

# Sib-pair genetic longitudinal studies with missing not at random data

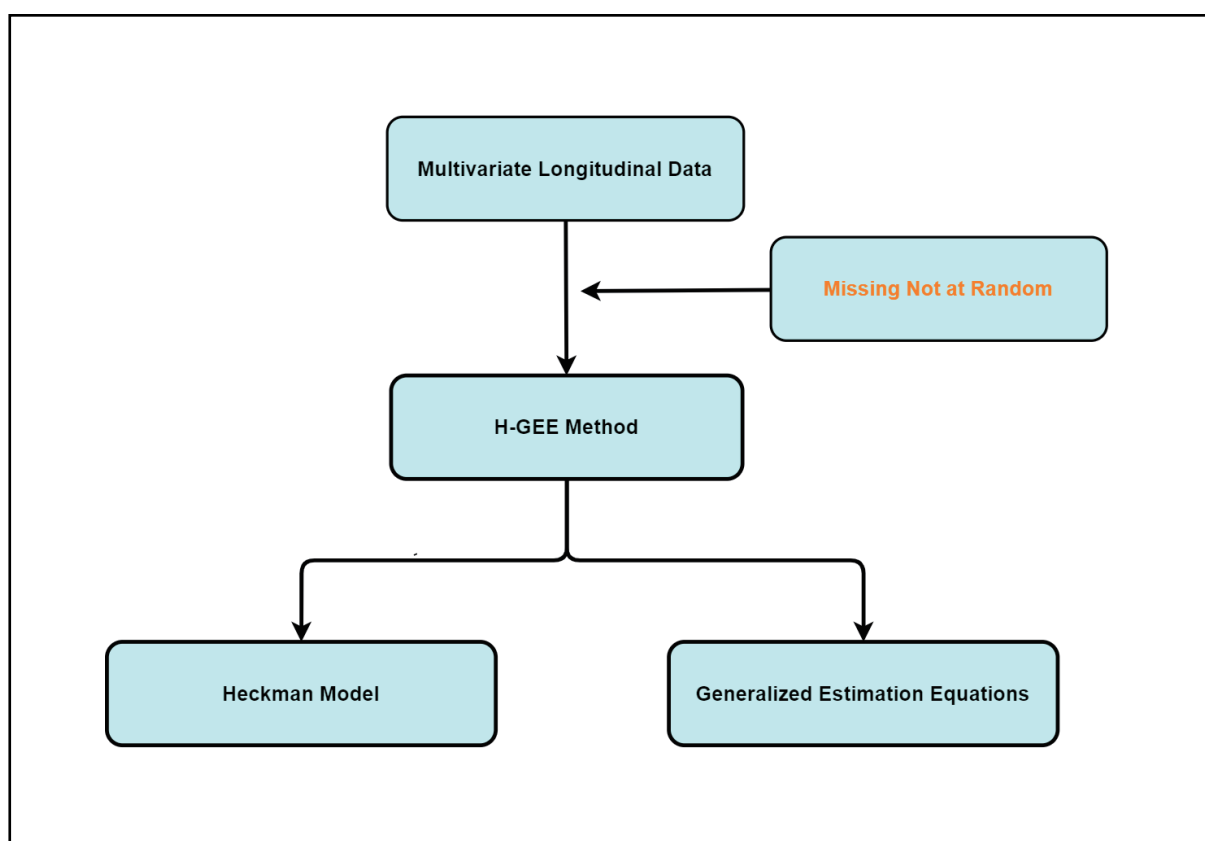
Siyu Jiang, and Hong Zhang 

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

 Correspondence: Hong Zhang, E-mail: [zhangh@ustc.edu.cn](mailto:zhangh@ustc.edu.cn)

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## Graphical abstract



*H-GEE method flowchart.*

## Public summary

- The challenge in longitudinal studies lies in effectively addressing missing-not-at-random (MNAR) data, which complicates data analysis.
- The proposed H-GEE method, combining the Heckman model and generalized estimating equations (GEE), aims to overcome MNAR challenges for robust data analysis.
- Extensive simulations validate H-GEE's effectiveness in handling MNAR data, highlighting its potential for advancing genetic and epidemiological research.

# Sib-pair genetic longitudinal studies with missing not at random data

Siyu Jiang, and Hong Zhang 

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

 Correspondence: Hong Zhang, E-mail: [zhangh@ustc.edu.cn](mailto:zhangh@ustc.edu.cn)

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Cite This: *JUSTC*, 2024, 54(12): 1203 (6pp)



Read Online

**Abstract:** In the interdisciplinary realm of statistics, genetics, and epidemiology, longitudinal sibling pair data offers a unique perspective for investigating complex diseases and traits, allowing the exploration of the dynamic processes of gene expression over time by controlling numerous confounding factors. Missing-not-at-random (MNAR) data are commonly used in such types of studies, but no statistical methods specifically tailored have been developed to handle MNAR data in complex longitudinal data in the literature. Here, we propose a new statistical method by jointly modeling longitudinal data from sib-pairs and MNAR data. Extensive simulations demonstrate the excellent finite sample properties of the proposed method.

**Keywords:** complex diseases; gene-time interaction; longitudinal data; missing-not-at-random data

**CLC number:** O212.1

**Document code:** A

**2020 Mathematics Subject Classification:** 62P10

## 1 Introduction

Longitudinal sibling pair data refers to genetic and phenotypic data collected from siblings within a family over time by controlling numerous confounding factors. This type of data is widely applied in fields such as genetic epidemiology and behavioral genetics, which can capture the changes in phenotype states and their genetic and environmental influences over time. Researchers leverage such data, especially those collected at multiple time points, to track phenotypic trajectories, uncover potential patterns or trends, and assess differences in responses to specific events or interventions among siblings. This study design proves particularly effective in exploring genes related to complex phenotypes such as hypertension, diabetes, heart disease, and height, as it allows researchers to control for many difficult-to-measure potential confounding factors, such as family background and other environmental influences<sup>[1]</sup>.

In summary, longitudinal sibling pair data provides a unique perspective for exploring the genetic and environmental factors underlying complex phenotypes. For instance, Friedlander et al.<sup>[2]</sup> proposed a method to adjust covariates in longitudinal sibling pair data and applied it to blood pressure data from the Framingham Heart Study. Guo et al.<sup>[3]</sup> employed stepwise discriminant analysis techniques to analyze chromosome data from the Framingham Heart Study, revealing interactions between genes and genes, as well as genes and the environment related to hypertension. Additionally, Keyes et al.<sup>[4]</sup> analyzed twin data to study the developmental trajectories of illicit drug use, while Silventoinen et al.<sup>[5]</sup> used twin data from multiple countries to investigate the heritability

of adult height. These studies underscore the importance of longitudinal sibling pair data in exploring how genetic and environmental factors collectively influence phenotypic development.

However, collecting such data poses challenges, with the most common issue being data missingness, especially non-random missingness. Discarding such data may result in information loss and distorted inference results. Therefore, specialized statistical methods are necessary for the analysis of longitudinal sibling pair data to elucidate the correlation between siblings, make full use of the observed information, and achieve reliable statistical inferences. Despite these challenges, longitudinal sibling pair data is crucial for understanding the genetic and environmental factors involved in the development of complex diseases and traits, providing a theoretical foundation for future personalized therapies and intervention strategies.

Most existing methods focus on the analysis of complete longitudinal sibling data, neglecting the issue of missing data. For complete longitudinal sibling data, methods for handling longitudinal data can be utilized, taking into account the dependency between siblings. There is extensive literature on the analysis of univariate longitudinal data<sup>[6–9]</sup>. Everitt and Dunn<sup>[10]</sup> provided a brief yet systematic method for analyzing continuous and categorical responses to longitudinal measurements. Dunlop<sup>[11]</sup> offered a clear description of regression methods for longitudinal data. Laird et al.<sup>[12]</sup> described formulas for longitudinal data random-effects models. Detailed discussions on maximum likelihood estimation methods based on the expectation-maximization (EM) algorithm were provided by Laird et al.<sup>[13]</sup> and others. Liang and Zeger<sup>[14]</sup> dis-

cussed the use of generalized estimating equations (GEE).

There is limited research in the existing literature regarding the study of longitudinal sibling data with missing data. Missing data mechanisms are generally categorized as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)<sup>[15]</sup>. For managing MCAR and MAR data, interpolation methods were commonly employed, as discussed in the works of Little et al.<sup>[16]</sup> and Sterne et al.<sup>[17]</sup>. Notably, Schafer and Graham<sup>[18]</sup> along with Graham<sup>[19]</sup> provided comprehensive descriptions of various existing methods. In this paper, we focus on MNAR data and propose a two-step method. In the first step, we estimate missing data based on the Heckman model<sup>[20]</sup>. In the second step, we propose to use generalized estimation equations to estimate unknown parameters. A bootstrap strategy is adopted to deal with the variability due to the estimated missing data. The validity of our proposed method was confirmed through extensive simulations.

## 2 Method

### 2.1 Notations and models

Let  $Y_{ijk}$  represent the continuous phenotype of the  $j$ -th sibling in the  $i$ -th family at time point  $t_{ijk}$  ( $k = 1 \cdots, n_{ij}$ ). This paper focuses on the association between phenotypes and genetic markers (single nucleotide polymorphisms, SNPs). Subsequently, the term “test locus” is used to refer to such genetic markers. Let  $A$  and  $a$  denote the major and minor alleles of the test locus, respectively. Let  $X_{ij}$  represent the genotype of the  $j$ -th sibling in the  $i$ -th family (0 for  $AA$ , 1 for  $Aa$ , and 2 for  $aa$ ).

We use the following linear model to characterize the influence of genotypes on phenotypes (outcome model):

$$Y_{ijk} = \alpha + \beta_x X_{ij} + \beta_t t_{ijk} + \gamma X_{ij} t_{ijk} + \eta_i + \eta_{ij} + \eta_k t_{ijk} + \varepsilon_{ijk}. \quad (1)$$

Here,  $\beta_f = (\alpha, \beta_x, \beta_t, \gamma)'$  is a fixed effect vector;  $\eta_i$ ,  $\eta_{ij}$ , and  $\eta_k$  are independent normally distributed random effects with mean zero and variances  $\sigma_a^2$ ,  $\sigma_b^2$ , and  $\sigma_c^2$ , respectively;  $\varepsilon_{ijk}$  is the random error following the normal distribution with mean 0 and variance  $\sigma^2$ . Note that the three random effects reflect variability due to family, individual, and time.

The MNAR mechanism of the phenotype is modeled through the following Heckman model (selection model)<sup>[20]</sup>:

$$\begin{cases} z_{ijk}^* = \beta_{s0} + \beta_{s1} X_{ij} + \beta_{s2} t_{ijk} + \beta_{s3} X_{ij} t_{ijk} + u_{ijk}, \\ z_{ijk} = I(z_{ijk}^* > 0), \end{cases} \quad (2)$$

where  $z_{ijk}^*$  is an underlying continuous variable and  $z_{ijk}$  is the missingness indicator ( $z_{ijk} = 1$  if  $Y_{ijk}$  is observed and  $z_{ijk} = 0$  if  $Y_{ijk}$  is missing);  $I$  is the indicator function ( $I(A) = 1$  if  $A$  occurs and  $I(A) = 0$  otherwise);  $u_{ijk}$  is the error term following the standard normal distribution;  $\beta_s = (\beta_{s0}, \beta_{s1}, \beta_{s2}, \beta_{s3})$  is a vector of fixed effects.

The joint model comprises the outcome model (1) and the selection model (2). Let  $\rho$  denote the correlation coefficient between  $u_{ijk}$  and  $\varepsilon_{ijk}$ . Let  $Y_{ijk}^*$  denote the observed outcome ( $Y_{ijk}^* = Y_{ijk} z_{ijk}$ ), then all observed data are  $(Y_{ijk}^*, t_{ijk}, X_{ij}, z_{ijk})$ ,  $i = 1, \dots, n$ ;  $j = 1, 2$ ;  $k = 1, \dots, n_{ij}$ . All unknown parameters

are  $\beta_s, \beta_f, \sigma_a^2, \sigma_b^2, \sigma_c^2, \sigma^2$ .

The joint model (1)-(2) extends the Heckman model from several aspects. First, the phenotype model utilizes three random effects to characterize variability at the levels of family, individual, and time. Second, the joint model does not require  $\varepsilon_{ijk}$  to be normally distributed. These two distinctions significantly broaden the applicability of the Heckman model. In the following subsection, we propose to use a generalized estimating equation (GEE) to accommodate the non-normal distribution assumption.

### 2.2 A two-step estimation method

It follows from (1)-(2) that the conditional expectation of the observed outcome is

$$\begin{aligned} E(Y_{ijk}^* | Y_{ijk}^* \neq 0) &= E(Y_{ijk} | z_{ijk}^* > 0) = \\ &= E(Y_{ijk} | u_{ijk} > -\beta_{s0} - \beta_{s1} X_{ij} - \beta_{s2} t_{ijk} - \beta_{s3} X_{ij} t_{ijk}) = \\ &= \alpha + \beta_x X_{ij} + \beta_t t_{ijk} + \gamma X_{ij} t_{ijk} + \\ &= E(\varepsilon_{ijk} | u_{ijk} > -\beta_{s0} - \beta_{s1} X_{ij} - \beta_{s2} t_{ijk} - \beta_{s3} X_{ij} t_{ijk}) = \\ &= \beta_f' X_{ijk}^s + \rho \sigma \lambda_{ijk}, \end{aligned}$$

where

$$\lambda_{ijk} = \frac{\phi(\beta_s' X_{ijk}^s)}{\Phi(\beta_s' X_{ijk}^s)},$$

$X_{ijk}^s = (1, X_{ij}, t_{ijk}, X_{ij} t_{ijk})$ ,  $\phi$  is the probability density function of the normal distribution, and  $\Phi$  is the cumulative distribution function of the standard normal distribution.

We employ a two-step method to separately estimate the parameter vectors  $\beta_s$  and  $\beta_f$ . In the first step, we can estimate  $\beta_s$  by the maximum likelihood estimator, denoted by  $\hat{\beta}_s$ , based on the selection model (2). In the second step, we first

replace  $\lambda_{ijk}$  by  $\hat{\lambda}_{ijk} = \frac{\phi(X_{ijk}^s \hat{\beta}_s)}{\Phi(X_{ijk}^s \hat{\beta}_s)}$ , yielding the linear model

$$Y_{ijk}^* = \alpha + \beta_x X_{ij} + \beta_t t_{ijk} + \gamma X_{ij} t_{ijk} + \rho \sigma \hat{\lambda}_{ijk} + e_{ijk},$$

where  $e_{ijk}$  is a random error with mean 0. Then, we employ the GEE method to estimate  $\beta_f$  and  $\rho \sigma$ . This two-step method does not need to specify the distributional form of  $Y_{ijk}$ , thus it is robust to a certain degree. In what follows, we provide the detailed parameter estimation procedure and discuss the corresponding asymptotic properties.

The first step of the estimation algorithm is to estimate the parameter vector  $\beta_s$  by the maximum likelihood method, which is based on the log-likelihood function for the observed data  $(z_{ijk}, X_{ij}, t_{ijk})$ ,  $i = 1, \dots, n$ ;  $j = 1, 2$ ;  $k = 1, \dots, n_{ij}$ :

$$\begin{aligned} \ln L(\beta_s) &= \sum_{i=1}^n \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} \ln f(z_{ijk} | X_{ij}, t_{ijk}, \beta_s) = \\ &= \sum_{i=1}^n \sum_{j=1}^2 \left[ \sum_{k: z_{ijk}=0} \ln \Phi(-\beta_s' X_{ijk}^s) + \sum_{k: z_{ijk}=1} \ln \Phi(\beta_s' X_{ijk}^s) \right]. \end{aligned}$$

The maximum likelihood estimator  $\hat{\beta}_s$  of  $\beta_s$  is defined as the

solution to the following score equation:

$$g(\beta_s) := \frac{\partial \ln L(\beta_s)}{\partial \beta_s} = \sum_{i=1}^n \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} \frac{\partial \ln f(z_{ijk} | X_{ijk}, t_{ijk}, \beta_s)}{\partial \beta_s} = 0.$$

According to standard likelihood theory, under certain regularity conditions,  $\hat{\beta}_s$  is consistent for  $\beta_s$  and asymptotically normal:

$$n^{1/2}(\hat{\beta}_s - \beta_s) \xrightarrow{d} N(0, [-H(\beta_s)]^{-1}),$$

$$\text{with } H(\beta_s) = E \left[ \frac{1}{n} \frac{\partial^2 \ln L(\beta_s)}{\partial \beta_s \partial \beta_s'} \right].$$

In the second step of the algorithm, a GEE is constructed based on the following linear model:

$$Y_{ijk}^* = \alpha + \beta_x X_{ij} + \beta_t t_{ijk} + \gamma X_{ij} t_{ijk} + \rho \sigma \hat{\lambda}_{ijk} + e_{ijk}, \quad (3)$$

$$i = 1, \dots, n; j = 1, 2; k = 1, \dots, n_{ij}.$$

Specifically, first define

$$\beta_f = (\alpha, \beta_x, \beta_t, \gamma, \rho \sigma)'$$

$$Y_i^* = (Y_{i11}^*, \dots, Y_{i1n_{ij}}^*, Y_{i21}^*, \dots, Y_{i2n_{ij}}^*)'$$

$$e_i = (e_{i11}, \dots, e_{i1n_{ij}}, e_{i21}, \dots, e_{i2n_{ij}})'$$

$$Z_i = (Z_{i11}, \dots, Z_{i1n_{ij}}, Z_{i21}, \dots, Z_{i2n_{ij}})'$$

$$\mu_i = Z_i \beta_f,$$

where

$$Z_{i11} = (1, X_{ij}, t_{ijk}, X_{ij} t_{ijk}, \hat{\lambda}_{ijk})'.$$

Model (3) can be expressed in the following form:

$$Y_i^* = Z_i \beta_f + e_i, \quad i = 1, \dots, n. \quad (4)$$

The corresponding GEE is

$$S(\beta) := \sum_{i=1}^n \Sigma_{Y_i}^{-1} (Y_i - \mu_i) = 0,$$

where  $\Sigma_{Y_i}$  is a working covariance matrix of  $Y_i$ . In practice, we can specify the correlation matrix  $R_i(\xi)$  of  $Y_i$ , leading to the working covariance matrix

$$\tilde{\Sigma}_{Y_i} = \sigma^2 R_i(\xi).$$

Some commonly used correlation coefficient structures include independence structure, exchangeable correlation structure, one-dependent structure, autocorrelation, and unstructured correlation. Regardless of the choice of correlation coefficient structure, under suitable regularity conditions (eg., the variance of  $Y_{ijk}$ , denoted by  $\sigma^2$ , is independent of  $i, j$ , and  $k$ ), the corresponding GEE estimator  $\hat{\beta}$  is consistent for  $\beta$  and asymptotically normal:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, G_\beta), \text{ with } G_\beta = \sigma^2 \left[ E \left\{ \frac{1}{n} \frac{\partial S(\beta)}{\partial \beta'} \right\} \right]^{-1}.$$

Estimates of  $\beta$  and  $G_\beta$  can be obtained by applying the existing GEE procedures implemented existing programs. Note that in model (3),  $\hat{\lambda}_{ijk}$  shares a common random variable  $\hat{\beta}$ , resulting in correlation between  $Y_{ijk}^*$ . While the existing

standard GEE procedures can produce consistent parameter estimates, they may underestimate the covariance matrix  $G_\beta$  due to the correlation between  $Y_{ijk}^*$ . To address this issue, we propose to adopt the bootstrap method, which can provide a consistent estimate of  $G_\beta$ .

### 3 A simulation study

We conducted a simulation study to evaluate the performance of our Heckman-model-based method, referred to as H-GEE. We also considered two additional methods C-GLM and D-GLM. C-GLM directly fitted the generalized linear model to complete data. On the other hand, D-GLM fitted the generalized linear model to the observed data by discarding missing data. Both C-GLM and D-GLM were implemented by the R function *lmer* in the R package *lme4*. The R functions *nlminb* and *lmer* were adopted to in the first step and the second step of H-GEE, respectively. It is noteworthy that C-GLM is not applicable in practical scenarios since it assumes that missing data are observable. C-GLM was simply used as a gold standard method.

The simulated data were generated as follows. First, under the assumption of Hardy-Weinberg equilibrium, genotypes of a single nucleotide polymorphism with minor allele frequency 0.25 were independently generated for 100N individuals. Genotypes of  $N$  couples were then drawn from these genotypes by assuming random mating. Genotypes ( $X_{i1}, X_{i2}$ ) of two offspring were generated for each of  $N$  couples based on Mendel's laws of inheritance. Then, set the model parameters  $\beta_f, \sigma_a^2, \sigma_b^2, \sigma_c^2, \beta_s, \rho$ , and  $\sigma^2$  in models (1)-(2), such that the proportion of error variance to total variance was approximately 80% by setting  $\sigma^2 = 2(\beta_0^2 + \beta_1^2 + \gamma^2 + \sigma_a^2 + \sigma_b^2 + \sigma_c^2)$ . Next, the age  $t_{ijk}$  was randomly drawn from the uniform distribution  $U(45, 75)$ . Finally,  $Y_{ijk}$  and  $z_{ijk}^*$  were generated according to models (1)-(2). The number of time points,  $k$ , followed the uniform distribution in  $\{2, 4, 5\}$ :  $P(k=3) = P(k=4) = P(k=5) = 1/3$ . The parameters were fixed as  $\sigma_a^2 = 0.5, \sigma_b^2 = 0.25, \sigma_c^2 = 0.05, \beta_s = (\beta_{s0}, 1, 1, 1), \rho = 0.75$ , and  $\beta_f = (5, 2, 1, 0)$ .

In all simulations, the default sample size, interaction effect, and missing rate were fixed at  $n = 100, \gamma = 0$ , and 0.2 (corresponding to  $\beta_{s0} = 1.5$ ), unless specified otherwise. We examined the performance of the considered methods under various effect sizes, sample sizes, missing rates, and error distributions. We also evaluated the validity of the bootstrap method for estimating the standard errors of parameter estimates.

First, we examined the estimation and test results of various genotype-time interaction effects for the three considered methods. We considered four different values of interaction effects ( $\gamma = 0, 0.25, 0.5$ , and 1). The sample size was fixed at  $n = 100$ . The simulation results based on 500 replicates are presented in Table 1. Both C-GLM and H-GEE produced very virtually unbiased estimates for the fixed effects  $\beta_f$ , while D-GLM had considerable estimation biases, especially when  $\gamma$  was large. These results demonstrated that simply deleting incomplete data could lead to very biased estimation results in the presence of MNAR data, while the proposed method H-GEE could effectively adjust such biases. Due to

missing data, H-GEE was less efficient than C-GLM. For example, the standard errors of the  $\gamma$  estimates by H-GEE were around 1.2 times those of C-GLM. The estimated standard errors of H-GEE were considerably smaller than the empirical ones (results not shown), indicating the necessity of employing the bootstrap method. The performance of the bootstrap method will be examined at the end of this section.

Second, we evaluated the performance of the considered methods under different sample sizes ( $n = 50, 100, 200, 300$ , and  $400$ ). As shown in Table 2, both C-GLM and H-GEE were virtually unbiased under various sample sizes. As

expected, the standard errors (SE) decreased when the sample size increased. The estimation bias of D-GLM tended to be stable with the increasing sample size, indicating systematic estimation biases of D-GLM.

Third, we examined the performance of the considered methods under different missing rates (0.05, 0.1, 0.2, 0.3, and 0.6), with the corresponding values of  $\beta_{s0}$  being 3, 2, 1.5, 0, and  $-1.5$ , respectively. As shown in Table 3, the estimation bias of D-GLM dramatically increased as the missing rate increased, while the estimation bias of H-GEE remained to be very minor.

**Table 1.** Estimation biases (standard errors) with various interaction effects.

| $\gamma^a$ | Method <sup>b</sup> | $\hat{\alpha}_0^c$ | $\hat{\beta}_0^d$ | $\hat{\beta}_1^e$ | $\hat{\gamma}^f$ | $\hat{\rho}\sigma^g$ |
|------------|---------------------|--------------------|-------------------|-------------------|------------------|----------------------|
| 0.00       | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | -0.00 (0.17)     | —                    |
|            | D-GLM               | 1.15 (0.18)        | -0.39 (0.18)      | -0.56 (0.17)      | -0.25 (0.19)     | —                    |
|            | H-GEE               | 0.02 (0.46)        | -0.00 (0.22)      | -0.02 (0.27)      | -0.00 (0.21)     | -0.04 (1.31)         |
| 0.25       | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | —                    |
|            | D-GLM               | 1.15 (0.18)        | -0.39 (0.18)      | -0.56 (0.18)      | -0.25 (0.19)     | —                    |
|            | H-GEE               | 0.02 (0.46)        | 0.00 (0.22)       | -0.02 (0.27)      | 0.00 (0.21)      | -0.03 (1.32)         |
| 0.50       | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | —                    |
|            | D-GLM               | 1.17 (0.18)        | -0.40 (0.18)      | -0.57 (0.18)      | -0.25 (0.20)     | —                    |
|            | H-GEE               | 0.02 (0.47)        | 0.00 (0.23)       | -0.02 (0.27)      | 0.00 (0.21)      | 0.03 (1.33)          |
| 1.00       | C-GLM               | 0.00 (0.18)        | 0.00 (0.17)       | -0.01 (0.17)      | 0.00 (0.18)      | —                    |
|            | D-GLM               | 1.24 (0.19)        | -0.42 (0.19)      | -0.60 (0.19)      | -0.27 (0.21)     | —                    |
|            | H-GEE               | 0.02 (0.50)        | 0.00 (0.24)       | -0.02 (0.29)      | 0.00 (0.22)      | 0.25 (1.41)          |

<sup>a</sup>The true value of  $\gamma$ . <sup>b</sup>C-GLM, the method fitting GLM with complete data; D-GLM, the method fitting GLM dropping incomplete data; H-GEE, the proposed method based on the Heckman model. <sup>c</sup>The true value of  $\alpha_0$  was 5. <sup>d</sup>The true value of  $\beta_0$  was 2. <sup>e</sup>The true value of  $\beta_1$  was 1. <sup>f</sup>The true value of  $\rho\sigma$  was 3.554.

**Table 2.** Estimation biases (standard errors) with different sample sizes.

| $N^a$ | Method <sup>b</sup> | $\hat{\alpha}_0^c$ | $\hat{\beta}_0^d$ | $\hat{\beta}_1^e$ | $\hat{\gamma}^f$ | $\hat{\rho}\sigma^g$ |
|-------|---------------------|--------------------|-------------------|-------------------|------------------|----------------------|
| 50    | C-GLM               | -0.01 (0.23)       | 0.01 (0.23)       | 0.00 (0.23)       | -0.01 (0.22)     | —                    |
|       | D-GLM               | 1.13 (0.24)        | -0.40 (0.26)      | -0.55 (0.26)      | -0.25 (0.28)     | —                    |
|       | H-GEE               | 0.04 (0.59)        | -0.01 (0.36)      | -0.03 (0.30)      | -0.02 (1.66)     | -0.16 (0.32)         |
| 100   | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | —                    |
|       | D-GLM               | 5.15 (0.18)        | -0.39 (0.18)      | -0.56 (0.17)      | -0.25 (0.19)     | —                    |
|       | H-GEE               | 0.02 (0.46)        | 0.00 (0.22)       | -0.02 (0.27)      | 0.00 (0.21)      | -0.05 (1.31)         |
| 200   | C-GLM               | 0.00 (0.12)        | 0.00 (0.11)       | 0.01 (0.12)       | 0.00 (0.11)      | —                    |
|       | D-GLM               | 5.14 (0.12)        | -0.38 (0.13)      | -0.54 (0.12)      | -0.24 (0.14)     | —                    |
|       | H-GEE               | 0.02 (0.33)        | 0.00 (0.16)       | 0.00 (0.19)       | 0.00 (0.14)      | -0.04 (0.91)         |
| 300   | C-GLM               | 0.00 (0.09)        | 0.00 (0.10)       | -0.01 (0.10)      | 0.00 (0.09)      | —                    |
|       | D-GLM               | 1.15 (0.10)        | -0.37 (0.11)      | -0.56 (0.10)      | -0.24 (0.11)     | —                    |
|       | H-GEE               | 0.01 (0.24)        | 0.00 (0.12)       | -0.01 (0.15)      | 0.00 (0.12)      | -0.01 (0.67)         |
| 400   | C-GLM               | 0.00 (0.08)        | -0.01 (0.08)      | 0.00 (0.08)       | 0.00 (0.08)      | —                    |
|       | D-GLM               | 1.15 (0.09)        | -0.39 (0.09)      | -0.55 (0.09)      | -0.25 (0.10)     | —                    |
|       | H-GEE               | 0.01 (0.23)        | -0.01 (0.11)      | 0.00 (0.14)       | 0.00 (0.11)      | -0.03 (0.64)         |

<sup>a</sup>The number of families. <sup>b</sup>C-GLM, the method fitting GLM with complete data; D-GLM, the method fitting GLM dropping incomplete data; H-GEE, the proposed method based on the Heckman model. <sup>c</sup>The true value of  $\alpha_0$  was 5. <sup>d</sup>The true value of  $\beta_0$  was 2. <sup>e</sup>The true value of  $\beta_1$  was 1. <sup>f</sup>The true value of  $\gamma$  was 0. <sup>g</sup>The true value of  $\rho\sigma$  was 3.554.

Fourth, we examined the performance of the considered methods under several typical non-normal error distributions. Two non-normal distributions were considered: the  $t$ -distribution with 5 degrees of freedom and the skewed normal distribution with skewness parameter 5. As shown in Table 4, C-GLM and H-GEE were again virtually unbiased, while D-GLM had systematic estimation biases.

Finally, we examined the bootstrap method for estimating standard errors in H-GEE. Let the version of H-GEE incorporating the bootstrap method be denoted by H-GEE-B. The simulation results corresponding to Table 1 are presented in Table 5. Evidently, the estimated standard errors were close to the empirical ones for all parameters, and the corresponding coverage probabilities were close to the nominal level 0.95. This demonstrated the validity of the bootstrap method.

## 4 Discussion

With the progress of bioinformatics and epidemiology, the understanding of complex diseases and traits continues to deepen. The longitudinal sibling pair data provides researchers with the opportunity to explore in detail the patterns of genetic changes over time, offering a crucial means for interpreting the connection between genes and complex diseases and traits. Its profound value lies in its ability to help researchers delve into the interaction between genes and time, bringing new research directions to the fields of genetics and epidemiology.

To further explore this dynamic relationship, this study adopts a longitudinal research strategy and collects sibling pair data at multiple time points. However, this research strategy

**Table 3.** Estimation biases (standard errors) with different missing rates.

| Rate <sup>a</sup> | Method <sup>b</sup> | $\hat{\alpha}_0^c$ | $\hat{\beta}_0^d$ | $\hat{\beta}_1^e$ | $\hat{\gamma}^f$ | $\hat{\rho}\sigma^g$ |
|-------------------|---------------------|--------------------|-------------------|-------------------|------------------|----------------------|
| 0.05              | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | -                    |
|                   | D-GLM               | 0.32 (0.17)        | -0.03 (0.16)      | -0.24 (0.16)      | -0.14 (0.17)     | -                    |
|                   | H-GEE               | 0.02 (0.26)        | 0.00 (0.16)       | -0.02 (0.22)      | -0.01 (0.20)     | -0.09 (2.26)         |
| 0.10              | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00(0.17)       | -                    |
|                   | D-GLM               | 0.66 (0.17)        | -0.16 (0.17)      | -0.40 (0.17)      | -0.20 (0.18)     | -                    |
|                   | H-GEE               | 0.02 (0.34)        | 0.00 (0.18)       | -0.02 (0.24)      | -0.01 (0.20)     | -0.05 (1.61)         |
| 0.20              | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | -                    |
|                   | D-GLM               | 5.15 (0.18)        | -0.39 (0.18)      | -0.56 (0.17)      | -0.25 (0.19)     | -                    |
|                   | H-GEE               | 0.02 (0.46)        | 0.00 (0.22)       | -0.02 (0.27)      | 0.00 (0.21)      | -0.05 (1.31)         |
| 0.30              | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | -                    |
|                   | D-GLM               | 1.61 (0.18)        | -0.61 (0.18)      | -0.68 (0.18)      | -0.29 (0.21)     | -                    |
|                   | H-GEE               | 0.02 (0.56)        | 0.00 (0.27)       | -0.02 (0.28)      | 0.00(0.22)       | -0.03 (1.14)         |
| 0.60              | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | -                    |
|                   | D-GLM               | 3.40 (0.25)        | -1.41 (0.25)      | -1.14 (0.28)      | -0.42 (0.33)     | -                    |
|                   | H-GEE               | 0.02 (0.90)        | 0.00 (0.43)       | -0.02 (0.38)      | 0.00 (0.30)      | -0.03 (0.87)         |

<sup>a</sup>The proportion of missing data. <sup>b</sup>C-GLM, the method fitting GLM with complete data; D-GLM, the method fitting GLM dropping incomplete data; H-GEE, the proposed method based on the Heckman model. <sup>c</sup>The true value of  $\alpha_0$  was 5. <sup>d</sup>The true value of  $\beta_0$  was 2. <sup>e</sup>The true value of  $\beta_1$  was 1. <sup>f</sup>The true value of  $\gamma$  was 0. <sup>g</sup>The true value of  $\rho\sigma$  was 3.555.

**Table 4.** Estimation biases (standard errors) with different error distributions.

| Error <sup>a</sup> | Method <sup>b</sup> | $\hat{\alpha}_0^c$ | $\hat{\beta}_0^d$ | $\hat{\beta}_1^e$ | $\hat{\gamma}^f$ | $\hat{\rho}\sigma^g$ |
|--------------------|---------------------|--------------------|-------------------|-------------------|------------------|----------------------|
| N(0,1)             | C-GLM               | 0.00 (0.17)        | 0.00 (0.16)       | -0.01 (0.16)      | 0.00 (0.17)      | -                    |
|                    | D-GLM               | 1.15 (0.18)        | -0.39 (0.18)      | -0.56 (0.17)      | -0.25 (0.19)     | -                    |
|                    | H-GEE               | 0.02 (0.46)        | 0.00 (0.22)       | -0.02 (0.27)      | 0.00 (0.21)      | -0.05 (1.31)         |
| $t_5$              | C-GLM               | 0.00 (0.10)        | 0.01 (0.09)       | 0.00 (0.09)       | 0.00 (0.09)      | -                    |
|                    | D-GLM               | 0.49 (0.10)        | -0.15 (0.10)      | -0.23 (0.10)      | -0.11 (0.10)     | -                    |
|                    | H-GEE               | 0.01 (0.25)        | 0.01 (0.12)       | 0.00 (0.15)       | 0.00 (0.12)      | -2.07 (0.69)         |
| SN(5)              | C-GLM               | 0.01 (0.09)        | 0.00 (0.07)       | 0.00 (0.06)       | 0.00 (0.06)      | -                    |
|                    | D-GLM               | 0.01 (0.09)        | 0.00 (0.08)       | -0.01 (0.07)      | 0.00 (0.07)      | -                    |
|                    | H-GEE               | 0.02 (0.20)        | 0.00 (0.10)       | -0.02 (0.11)      | 0.00 (0.09)      | -3.61 (0.56)         |

<sup>a</sup>The error distribution: N(0,1), the standard normal distribution;  $t_5$ , the  $t$ -distribution with 5 degrees of freedom; SN(5), the skewed normal distribution with the location parameter of 5. <sup>b</sup>C-GLM, the method fitting GLM with complete data; D-GLM, the method fitting GLM dropping incomplete data; H-GEE, the proposed method based on the Heckman model. <sup>c</sup>The true value of  $\alpha_0$  was 5. <sup>d</sup>The true value of  $\beta_0$  was 2. <sup>e</sup>The true value of  $\beta_1$  was 1. <sup>f</sup>The true value of  $\gamma$  was 0. <sup>g</sup>The true value of  $\rho\sigma$  was 3.555.

**Table 5.** Simulation results using the bootstrap method.

| $\gamma^a$ | $\hat{\alpha}_0$ |                  |                 | $\hat{\beta}_0$ |      |      | $\hat{\beta}_1$ |      |      | $\hat{\gamma}$ |      |      |
|------------|------------------|------------------|-----------------|-----------------|------|------|-----------------|------|------|----------------|------|------|
|            | SE <sup>b</sup>  | SEE <sup>c</sup> | CP <sup>d</sup> | SE              | SEE  | Bias | SE              | SEE  | CP   | SE             | SEE  | CP   |
| 0.00       | 0.45             | 0.45             | 0.93            | 0.23            | 0.22 | 0.94 | 0.26            | 0.27 | 0.95 | 0.21           | 0.21 | 0.94 |
| 0.25       | 0.46             | 0.45             | 0.93            | 0.23            | 0.22 | 0.94 | 0.26            | 0.27 | 0.95 | 0.21           | 0.21 | 0.93 |
| 0.50       | 0.46             | 0.46             | 0.93            | 0.23            | 0.23 | 0.94 | 0.27            | 0.28 | 0.95 | 0.21           | 0.21 | 0.93 |
| 1.00       | 0.48             | 0.47             | 0.93            | 0.24            | 0.23 | 0.94 | 0.27            | 0.28 | 0.95 | 0.22           | 0.22 | 0.94 |

<sup>a</sup>True value of  $\gamma$ ; <sup>b</sup>empirical standard error; <sup>c</sup>mean estimated standard error; <sup>d</sup>coverage probability of 95% confidence interval.

comes with challenges, such as MNAR outcomes. There is a lack of literature on methods for handling MNAR outcomes in longitudinal studies. To address this issue, this paper proposes a novel method H-GEE, which combines the Heckman model and generalized estimating equations (GEE). H-GEE enables effective statistical analysis of longitudinal sibling pair data in the presence of MNAR outcomes. Extensive simulation studies confirmed the feasibility of this method under limited samples. H-GEE exhibited a reasonably good performance in our simulation studies.

H-GEE has its limitations. Currently, it can only handle continuous response variables, and it deserves to be extended to other types of response variables such as binary response variables, count response variables, and survival time response variables. Additionally, although the bootstrap method is valid in estimating standard errors, it is time consuming. It deserves further investigation to derive the asymptotic distribution of estimators, so that an explicit standard error estimator can be obtained.

## Acknowledgements

The authors thank the students in Prof. Hong Zhang's laboratory for their support. This work was supported by the National Natural Science Foundation of China (12171451).

## Conflict of interest

The authors declare that they have no conflict of interest.

## Biographies

**Siyu Jiang** is currently a graduate student under the tutelage of Prof. Hong Zhang at the University of Science and Technology of China. His research mainly focuses on statistical genetics.

**Hong Zhang** is a Full Professor at the University of Science and Technology of China (USTC). He received his Bachelor's degree in Mathematics and Ph.D. degree in Statistics from USTC in 1997 and 2003, respectively. His research mainly focuses on statistical genetics, causal inference, and machine learning.

## References

- [1] Rutter M. Nature, nurture, and development: From evangelism through science toward policy and practice. *Child Development*, 2002, 73 (1): 1–21.

- [2] Friedlander Y, Talmud P J, Edwards K L, et al. Sib-pair linkage analysis of longitudinal changes in lipoprotein risk factors and lipase genes in women twins. *Journal of Lipid Research*, 2000, 41 (8): 1302–1309.
- [3] Guo Z, Li X, Rao S Q, et al. Multivariate sib-pair linkage analysis of longitudinal phenotypes by three step-wise analysis approaches. *BMC Genetics*, 2003, 4 (1): 1–7.
- [4] Keyes M A, Malone S M, Elkins I J, et al. The enrichment study of the Minnesota twin family study: increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Research and Human Genetics*, 2009, 12 (5): 489–501.
- [5] Silventoinen K, Sammalisto S, Perola M, et al. Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Research*, 2003, 6 (5): 399–408.
- [6] Hand D M, Crowder M J. Practical Longitudinal Data Analysis. New York: Chapman & Hall/CRC, 1996.
- [7] Verbeke G. Linear mixed models for longitudinal data. In: Linear Mixed Models in Practice. New York: Springer, 1997.
- [8] Diggle P, Heagerty P, Liang K Y, et al. Analysis of Longitudinal Data. New York: Oxford University Press, 2002.
- [9] Fitzmaurice G M, Laird N M, Ware J H. Applied Longitudinal Analysis. Hoboken, USA: Wiley, 2012.
- [10] Everitt B S, Dunn G. Applied Multivariate Data Analysis. Second Edition. Chichester, UK: Wiley, 2001.
- [11] Dunlop D D. Regression for longitudinal data: a bridge from least squares regression. *The American Statistician*, 1994, 48 (4): 299–303.
- [12] Laird N M, Ware J H. Random-effects models for longitudinal data. *Biometrics*, 1982, 38 (4): 963–974.
- [13] Laird N M, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 1987, 82 (397): 97–105.
- [14] Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986, 73 (1): 13–22.
- [15] Little R J, Rubin D B. Statistical Analysis with Missing Data. Third Edition. Hoboken, USA: Wiley, 2019.
- [16] Little R J. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 1993, 88 (421): 125–134.
- [17] Sterne J A, Carlin J B, Royston P, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 2009, 338: b2393.
- [18] Schafer J L, Graham J W. Missing data: Our view of the state of the art. *Psychological Methods*, 2002, 7 (2): 147–177.
- [19] Graham J W. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 2009, 60: 549–576.
- [20] Heckman J J. Sample selection bias as a specification error. *The Econometric Society*, 1979, 47 (1): 153–161.