1}(t)$ and the linear predictor η_{ij}, i.e., $g(\mu_{ij}(\boldsymbol{\beta})) = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$. When $u(s) = s$, we say that $g(t) = (\dot{a})^{-1}(t)$ is the natural link function. For example, the logistic regression model is defined with the natural link function. We call the model with $\boldsymbol{x}_{*,ij}$ or $\boldsymbol{x}_{ij}$ as the full or candidate model, respectively. We assume that the true probability density function of y_{ij} can be written as (2.1), i.e., the true model is one of the candidate models.

Denote $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{im}(\boldsymbol{\beta}))'$, $\boldsymbol{A}_i(\boldsymbol{\beta}) = \mathrm{diag}(\sigma^2_{i1}(\boldsymbol{\beta}), \ldots, \sigma^2_{im}(\boldsymbol{\beta}))$ and $\boldsymbol{\Delta}_i(\boldsymbol{\beta}) = \mathrm{diag}(\partial\theta_{i1}/\partial\eta_{i1}, \ldots, \partial\theta_{im}/\partial\eta_{im})$, where $\mathrm{diag}(a_1, \ldots, a_s)$ denotes the $s \times s$ diagonal matrix whose the (i,i)th element is a_i. We write $\boldsymbol{D}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i(\boldsymbol{\beta})\boldsymbol{\Delta}_i(\boldsymbol{\beta})\boldsymbol{X}_i$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_i^*\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})$, where $\boldsymbol{R}_i^*$ is the true correlation matrix, and $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})$. Here, $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the working correlation matrix that one can choose freely, which may possibly have a nuisance parameter $\boldsymbol{\alpha}$. Depending on the situation, we can choose some useful working correlation matrices. For example, "Independence" (i.e., $(\boldsymbol{R})_{jk} = 0$, if $j \neq k$), "Exchangeable" (i.e., $(\boldsymbol{R})_{jk} = \alpha$, if $j \neq k$), "(first-order) Autoregressive" (i.e., $(\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha^{j-k}$, if $j > k$), "1-dependent" (i.e., $(\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha$, if $j = k+1$) and "Unstructured" (i.e., $(\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha_{jk}$, if $j > k$). If $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is equal to $\boldsymbol{R}_i^*$, then $\boldsymbol{V}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}) = \boldsymbol{\Sigma}_i(\boldsymbol{\beta}_0) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\boldsymbol{R}_i^*\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0) = \mathrm{Cov}[\boldsymbol{y}_i]$, where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$.

Liang and Zeger (1986) proposed the GEE as follows:

$$\boldsymbol{s}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}'_i(\boldsymbol{\beta})\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}_p, \tag{2.2}$$

where $\boldsymbol{0}_p$ is a p-dimensional vector of zeros. An estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is defined as a solution of the equation (2.2), and the estimator is called the GEE estimator. In the present paper, we assume that $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{R}(\boldsymbol{\alpha})$ and $\boldsymbol{R}_i^* = \boldsymbol{R}_0$. Moreover, to simplify our discussion, we also assume that the nuisance parameter $\boldsymbol{\alpha}$ is known. Hence, we write $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{V}_i(\boldsymbol{\beta})$.

In order to propose a new model selection criterion at Section 3, a stochastic expansion of $\hat{\boldsymbol{\beta}}$ is needed. In this section, we obtain the stochastic expansion of $\hat{\boldsymbol{\beta}}$ up to the order n^{-1}. For simplicity, we omit $(\boldsymbol{\beta})$ from functions of $\boldsymbol{\beta}$ like $\mu_{ij}(\boldsymbol{\beta}) = \mu_{ij}$. Furthermore, in order to distinguish a function of $\boldsymbol{\beta}$ evaluated at the true parameter $\boldsymbol{\beta}_0$ and GEE estimator $\hat{\boldsymbol{\beta}}$, we write such as $\mu_{ij}(\boldsymbol{\beta}_0) = \mu_{ij,0}$ and $\mu_{ij}(\hat{\boldsymbol{\beta}}) = \hat{\mu}_{ij}$, respectively. Furthermore, in order to assure asymptotic properties of the GEE estimator, we consider the following regularity assumptions (see, e.g., Xie and Yang, 2003):

C1. $\boldsymbol{\beta}_0$ is in an admissible set $\mathcal{B}$, where $\mathcal{B}$ is an open set in $\mathbb{R}^p$ for the parameter $\boldsymbol{\beta}$.

C2. $\boldsymbol{x}'_{ij}\boldsymbol{\beta} \in g(\mathcal{M})$ for all $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{M}$ is the image of $\dot{a}(\Theta^0)$.

C3. $u(\eta_{ij})$ is four times continuously differentiable and $\dot{u}(\eta_{ij}) > 0$ in $g(\mathcal{M})^0$.

C4. $\boldsymbol{H}_{n,0}$ and $\boldsymbol{M}_{n,0}$ are positive definite when n is large, where

$$\boldsymbol{H}_n = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \ \boldsymbol{M}_n = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i.$$

Condition C1 and C2 are necessary to have the GLM for all $\boldsymbol{\beta}$. Condition C3 and C4 are necessary to calculate the bias. In addition, in order to assure the strong consistency, asymptotic normality and uniqueness of the GEE estimator, we consider the following additional assumptions, which can be derived slightly modifying the results reported by Xie and Yang (2003):

C5. $\liminf_{n\to\infty} \lambda_{\min}(\boldsymbol{H}_{n,0}/n) > 0$, where $\lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of symmetric matrix $\boldsymbol{A}$.

C6. Sequence $\{\boldsymbol{x}_{ij}\}$ lies in $\mathcal{X}$ with $u(\boldsymbol{x}_{ij}\boldsymbol{\beta}) \in \Theta^0, \ \boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{X}$ is a compact set for regressors $\boldsymbol{x}_{ij}$.

C7. In a neighborhood of $\boldsymbol{\beta}_0$, say N, there exist a constant $c_0 > 0$ and some $\delta > 0$, independent of n, such that, when $n \to \infty$, for any p-dimensional vector $\boldsymbol{\lambda}$, $\|\boldsymbol{\lambda}\| = 1$,

$$\boldsymbol{\lambda}' \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}'} \boldsymbol{\lambda} \geq c_0 \lambda_{\max}^{(1/2)+\delta}(\boldsymbol{M}_{n,0}), \quad \text{a.s. for } \boldsymbol{\beta} \in N.$$

C8. The equation (2.2) has a unique solution when n is large.

Here $\lambda_{\max}(\boldsymbol{A})$ is the largest eigenvalue of symmetric matrix $\boldsymbol{A}$. According to THEOREM 7 and COROLLARY 1 in Xie and Yang (2003), $\hat{\boldsymbol{\beta}}$ has the strong consistency and asymptotic normality under these conditions. Furthermore, from C5, $\boldsymbol{H}_{n,0} = O(n)$.

Based on the above conditions, to perform the stochastic expansion of $\hat{\boldsymbol{\beta}}$, we focus on the fact that $\hat{\boldsymbol{s}}_n = \boldsymbol{0}_p$. By applying a Taylor expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ to this equation, the GEE is expanded as follows:

$$\begin{aligned}
\boldsymbol{0}_p &= \boldsymbol{s}_{n,0} + \left.\frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_p\} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}'} \right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= \boldsymbol{s}_{n,0} - \boldsymbol{H}_{n,0}(\boldsymbol{I}_p + \boldsymbol{G}_{1,0} + \boldsymbol{G}_{2,0})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_p\} \boldsymbol{L}_1(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \qquad (2.3)
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, $\boldsymbol{I}_p$ is a p-dimensional identity matrix, $\boldsymbol{G}_{1,0}$, $\boldsymbol{G}_{2,0}$ and $\boldsymbol{L}_1(\boldsymbol{\beta}^*)$ are defined by

$$\begin{aligned}
\boldsymbol{G}_{1,0} &= -\boldsymbol{H}_{n,0}^{-1} \sum_{i=1}^{n} \boldsymbol{D}_{i,0}' \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{V}_i^{-1} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) \{\boldsymbol{I}_p \otimes (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\}, \\
\boldsymbol{G}_{2,0} &= -\boldsymbol{H}_{n,0}^{-1} \sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{D}_i' \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) [\boldsymbol{I}_p \otimes \{\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\}], \\
\boldsymbol{L}_1(\boldsymbol{\beta}^*) &= \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}'} \right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.
\end{aligned}$$

Note that for a matrix $\boldsymbol{W} = (w_{ij})$, the derivative of $\boldsymbol{W}$ by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and β_k are respectively defined as follows:

$$\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{W} = \left(\frac{\partial \boldsymbol{W}}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{W}}{\partial \beta_p} \right), \quad \frac{\partial}{\partial \boldsymbol{\beta}} \otimes \boldsymbol{W} = \left(\frac{\partial \boldsymbol{W}'}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{W}'}{\partial \beta_p} \right)', \quad \frac{\partial \boldsymbol{W}}{\partial \beta_k} = \left(\frac{\partial w_{ij}}{\partial \beta_k} \right).$$

Also note that $\boldsymbol{L}_1(\boldsymbol{\beta}^*) = O_p(n)$, $\boldsymbol{s}_{n,0} = O_p(n^{1/2})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \boldsymbol{G}_{1,0}, \boldsymbol{G}_{2,0} = O_p(n^{-1/2})$. Thus, (2.3) yields

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{H}_{n,0}^{-1} \boldsymbol{s}_{n,0} + O_p(n^{-1}) = \boldsymbol{b}_{1,0} + O_p(n^{-1}). \tag{2.4}$$

Similarly, the GEE can also be expanded as follows:

$$\begin{aligned} \boldsymbol{s}_{n,0} = \boldsymbol{H}_{n,0}(\boldsymbol{I}_p + \boldsymbol{G}_{1,0} + \boldsymbol{G}_{2,0})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_p\}\{\boldsymbol{G}_{3,0} + (\boldsymbol{L}_1(\boldsymbol{\beta}_0) - \boldsymbol{G}_{3,0})\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ - \frac{1}{6}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_p\} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}'} \right) \right\} \Bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{**}} \{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\beta}^{**}$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{G}_{3,0} = \mathrm{E}[\boldsymbol{L}_1(\boldsymbol{\beta}_0)]$.

Note that the order of the last term of equation (2.5) is $O_p(n^{-1/2})$. Also note that $\boldsymbol{G}_{3,0} = O(n)$ and $\boldsymbol{L}_1(\boldsymbol{\beta}_0) - \boldsymbol{G}_{3,0} = O_p(n^{1/2})$. Therefore, by using equation (2.4) and (2.5), $\hat{\boldsymbol{\beta}}$ can be expanded as follows:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{b}_{1,0} + \boldsymbol{b}_{2,0} + O_p(n^{-3/2}), \tag{2.6}$$

where $\boldsymbol{b}_{2,0} = \boldsymbol{H}_{n,0}^{-1}(\boldsymbol{b}_{1,0}' \otimes \boldsymbol{I}_p)\boldsymbol{G}_{3,0}\boldsymbol{b}_{1,0}/2 - \boldsymbol{G}_{1,0}\boldsymbol{b}_{1,0} - \boldsymbol{G}_{2,0}\boldsymbol{b}_{1,0}$. Note that $\boldsymbol{b}_{1,0}$ and $\boldsymbol{b}_{2,0}$ are $O_p(n^{-1/2})$ and $O_p(n^{-1})$, respectively.

3. MAIN RESULT

In this section, we propose a new model selection criterion. The goodness of fit of the model is measured by the risk function based on the PMSE normalized by the covariance matrix as follows:

$$\mathrm{Risk_P} = \mathrm{PMSE} - mn = \mathrm{E}_y \left[\mathrm{E}_z \left[\sum_{i=1}^{n} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)' \boldsymbol{\Sigma}_{i,0}^{-1} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i) \right] \right] - mn,$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})'$ is an m-dimensional random vector that is independent of $\boldsymbol{y}_i$ and has the same distribution as $\boldsymbol{y}_i$. It is easy to show that if $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}_0$, then $\mathrm{Risk_P}$ has the minimum value of zero, i.e., the PMSE has the minimum value of mn. For this reason, we would like to select the model, which has the minimum PMSE. However, since the PMSE is typically unknown, we must estimate it.

We define $\boldsymbol{R}_0(\boldsymbol{\beta})$, $\mathcal{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\mathcal{L}^*(\boldsymbol{\beta})$ as follows:

$$\begin{aligned}
\boldsymbol{R}_0(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1/2}, \\
\mathcal{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1))'\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2)\boldsymbol{R}_0^{-1}(\boldsymbol{\beta}_2)\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2)(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)), \\
\mathcal{L}^*(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i).
\end{aligned}$$

Based on the above, we estimate PMSE by $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$, where $\hat{\boldsymbol{\beta}}_f$ is a GEE estimator that is obtained under the "full" model. Specifically, $\hat{\boldsymbol{\beta}}_f$ is defined as the solution to the following equation:

$$\boldsymbol{s}_{f,n}(\boldsymbol{\beta}_*) = \sum_{i=1}^{n} \boldsymbol{D}_i'(\boldsymbol{\beta}_*)\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_*, \boldsymbol{\alpha}^*)(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_*)) = \boldsymbol{0}_l,$$

where $\boldsymbol{D}_i(\boldsymbol{\beta}_*) = \boldsymbol{A}_i(\boldsymbol{\beta}_*)\boldsymbol{\Delta}_i(\boldsymbol{\beta}_*)\boldsymbol{X}_{*,i}$, $\boldsymbol{V}_i(\boldsymbol{\beta}_*, \boldsymbol{\alpha}^*) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_*)\bar{\boldsymbol{R}}_i(\boldsymbol{\alpha}^*)\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_*)$ and $\bar{\boldsymbol{R}}_i(\boldsymbol{\alpha}^*)$ is a positive definite working correlation matrix that one can choose freely. We assume that the nuisance parameter $\boldsymbol{\alpha}^*$ is known. Note that $\boldsymbol{\beta}_*$ is an l-dimensional unknown parameter under the full model. Also note that $\bar{\boldsymbol{R}}_i(\boldsymbol{\alpha}^*)$ is the same for all candidate models. The reason for using $\hat{\boldsymbol{\beta}}_f$ is discussed later. For simplicity, we write $\mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_2) = \mathcal{L}(\boldsymbol{\beta}_2)$ and $\mathcal{L}^*(\boldsymbol{\beta}_0) = \mathcal{L}^*$.

Since $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$ is not an asymptotically unbiased estimator of PMSE, we evaluate the asymptotic bias in order to propose the new model selection criterion. The bias when we estimate the PMSE by $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$ is given as

$$\begin{aligned}
\text{Bias} = \text{PMSE} - \text{E}_y[\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)] &= \{\text{Risk}_\text{P} - \text{E}_y[\mathcal{L}^*(\hat{\boldsymbol{\beta}})]\} + \{\text{E}_y[\mathcal{L}^*(\hat{\boldsymbol{\beta}}) - \mathcal{L}^*)]\} \\
&\quad + \{\text{E}_y[\mathcal{L}^* - \mathcal{L}(\hat{\boldsymbol{\beta}}_f)]\} + \{\text{E}_y[\mathcal{L}(\hat{\boldsymbol{\beta}}_f) - \mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)]\} \\
&= \text{Bias1} + \text{Bias2} + \text{Bias3} + \text{Bias4}. \qquad (3.1)
\end{aligned}$$

Here, we can see that Bias3 in (3.1) satisfies

$$\begin{aligned}
\text{Bias3} &= \text{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})'\{\boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\right] \\
&= mn - \text{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\right].
\end{aligned}$$

Therefore, we can ignore the calculation of Bias3 because it is not dependent on candidate models.

Similarly, Bias1 in (3.1) is expanded as

$$\begin{aligned}
\text{Bias1} &= \text{E}_y\left[\text{E}_z\left[\sum_{i=1}^{n} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)\right] - \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)\right] \\
&= 2\text{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})'\boldsymbol{\Sigma}_{i,0}^{-1}(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_{i,0})\right]. \qquad (3.2)
\end{aligned}$$

Applying a Taylor expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ to $\hat{\boldsymbol{\mu}}_i$ yields

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_i &= \boldsymbol{\mu}_{i,0} + \left.\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_m\} \left.\left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right)\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \frac{1}{6}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_m\} \left.\left\{\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right)\right\}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{***}} \{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} \\
&= \boldsymbol{\mu}_{i,0} + \boldsymbol{D}_{i,0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \otimes \boldsymbol{I}_m\} \boldsymbol{D}_{i,0}^{(1)} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O_p(n^{-3/2}), \qquad (3.3) \\
\boldsymbol{D}_{i,0}^{(1)} &= \left.\left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \boldsymbol{D}_i\right)\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.
\end{aligned}$$

Here, $\boldsymbol{\beta}^{***}$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Substituting the stochastic expansion of $\hat{\boldsymbol{\beta}}$ in (2.6) into (3.3) yields the following:

$$\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_{i,0} = \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0} + \left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\} + O_p(n^{-3/2}). \qquad (3.4)$$

By combining (3.2) and (3.4), we obtain

$$\begin{aligned}
\frac{1}{2}\text{Bias1} &= \mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
&\quad + \mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\}\right] \qquad (3.5) \\
&\quad + \mathrm{E}_y[O_p(n^{-1/2})].
\end{aligned}$$

Note that $\mathrm{E}[(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})'(\boldsymbol{y}_j - \boldsymbol{\mu}_{j,0})] = 0,\ (i \neq j)$, the first term of (3.5) can be calculated as

$$\begin{aligned}
\mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] &= \mathrm{E}_y\left[\sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{j,0}'\boldsymbol{V}_{j,0}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_{j,0})\right] \\
&= \mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}'\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\right] \\
&= \mathrm{tr}\left[\boldsymbol{H}_{n,0}^{-1}\sum_{i=1}^{n} \boldsymbol{D}_{i,0}'\boldsymbol{V}_{i,0}^{-1}\boldsymbol{D}_{i,0}\right] = \mathrm{tr}[\boldsymbol{H}_{n,0}^{-1}\boldsymbol{H}_{n,0}] = p. \qquad (3.6)
\end{aligned}$$

Similarly, since $\mathrm{E}[(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0}) \otimes (\boldsymbol{y}_j - \boldsymbol{\mu}_{j,0})'(\boldsymbol{y}_k - \boldsymbol{\mu}_{k,0})] = \boldsymbol{0}_m$, (not $i = j = k$), the second term of (3.5) can be expanded as

$$\begin{aligned}
&\mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\}\right] \\
&= \mathrm{E}_y\left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2i,0} + \frac{1}{2}(\boldsymbol{b}_{1i,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1i,0}\right\}\right],
\end{aligned}$$

where $\boldsymbol{b}_{1i,0} = \boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}'\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})$, $\boldsymbol{b}_{2i,0} = \boldsymbol{H}_{n,0}^{-1}(\boldsymbol{b}_{1i,0}' \otimes \boldsymbol{I}_p)\boldsymbol{G}_{3,0}\boldsymbol{b}_{1i,0}/2 - \boldsymbol{G}_{1i,0}\boldsymbol{b}_{1i,0} - \boldsymbol{G}_{2i,0}\boldsymbol{b}_{1i,0}$,

$$\boldsymbol{G}_{1i,0} = -\boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}' \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{V}_i^{-1} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) \{\boldsymbol{I}_p \otimes (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\},$$

$$\boldsymbol{G}_{2i,0} = -\boldsymbol{H}_{n,0}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{D}_i' \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) [\boldsymbol{I}_p \otimes \{\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\}].$$

Note that $\boldsymbol{D}_{i,0}\boldsymbol{b}_{2i,0} + (\boldsymbol{b}_{1i,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1i,0}/2 = O_p(n^{-2})$, the second term of (3.5) can be obtained as

$$\mathrm{E}_y \left[\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{\Sigma}_{i,0}^{-1} \left\{ \boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}' \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0} \right\} \right] = O(n^{-1}). \tag{3.7}$$

Under certain conditions, the limit of the expectation is equal to the expectation of the limit. Furthermore, in many cases of practical interest, a moment of statistic can be expanded as power series in n^{-1} (see e.g., Hall, 1992). Therefore, by substituting (3.6) and (3.7) into (3.5), we obtain the asymptotic expansion of Bias1 up to order 1 as

$$\mathrm{Bias1} = 2p + O(n^{-1}). \tag{3.8}$$

Moreover, by using the same argument of the calculation of Bias1, we obtain

$$\mathrm{Bias2} + \mathrm{Bias4} = O(n^{-1}). \tag{3.9}$$

The derivation of (3.9) is shown in Appendix.

Consequently, by substituting (3.8) and (3.9) into (3.1), we obtain the asymptotic expansion of Bias up to order 1 as

$$\mathrm{Bias} = 2p + \mathrm{Bias3} + O(n^{-1}). \tag{3.10}$$

Recall that Bias3 is not dependent on candidate models. Hence, the PMSEG can then be defined by adding an estimated (Bias − Bias3) to $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$, i.e.,

$$\mathrm{PMSEG} = \mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f) + 2p. \tag{3.11}$$

(3.11) is our proposed model selection criterion called the PMSEG (the prediction mean squared error in the GEE). Recall that $\hat{\boldsymbol{\beta}}_f$ is estimated under the full model and it is not dependent on candidate models. Since the covariance matrix in the PMSEG is estimated by $\hat{\boldsymbol{\beta}}_f$, the PMSEG can be simply defined. If the covariance matrix is estimated by $\hat{\boldsymbol{\beta}}$, Bias3 in (3.11) is different for each candidate model. Unfortunately, it is difficult and too expensive to calculate Bias3. This is one of the advantages of using $\hat{\boldsymbol{\beta}}_f$ for estimating the covariance matrix. For actual use, we recommend to use the working independence matrix in order to obtain $\hat{\boldsymbol{\beta}}_f$ since we can get $\hat{\boldsymbol{\beta}}_f$ easily by omitting the calculations of the working correlation matrix. Fitzmaurice (1995) mentioned that the GEE estimator under the working independence assumption is often inefficient. Nevertheless, from some simulation results, we confirmed that the estimation of the covariance matrix using the inefficient estimator does not dramatically influence the performance of the PMSEG.

4. NUMERICAL STUDIES

In this section, we confirm a usefulness of the PMSEG through comparisons with the QIC, which is a representative previous study. The QIC is an estimator of a risk function based on the quasi likelihood under the independence assumption. The risk function that is estimated by the QIC, which is called $\mathrm{Risk_Q}$ in this paper, and the quasi likelihood $Q(\cdot)$ are defined as follows:

$$\mathrm{Risk_Q} = \mathrm{E}_y\left[\mathrm{E}_z\left[-2\sum_{i=1}^{n}\sum_{j=1}^{m} Q(\hat{\boldsymbol{\beta}}; z_{ij})\right]\right],\ Q(\hat{\boldsymbol{\beta}}; z_{ij}) = \int_{z_{ij}}^{g^{-1}(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}})} \frac{z_{ij}-t}{\nu(t)}dt,$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})'$ is an m-dimensional random vector that is independent of $\boldsymbol{y}_i$ and has the same distribution as $\boldsymbol{y}_i$. Note that since the $\mathrm{Risk_Q}$ is considered under the independence assumption, both the $\mathrm{Risk_Q}$ and QIC are not reflective of the correlation between responses. Also note that the PMSEG and QIC are estimators of the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$, respectively. In general, the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$ are different, i.e., the PMSEG and QIC are criteria from different viewpoints.

At the beginning, we examine the numerical studies for the frequencies of candidate models and the prediction error of the best models selected by the PMSEG and QIC. We prepared the fifteen candidate models with $n = 100$ and $m = 3$. First, we constructed a 3×5 explanatory variable matrix $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \boldsymbol{x}_{*,i2}, \boldsymbol{x}_{*,i3})'$, $i = 1, \ldots, 100$. The first column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_3$, where $\mathbf{1}_3$ is a 3-dimensional vector of ones, and the second column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_3 \times \varsigma_i$, where $\varsigma_1, \ldots, \varsigma_{100}$ are i.i.d. binomial distribution $B(1, 0.5)$. The third column of $\boldsymbol{X}_{*,i}$ is $(0, 1, 2)'$, and the remaining six elements of $\boldsymbol{X}_{*,i}$ were defined by realizations of independent variables with uniform distribution on the interval $[-1, 1]$.

In this simulation, we prepared two situations, as follows:

$$\begin{aligned}
&\text{Case 1}: \ \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.85^{|j-j^*|},\ \boldsymbol{\beta}_0 = (0.25, -0.25, -0.25, 0, 0)',\\
&\text{Case 2}: \ \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.35,\ \boldsymbol{\beta}_0 = (0.25, -0.25, -0.25, 0, 0)'.
\end{aligned}$$

The explanatory variables matrix for the ith subject in the $(5k + k^*)$th model consists of the first k^* column of $\boldsymbol{X}_{*,i}$, $k = 0, 1, 2$, $k^* = 1, \ldots, 5$. For the working correlation matrix, we prepare three different matrices, working exchangeable matrix ($k = 0$), working autoregressive matrix ($k = 1$) and working independence matrix ($k = 2$). Thus, in case 1, the true model is the eighth model, and in case 2, the true model is the third model. We simulated 10,000 realizations of $\boldsymbol{y} = (y_{11}, \ldots, y_{1m}, \ldots, y_{100,1}, \ldots, y_{100,m})'$ in the logistic regression model, i.e., $y_{ij} \sim$ indep. $B(1, p_{ij})$, where $p_{ij} = \mathrm{logit}^{-1}(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_0)$, $i = 1, \ldots, 100$, $j = 1, 2, 3$. Note that we used the working independence matrix for obtaining $\hat{\boldsymbol{\beta}}_f$ in this simulation. Tables 1 and 2 list the following characteristics.

(1) Prediction error of the best model in the $\mathrm{Risk_P}$/$\mathrm{Risk_Q}$ ($\mathrm{PEB_P}$/$\mathrm{PEB_Q}$): the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$

Table 1: Selection probability and prediction error for the Case 1

Criterion	k^*	1	2	**3**	4	5	$\mathrm{SP_C}$	$\mathrm{SP_M}$	$\mathrm{PEB_P}$	$\mathrm{PEB_Q}$
PMSEG	Exchangeable	0.95	0.08	11.02	2.62	1.10	70.49	**95.61**	**4.37**	**416.66**
	Autoregressive	2.61	0.48	**50.71**	9.77	6.92				
	Independence	0.22	0.05	12.72	0.67	0.08				
QIC	Exchangeable	4.36	0.01	2.60	3.46	3.60	**72.79**	55.50	10.54	418.09
	Autoregressive	38.36	0.26	**22.20**	6.11	5.86				
	Independence	1.51	0.00	3.09	5.17	3.41				

Table 2: Selection probability and prediction error for the Case 2

Criterion	k^*	1	2	**3**	4	5	$\mathrm{SP_C}$	$\mathrm{SP_M}$	$\mathrm{PEB_P}$	$\mathrm{PEB_Q}$
PMSEG	**Exchangeable**	13.97	0.89	**26.28**	8.06	6.76	55.96	**72.79**	**4.88**	**416.24**
	Autoregressive	7.67	1.77	9.10	1.97	1.20				
	Independence	2.06	0.85	17.00	1.73	0.69				
QIC	**Exchangeable**	29.70	1.14	**20.06**	7.46	5.64	**64.00**	57.93	5.53	416.61
	Autoregressive	2.55	0.19	1.70	1.17	1.14				
	Independence	7.53	0.96	16.64	2.56	1.56				

of the model selected by the criterion as the best model, which are respectively estimated as

$$\mathrm{PEB_P} = \frac{1}{10000}\sum_{\upsilon=1}^{10000} \mathrm{E}_z\left[\sum_{i=1}^{n}(\boldsymbol{z}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}))'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}))\right] - mn,$$

$$\mathrm{PEB_Q} = \frac{1}{10000}\sum_{\upsilon=1}^{10000} \mathrm{E}_z\left[-2\sum_{i=1}^{n}\sum_{j=1}^{m} Q(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}; z_{ij})\right].$$

(2) Selection probability: the frequency of the best model chosen by minimizing each criterion. In particular, the $\mathrm{SP_C}$/$\mathrm{SP_M}$ is the frequency that the working correlation matrix/mean structure of the selected model is correctly specified.

Here $\boldsymbol{z}_i$ is a future observation, and $\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}$ is the value of $\hat{\boldsymbol{\beta}}$ of the selected model at the υth iteration. In particular, both the $\mathrm{PEB_P}$ and $\mathrm{PEB_Q}$ are important properties because these are equivalent to the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$ of the best model selected by the criterion, respectively. We would like to note that the PMSEG selects the model which minimizes the $\mathrm{Risk_P}$, and the QIC selects the model which minimizes the $\mathrm{Risk_Q}$. Thus, the model selected by the PMSEG does not necessarily minimize the $\mathrm{Risk_Q}$ and the model selected by the QIC does not necessarily minimize the $\mathrm{Risk_P}$. In other words, in order to evaluate the goodness of the criterion, the $\mathrm{PEB_P}$ and $\mathrm{PEB_Q}$ are favorable indicators for the PMSEG and QIC, respectively.

From Tables 1 and 2, we can see that the value of the $\mathrm{PEB_P}$ from the model selected by the

PMSEG is smaller than that from the model selected by the QIC. This result is justified since the $\mathrm{PEB_P}$ is the favorable indicator for the PMSEG. However, surprisingly, although the $\mathrm{PEB_Q}$ is the favorable indicator for the QIC, the value of the $\mathrm{PEB_Q}$ from the model selected by the PMSEG is also smaller. This result means that the PMSEG is better than the QIC whether evaluating the goodness of the criterion by the $\mathrm{PEB_P}$ or $\mathrm{PEB_Q}$. Moreover, both the frequency of selecting the true model and $\mathrm{SP_M}$ of the PMSEG are larger than those of the QIC in two cases. On the other hand, the $\mathrm{SP_C}$ of the QIC is larger in two cases. Furthermore, by comparing Tables 1 and 2, we can see that the difference between the PMSEG and QIC is more salient when the correlation is large. We simulated several other models and obtained similar results.

Next, for the purpose of analyzing the GEE method, we consider the Mother's Stress and Child Morbidity (MSCM) data reported in Alexander and Markowitz (1986), who studied the relationship between maternal employment and pediatric health care utilization. The MSCM data contain the information of mothers and children in the study for 28 days, and there are 167 mothers and preschool children enrolled. In this analysis, we focus on the child illness for the first 9 days. The response variable is the child illness on the study day (1=yes, 0=no), and there were eight predictor variables: Married (1=married, 0=other), Employed (1=employed, 0=unemployed), Race (child race, 1=non-white, 0=white), Household (size of household, 1=more than 3 people, 0=other), Stress (today's mother's stress, 1=yes, 0=no), and additionally, St1, St2 and St3 are the mother's stress of one, two and three days before, respectively. This data have a few missing value, and we assume that the missingness mechanism is missing completely at random. We use complete 146 mothers and children data. We also assume that the response variable y_{ij} is distributed according to $B(1, p_{ij})$, $i = 1, \ldots, 146$, $j = 1, \ldots, 9$. For the link function, we prepare the logistic link function. For the working correlation matrix, we prepare four matrices: the working 1-dependent, autoregressive, exchangeable and independence matrices. In this analysis, we select the working correlation matrix and variables. Table 3 shows the selection probability of the model selected by minimizing the criterion and the estimated prediction error of the best model selected by the criterion. We divide the MSCM data into calibration data and validation data. The numbers of subjects in the calibration data and validation data were 136 and 10, respectively. The best subset of variables and working correlation matrices were selected by each criterion derived from the calibration data. The selection probabilities were obtained from only the calibration data. The estimated prediction errors were obtained as follows. Let $\boldsymbol{d}_t = (d_{1t}, \ldots, d_{146t})'$ be a 146-dimensional binary vector that contains 136 zeros and 10 ones at the tth iteration, $t = 1, \ldots, 1000$, i.e., $d_{it} = 0$ or 1 and $\sum_{i=1}^{146} d_{it} = 10$. In addition, we denote that $\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]}$ is a GEE estimator $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_t]}$ of the best model selected from the calibration data, where $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_t]}$ is the solution of the following equation:

$$\sum_{i=1}^{146}(1 - d_{it})\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_p.$$

Table 3: Selection probability and estimated prediction error with the MSCM data

Variables	PMSEG				QIC			
	Auto	Ex	1-dep	Ind	Auto	Ex	1-dep	Ind
Race, Household, Stress, St1	0	6	77	0	0	0	2	8
Race, Household, Stress, St3	0	11	0	0	0	0	0	533
Race, Household, Stress, St1, St2	0	2	1	0	0	0	0	0
Race, Household, Stress, St1, St3	0	135	739	10	25	0	2	429
Race, Household, Stress, St2, St3	0	3	0	0	0	0	0	0
Married, Race, Household, Stress, St1, St3	0	1	3	0	0	0	0	0
Employed, Race, Household, Stress, St1, St3	0	0	1	0	0	0	0	0
Race, Household, Stress, St1, St2, St3	0	8	1	0	0	0	0	1
Married, Race, Household, Stress, St1, St2, St3	0	2	0	0	0	0	0	0
$\widehat{\mathrm{PEB_P}}$	**230.11**				234.89			
$\widehat{\mathrm{PEB_Q}}$	**1018.41**				1020.72			

Finally, the estimated $\mathrm{PEB_P}$ and $\mathrm{PEB_Q}$ are given as

$$\widehat{\mathrm{PEB_P}} = 136 \times \left(\frac{1}{1000} \sum_{t=1}^{1000} \frac{1}{10} \mathcal{L}(\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]}, \hat{\boldsymbol{\beta}}_{f,t}, t) - 9 \right),$$

$$\widehat{\mathrm{PEB_Q}} = 136 \times \frac{1}{1000} \sum_{t=1}^{1000} \frac{1}{10} \left[-2 \sum_{i=1}^{146} \sum_{j=1}^{9} d_{it} \times Q(\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]}; y_{ij}) \right],$$

where $\mathcal{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, t)$ is defined as follows:

$$\mathcal{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, t) = \sum_{i=1}^{146} d_{it} \times (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1))' \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2) \boldsymbol{R}_0^{-1}(\boldsymbol{\beta}_2) \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2) (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)).$$

Note that we used the working independence matrix for obtaining $\hat{\boldsymbol{\beta}}_f$ from the calibration data in this study. Also note that we denote $\hat{\boldsymbol{\beta}}_{f,t}$ as the value of $\hat{\boldsymbol{\beta}}_f$ from the calibration data at the tth iteration. From Table 3, we can see that the model most selected by each criterion is different. However, both the $\widehat{\mathrm{PEB_P}}$ and $\widehat{\mathrm{PEB_Q}}$ of the PMSEG were smaller than those of the QIC. Hence, using the PMSEG is better than using the QIC for selecting models in this study.

Consequently, from Tables 1, 2 and 3, we recommend the use of the PMSEG rather than the QIC for selecting models in the GEE method.

5. CONCLUSION AND DISCUSSION

In the present paper, we proposed the PMSEG as a model selection criterion that reflects the correlation in the GEE method. The PMSEG is the simple criterion such as the AIC. Nowadays,

the GEE method is one of the mainstream of longitudinal analysis methods and many statistical softwares (e.g., SAS, R, etc) support the GEE method. For these reasons, it is important to propose a more useful criterion for analyzing longitudinal data using the GEE method.

We recall that, in deriving the PMSEG, we assume that the nuisance parameter $\boldsymbol{\alpha}$ is known. Actually, we often estimate $\boldsymbol{\alpha}$ because $\boldsymbol{\alpha}$ is unknown in many cases. In fact, we estimate $\boldsymbol{\alpha}$ in Section 4. However, Liang and Zeger (1986) showed that an estimator of $\boldsymbol{\alpha}$ is consistent under the standard assumption, and we confirmed that the estimation of $\boldsymbol{\alpha}$ dose not dramatically influence the performance of the PMSEG from some simulation results. Theoretical study of the influence of estimating $\boldsymbol{\alpha}$ to the PMSEG is left for the future work.

In all situations of the simulation results of Section 4, we showed that the PMSEG has better performance than the QIC for the variable selection, and the difference between the performances of the PMSEG and QIC is more salient when the correlation is large. Recall that the PMSEG reflects the correlation between responses, however, the QIC is not reflective. This is probably one of the reasons why the PMSEG has better performance than the QIC. In the study of the MSCM data, we also showed that the PMSEG is useful as same as the QIC. Nevertheless, computational costs of the PMSEG are lower than those of the QIC because the bias term of the PMSEG is 2 $\times$ the number of parameters. On the other hand, many previous studies including the QIC require the calculation of the bias term for each candidate model. Therefore, the PMSEG is useful and user friendly.

APPENDIX: Derivation of (3.9)

In this section, we calculate Bias2 + Bias4. Bias2 in (3.1) can be calculated as

$$\begin{aligned}
\text{Bias2} &= \mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)-\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\right] \\
&= \mathrm{E}_y\left[2\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right]+\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right],
\end{aligned}$$

and Bias4 in (3.1) can also be calculated as

$$\begin{aligned}
\text{Bias4} &= \mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\right] \\
&\quad-\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)\right] \\
&= -\mathrm{E}_y\left[2\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right] \\
&\quad-\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right].
\end{aligned}$$

Thus, we obtain Bias2 + Bias4 as follows:

$$
\begin{aligned}
&\text{Bias2} + \text{Bias4} \\
&= \mathrm{E}_y \left[2 \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \{ \boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) \boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f) \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) \} (\boldsymbol{\mu}_{i,0} - \hat{\boldsymbol{\mu}}_i) \right] \qquad \text{(A.1)} \\
&\quad + \mathrm{E}_y \left[\sum_{i=1}^{n} (\boldsymbol{\mu}_{i,0} - \hat{\boldsymbol{\mu}}_i)' \{ \boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) \boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f) \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) \} (\boldsymbol{\mu}_{i,0} - \hat{\boldsymbol{\mu}}_i) \right]. \qquad \text{(A.2)}
\end{aligned}
$$

In order to calculate (A.1) and (A.2), we perform the stochastic expansion of $\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$, $\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)$, $\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f)$ and $\hat{\boldsymbol{\beta}}_f$. Denote $\boldsymbol{D}_{*,i} = \boldsymbol{A}_i(\boldsymbol{\beta}_*)\boldsymbol{\Delta}_i(\boldsymbol{\beta}_*)\boldsymbol{X}_{*,i}$ and $\boldsymbol{D}_{*,i,0} = \boldsymbol{A}_{i,0}\boldsymbol{\Delta}_{i,0}\boldsymbol{X}_{*,i}$. Considering the same argument in section 2 and 3, $\hat{\boldsymbol{\beta}}_f$ and $\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f)$ can be expanded as

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{*,0} = \boldsymbol{b}_{f,0} + O_p(n^{-1}),\ \boldsymbol{b}_{f,0} = \boldsymbol{H}_{f,n,0}^{-1} \boldsymbol{s}_{f,n}(\boldsymbol{\beta}_{*,0}), \\
\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f) - \boldsymbol{\mu}_{i,0} = \boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0} + O_p(n^{-1}), \qquad \text{(A.3)}
\end{aligned}
$$

where $\boldsymbol{\beta}_{*,0}$ is the true value of $\boldsymbol{\beta}_*$, and $\boldsymbol{H}_{f,n,0}$ is defined as

$$
\boldsymbol{H}_{f,n,0} = \sum_{i=1}^{n} \boldsymbol{D}_{*,i,0}' \boldsymbol{A}_{i,0}^{-1/2} \bar{\boldsymbol{R}}_i^{-1}(\boldsymbol{\alpha}^*) \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{*,i,0}.
$$

Let $\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)$ denote the m-dimensional vector and the jth element of which is the (j,j)th element of $\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$. Note that $\mathrm{diag}(\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)) = \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$. By applying a Taylor expansion around $\hat{\boldsymbol{\beta}}_f = \boldsymbol{\beta}_{*,0}$, $\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)$ is expanded as

$$
\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f) = \boldsymbol{a}_{f,i}(\boldsymbol{\beta}_{*,0}) + \boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0} + O_p(n^{-1}),\ \boldsymbol{A}_{f,i,0}^{*} = \left. \frac{\partial}{\partial \boldsymbol{\beta}_*'} \boldsymbol{a}_{f,i}(\boldsymbol{\beta}_*) \right|_{\boldsymbol{\beta}_* = \boldsymbol{\beta}_{*,0}}.
$$

Hence, we obtain

$$
\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) = \mathrm{diag}(\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)) = \boldsymbol{A}_{i,0}^{-1/2} + \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0}) + O_p(n^{-1}). \qquad \text{(A.4)}
$$

Note that $\boldsymbol{b}_{f,0}, \boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0}, \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0}) = O_p(n^{-1/2})$. Moreover, substituting (A.3) and (A.4) into $\boldsymbol{R}_0(\hat{\boldsymbol{\beta}}_f)$ yields following:

$$
\begin{aligned}
\boldsymbol{R}_0(\hat{\boldsymbol{\beta}}_f) = &-\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_{i,0}^{-1/2} \{ \boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' + (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})(\boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0})' \} \boldsymbol{A}_{i,0}^{-1/2} \\
&+ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_{i,0}^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{A}_{i,0}^{-1/2} \\
&+ \frac{1}{n} \sum_{i=1}^{n} \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0})(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \boldsymbol{A}_{i,0}^{-1/2} \\
&+ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_{i,0}^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})' \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0}) + O_p(n^{-1}).
\end{aligned} \qquad \text{(A.5)}
$$

By the Lindeberg central limit theorem, the first term of (A.5) is $O_p(n^{-1})$. Thus, using this fact and (A.5), we obtain

$$\begin{aligned}\boldsymbol{R}_0^{-1/2}\boldsymbol{R}_0(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1/2} = \boldsymbol{I}_m - \boldsymbol{R}_0^{-1/2}\Bigg\{&\boldsymbol{R}_0 - \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1/2} + O_p(n^{-1}).\end{aligned}$$

Therefore, by calculating the inverse matrix of $\boldsymbol{R}_0^{-1/2}\boldsymbol{R}_0(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1/2}$, $\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)$ can be expanded as

$$\begin{aligned}\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f) = \boldsymbol{R}_0^{-1} + \boldsymbol{R}_0^{-1}\Bigg\{&\boldsymbol{R}_0 - \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1} + O_p(n^{-1}).\end{aligned} \tag{A.6}$$

Note that the second term of (A.6) is $O_p(n^{-1/2})$.

Next, we calculate (A.2). By using (A.4) and (A.6), we obtain

$$\begin{aligned}&\boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\\ =&-\mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2} - \boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})\\ &-\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\Bigg\{\boldsymbol{R}_0 - \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\boldsymbol{A}_{i,0}^{-1/2}\\ &-\frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})'\mathrm{diag}(\boldsymbol{A}_{f,i,0}^*\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2} + O_p(n^{-1}).\end{aligned} \tag{A.7}$$

Note that $\boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) = O_p(n^{-1/2})$ and $\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0} = O_p(n^{-1/2})$. Therefore, by substituting (3.4) and (A.7) into (A.2), we obtain

$$(\text{A.2}) = O(n^{-1}). \tag{A.8}$$

Recall that, in general, a moment of statistic can be expanded as power series in n^{-1}. Hence, the order of (A.8) is shown by $O(n^{-1})$, not $O(n^{-1/2})$.

Finally, we calculate (A.1). By substituting (3.4) and (A.7) into (A.1), we obtain

$$
\begin{aligned}
(\mathrm{A}.1) = & \mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\{\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}+\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
& -\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
& -\mathrm{E}_y\left[\frac{2}{n}\sum_{i,j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^{*}_{f,j,0}\boldsymbol{b}_{f,0})\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \qquad (\mathrm{A}.9) \\
& -\mathrm{E}_y\left[\frac{2}{n}\sum_{i,j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\mathrm{diag}(\boldsymbol{A}^{*}_{f,j,0}\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
& +\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{\Sigma}_{i,0}^{-1}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right]+O(n^{-1}).
\end{aligned}
$$

Here, $\boldsymbol{\kappa}_{i,0}=(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})$ and $\sum_{i,j}=\sum_{i=1}^{n}\sum_{j=1}^{n}$. Moreover, in order to simplify the calculation, we define the following notation:

$$
\sum_{i\neq j}=\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n},\quad \boldsymbol{b}_{f,i,0}=\boldsymbol{H}_{f,n,0}^{-1}\boldsymbol{D}'_{*,i,0}\boldsymbol{A}_{i,0}^{-1}\boldsymbol{\kappa}_{i,0}.
$$

Recall that $\mathrm{E}[\boldsymbol{\kappa}_{i,0}\otimes\boldsymbol{\kappa}'_{j,0}\boldsymbol{\kappa}_{k,0}]=\mathbf{0}_m$, (not $i=j=k$). Hence, the first term of (A.9) is as follows:

$$
\begin{aligned}
& \mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\{\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}+\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
= & \mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\{\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,i,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}+\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,i,0})\}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1i,0}\right] \\
= & O(n^{-1}). \qquad (\mathrm{A}.10)
\end{aligned}
$$

Similarly, since $\mathrm{E}_y[\boldsymbol{\kappa}'_{i,0}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{\kappa}_{k,0}]=0$ unless $i=k$, the second term of (A.9) can be calculated as

$$
\begin{aligned}
& -\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right] \\
= & -\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\frac{1}{n}\sum_{j=1,i\neq j}^{n}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1i,0}\right]+O(n^{-1}) \\
= & -\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{\Sigma}_{i,0}^{-1}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1i,0}\right]+O(n^{-1})=-2p+O(n^{-1}). \qquad (\mathrm{A}.11)
\end{aligned}
$$

Note that $-2p$ in (A.11) is obtained from (3.6). Furthermore, $\mathrm{E}_y[\boldsymbol{\kappa}'_{i,0}(\boldsymbol{\kappa}_{j,0}\otimes\boldsymbol{\kappa}'_{k,0})(\boldsymbol{\kappa}_{k,0}\otimes\boldsymbol{\kappa}_{l,0})]=0$ expect in the following cases:

$$
i=j=l \text{ or } i=j\neq k=l \text{ or } i=l\neq k=j \text{ or } j=l\neq k=i.
$$

Thus, the third term of (A.9) is given by

$$
\begin{aligned}
&-\mathrm{E}_y\left[\frac{2}{n}\sum_{i,j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^*_{f,j,0}\boldsymbol{b}_{f,0})\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right]\\
&=-\mathrm{E}_y\left[\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^*_{f,i,0}\boldsymbol{b}_{f,0})\boldsymbol{\kappa}_{i,0}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right]\\
&\quad-\mathrm{E}_y\left[\frac{2}{n}\sum_{i\neq j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^*_{f,j,0}\boldsymbol{b}_{f,i,0})\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1j,0}\right]\\
&\quad-\mathrm{E}_y\left[\frac{2}{n}\sum_{i\neq j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\mathrm{diag}(\boldsymbol{A}^*_{f,j,0}\boldsymbol{b}_{f,j,0})\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1i,0}\right]+O(n^{-1})\\
&=O(n^{-1}).
\end{aligned}
\tag{A.12}
$$

In the same manner as in the calculation of the third term of (A.9), the fourth term of (A.9) is calculated as

$$
-\mathrm{E}_y\left[\frac{2}{n}\sum_{i,j}\boldsymbol{\kappa}'_{i,0}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{j,0}^{-1/2}\boldsymbol{\kappa}_{j,0}\boldsymbol{\kappa}'_{j,0}\mathrm{diag}(\boldsymbol{A}^*_{f,j,0}\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right]=O(n^{-1}). \tag{A.13}
$$

Moreover, the fifth term of (A.9) is obtained from (3.6), as follows:

$$
\mathrm{E}_y\left[2\sum_{i=1}^{n}\boldsymbol{\kappa}'_{i,0}\boldsymbol{\Sigma}_{i,0}^{-1}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right]=2p. \tag{A.14}
$$

Substituting (A.10), (A.11), (A.12), (A.13) and (A.14) into (A.9), (A.1) is given by

$$
(\mathrm{A}.1)=O(n^{-1}). \tag{A.15}
$$

Consequently, from (A.8) and (A.15), we obtain $\mathrm{Bias2}+\mathrm{Bias4}=O(n^{-1})$.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Hirofumi Wakaki of Hiroshima University for their helpful comments and suggestions.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

Alexander, C. S. & Markowitz, R. (1986). Maternal employment and use of pediatric clinic services. *Medical care*, **24**, 134–147.

Cantoni, E., Flemming, J. M. & Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linear Models. *Biometrics*, **61**, 507–514.

Fitzmaurice, G. M. (1995). A Caveat Concerning independence Estimating Equations With Multivariate Binary Data. *Biometrics*, **51**, 309–317.

Gosho, M., Hamada, C. & Yoshimura, I. (2011). Modifications of QIC and CIC for selecting a Working Correlation Structure in the Generalized Estimating Equation Method. *Japanese Journal of Biometrics*, **32**, 1–12.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hin, L. Y. & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, **28**, 642–658.

Kullback, S. & Libler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. ser. A*, **135**, 370–384.

Nishii, R. (1984). Asymptotic Properties of Criteria for Selecting of Variables in Multiple Regression. *Ann. Statist.*, **12**, 758–765.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, **57**, 120–125.

Rao, C. R. & Wu, Y. (1989). A Strongly Consistent Procedure for Model Selection in a Regression Problem. *Biometrika*, **76**, 369–374.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss-Newton method. *Biometrika*, **61**, 439–447.

Xie, M. & Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.*, **31**, 310–347.