JGEE: Joint Modeling Mean and Overdispersion for Correlated Count Data

Xiecheng Gu [1], Liming Xiang[1*] and Andy H. Lee[2]

[1] *Division of Mathematical Sciences, School of Physical and Mathematical Sciences*

Nanyang Technological University, 21 Nanyang Link, Singapore 637371

[2] *Department of Epidemiology and Biostatistics, School of Public Health*

Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia

Abstract

Clustered/logitudinal count data are encountered in many medical and health studies. When overdispersion is present as often happens in practice, the existing methods for analysis of overdispersed correlated count data are limited to treating the overdispersion as a nuisance parameter and assuming it to be the same across all clusters. This may hamper the ability to understand covariate effects on the overdispersion. In this paper, we propose a GEE-based modeling approach for joint mean and overdispersion analysis. Besides a regression on the mean level, this approach particularly allows a regression on the dispersion level to account for varying impacts of covariates over clusters on extra variation in count data. A computational algorithm is carried out for estimation of parameters in the model. We show satisfactory performance of the proposed method via simulations and illustrate its practical utility by analyzing data from a study of diarrhea incidence in children.

Key words: GEE; Longitudinal data; Negative Binomial Distribution; Overdispersion; Poisson regression.

* Corresponding author. *E-mail:* LMXiang@ntu.edu.sg.

1. Introduction

The response of interest often consists of count data. An appealing factor that influences the estimation efficiency in repeated-measure data analysis is the overdispersion. It always occurs in the count data as a common phenomenon that the variation expected in the given specified model is less than computed from the actual data set. Lots of applications are addressed in the books written by Fleiss (1981).

When overdisersion exhibits in count data, i.e., the variability of the data exceeds that predicted by a Poisson model.

Overdispersed Poisson-type data was well investigated by Breslow (1984) and Lawless (1987). To account for overdispersion in modeling, we usually apply the negative binomial and some other mixed Poisson models to fit such extra variation that occurs among the observations. It is found that the dispersion often leads to a loss of efficiency in estimating regression parameters and a failure of the implicit mean-variance structure. As discussed by the paper of Liang & Zeger (1986), the standard GEE method is also able to handle the longitudinal data with overdispersion, but often ignored and treated it as nuisance parameter in a hierarchical model analysis. The ratio that measures the difference between the statistical model and empirical observations is allowed to be denoted by $\phi > 1$ and can be estimated by the Pearson residuals. For instance, we have a set of count data modeled by the Poisson distribution whose mean and variance are both equal to μ but the sample variance is recognized to exceed the nominal variance μ, denoted by $\phi\mu$. Cox (1983) found that when ϕ was modest, the high asymptotic relative efficiency was still remained in the overdispersed model. However his method does not perform well when ϕ is sufficiently large. Based on the common one-parameter exponential family, the sufficient condition to develop overdispersion was derived by Gelfand & Dalal (1990). Although the ways that overdispersion raised are still in discussion, it may be mainly caused by the unobserved variation in the response probabilities and the unknown interaction between responses. This motivates us to solve such kind of problem in the reverse way, that is, involve dispersion coefficients into the model to estimate the dispersion effects to reduce such negative influence on consistency of estimators.

Regarding the discrete responses, Jowaheer & Sutradhar (2002) made an effort on analyzing the longitudinal count data with overdispersion under the negative binomial model by GEE and developed a moment method to estimate the dispersion when the dispersion parameter was supposed to be constant over all subjects.

Recently many works have been done in estimating parameters on the covariance or variance level. Thall & Vail (1990) introduced several covariance models for overdispersed count data and discussed the large-sample properties of the estimators. Prentice & Zhao (1991) developed and extended the joint estimating equations for mean and covariance parameters based on quadratic exponential family. They assessed the performance of the estimating equations under several working covariance matrices with independent or gaussian structures. Pourahmadi (1999) used Cholesky decomposition to construct joint mean-covariance model, followed by Ye & Pan (2006) who introduced such decomposition method into covariance estimation within the framework of GEE. They applied simple regression model both on the mean and variance levels. Then Leng et al. (2010) continued their work by introducing the data-driven approach based on semiparametric models with same normal distribution assumption. However, the model specification they used in GEE is limited to the continuous distributions, especially for gaussian case and may not be applicable for discrete data.

In the context of likelihood approaches, the variance component model is difficult to specify as well as the multivariate analysis on discrete data. By contrast, GEE methods take full advantage of the moment structures of the discrete distribution from the exponential family. The estimating equations on variance level need only the second and fourth moment of responses. Fortunately, even the correlation between each subject on that level is unknown, the result is still asymptotic consistency.

Motivating example

Recently Binns et al.(2007) conducted a randomized trial with children of 1 to 3 years old attending childcare centres in Perth, Australia. The aim of this study was to evaluate the effect on diarrhoea incidence of a milk product in children clustered in 29 child-care centers through Perth metropolitan area. Initially, the centres were randomly selected from centres located in the southern city area of Perth. Children in the participating centres then were allocated into the intervention and control groups at random. The control product was identical to the intervention product except that it contained no prebiotics or probiotics. The indicator of diarrhoea, the number of days that children had four or more stools was of major interest of researchers. The number of such days was treated as a discrete response variable. Other appealing characteristics of participating children listed in the data sample include the average consumption of CUPDAY milk, days of family members in sick, besides the basic demographic factors like age and sex.

We found that the CUPDAY data collected from attended childcare centers were skewed and strongly overdispersed. The responses implying the diarrhoea rate of children did not follow Poisson distribution. Although in GEE, and due to the basic properties of exponential family, the constant dispersion parameter is set to measure the deviation that the observed variance is greater than the model variance, in fact it has no influence on estimating regression parameter as it is always be canceled out of the estimating equations. Therefore we doubt that the common count data structure, e.g, Poisson with equal mean and variance, is no longer suitable for this kind of data. We also expect that by jointly estimating parameters on both mean and dispersion levels, the yielding explanation becomes more reasonable and convinced.

The following contents of this paper are organized as below. Our proposed model is specified in Section . In Section we construct our joint estimating equations in which the regression is extended to the dispersion level and derive the new iterative algorithm. The asymptotic inference can be found in the next Section 0.2. A simulation study is designed and conducted for evaluating the performance of the proposed method in Section 0.2, followed by an illustrative example illustrated in Section 0.2. The conclusion and discussion are summarized in the last section.

2. Model specification

Independently over all subjects, the longitudinal repeated responses are defined as $\{Y_{ij},\ 1 \le i \le N,\ 1 \le j \le T\}$ with correlation within each subject i. One of popular approaches to model overdispersion is the negative binomial distribution. Suppose that the longitudinal responses Y_{ij} follows negative binomial distribution given by

$$f_{\mu_{ij},\phi_{ij}}(y_{ij}) = \frac{\Gamma(y_{ij}+\phi_{ij}^{-1})}{\Gamma(\phi_{ij}^{-1})\Gamma(y_{ij}+1)} \left(\frac{1}{1+\phi_{ij}\mu_{ij}}\right)^{\phi_{ij}^{-1}} \left(\frac{\phi_{ij}\mu_{ij}}{1+\phi_{ij}\mu_{ij}}\right)^{y_{ij}} \tag{0.1}$$

and

$$E(Y_{ij}) = \mu_{ij}, \quad Var(Y_{ij}) = \mu_{ij} + \phi_{ij}\mu_{ij}^2, \tag{0.2}$$

where $\phi_{ij} > 0$. In more general case, we can select some distributions from double exponential families, defined as

$$f_{\mu,\theta,n}(y) = c(\mu,\theta,n)\theta^{1/2} g_{\mu,n}(y)^{\theta} g_{y,n}(y)^{1-\theta}[dG_n(y)] \tag{0.3}$$

where μ denotes the mean of Y, θ the dispersion parameter and n the sample size. The first term in the right of (0.3), $c(\mu,\theta,n)$ make $\int_{-\infty}^{+\infty} \tilde{f}_{\mu,\theta,n}(y)dG_n(y) = 1$, while $g_{\mu,n}(y)$ refers to the existing density function from the one-parameter exponential distribution family and $g_{y,n}(y)$ has the same form as $g_{\mu,n}(y)$ in which μ is replaced by the value of y. The last term $G_n(y)$ is the carrier measure for the exponential family so that $\Pr(J) = \int_J g_{\mu,n}(y)dG_n(y)$ for measurable set J.

Remark 0.1. Besides negative binomial distribution, there are some alternative distributions which accommodate overdispersion. For example, the repeated counts Y_{ij}, can also be assumed to have the double Poisson distribution alternatively, whose probability mass function (pmf), from (0.3) is given by

$$f_{\mu_{ij},\theta_{ij}}(y_{ij}) = \theta_{ij}^{1/2} e^{-\theta_{ij}\mu_{ij}} \left(\frac{e^{-y_{ij}} y_{ij}^{y_{ij}}}{y_{ij}!} \right) \left(\frac{e\mu_{ij}}{y_{ij}} \right)^{\theta_{ij} y_{ij}} . \tag{0.4}$$

By the Fact 2 in Efron (1986), we have $E(Y_{ij}) \doteq \mu_{ij}$ and $Var(Y_{ij}) \doteq \mu_{ij}/\theta_{ij}$, where " $\doteq$ " represents approximation. As the value of Y_{ij} grows larger, this approximation is much closer to the true mean and variance.

Another alternative way to deal with overdispersed data is to use the form of Generalized Poisson introduced by Consul & Jain (1973), in which, under same notations as (0.2), the first two moments are defined as $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \mu_{ij}^2 \phi_{ij}^2$, respectively.

Based on the idea first provided by Liang & Zeger (1986), the estimating equations method drops the restrictive distribution assumption and only requires to specify the first and second moments that have several optional structures, especially in the overdispersion model. Therefore, our model formulation can be specified as follows:

- Let $Y_{ij} \sim NB(\mu_{ij}, \phi_{ij})$ with $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \mu_{ij} + \phi_{ij}\mu_{ij}^2$, where $\{Y_{ij},\ 1 \le i \le N,\ 1 \le j \le T\}$ are independent over all i and dependent with correlation matrix R_1 which is identical for all j. Assume R_1 is determined by the sole correlation parameter ρ_1, so that $R_1 = R_1(\rho_1)$. In each subject i, $Y_i = (Y_{i1}, \ldots, Y_{iT})$, $i = 1, 2, 3, \ldots, I$, represents T-vector of response variables, $i = 1, 2, 3, \ldots, I$.

- The linear predictor on the mean: $\zeta_{ij}^{\mu}(\beta) = x_{ij}^T \beta$
  and on the dispersion: $\zeta_{ij}^{\phi}(\gamma) = w_{ij}^T \gamma$,

where $\mathbf{X}_i$ and $\mathbf{W}_i$ are the $T \times p$ and $T \times q$ design matrices with jth row vectors x_{ij}^T and w_{ij}^T respectively. $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\gamma = (\gamma_1, \ldots, \gamma_q)^T$ denote the vector of the regression parameters on the mean and dispersion level, respectively.

- The corresponding logarithm link functions gives:

$$\mu_{ij} = E(Y_{ij}|\beta) = h(\zeta_{ij}^{\mu}(\beta)) = \exp(x_{ij}^T\beta) \tag{0.5}$$

$$\phi_{ij} = h(\zeta_{ij}^{\phi}) = \exp(w_{ij}^T\gamma) \tag{0.6}$$

Since the aim of this paper is to investigate the overdispersion among the count data, the log transformation of $\phi_{ij} > 0$ ensures the exceeded variance of our primary interest.

3. Estimation

0.1 Joint structure

As a marginal modeling approach in longitudinal data, GEE is much easier to compute the estimates of coefficients of interest, without concerning the issues related to the implicit likelihood for discrete responses. Recall the basic form of GEE for mean

$$S(\hat{\beta}) = \sum_{i=1}^{N} D_{i1}^T \tilde{V}_{i1}^{-1} r_{i1} = 0_p, \tag{0.7}$$

where r_{i1} is the residual $r_{i1} = y_i - \mu_i$ and $\mu_i = h(x_i^T\beta)$. $D_{i1} = \partial\mu_i/\partial\beta^T$ denotes the 1st derivative of mean μ_i with respect to β. In the equation (0.7), the true variance V_{i1} with entries $cov(Y_{ij}, Y_{ij^\star})$, for $j, j^\star = 1, \ldots, T$, is replaced by $\tilde{V}_{i1} = A_i^{1/2} R_1(\rho_1) A_i^{1/2}$ where $A_i = diag[Var(Y_i)]$ denotes a $T \times T$ marginal variance matrix, and $R_1(\rho_1)$ is a working correlation matrix with ρ_1 being the corresponding correlation paramter.

Therefore, the regression parameter β can be estimated by solving the estimating equations (0.7), through a numerical methods, like Newton-Raphson algorithm that clearly leads to clear asymptotic inference.

A few studies have been done to extend the estimating equations (0.7) to the variance or covariance structure, for example, Prentice & Zhao (1991) and Jowaheer & Sutradhar (2002). Both of them required a constant dispersion ϕ and the normality assumption which may be unrealistic or too restrictive to represent data in practice. We relax these assumptions and develop estimating equations at the mean and the dispersion levels simultaneously. Our model features separate regression functions for the mean and overdispersion of count responses, thus allowing for exploring covariate effects on mean as well as on extra variation of Y.

By the moment generating functions for Y_{ij}, the higher moments of Y_{ij} can be obtained by

$$E(Y_{ij}^2) = \mu_{ij} + (\phi_{ij} + 1)\mu_{ij}^2,$$

$$Var(Y_{ij}^2) = \mu_{ij} + (6 + 7\phi_{ij})\mu_{ij}^2 + (4 + 14\phi_{ij} + 12\phi_{ij}^2)\mu_{ij}^3 + (4\phi_{ij} + 10\phi_{ij}^2 + 6\phi_{ij}^3)\mu_{ij}^4$$

. Detailed derivations of above results are given in the Appendix.

In the rest of this paper, we focus on modeling overdispersed count data under the negative binomial distribution only. The modeling approaches using alternative double Poisson structure or other distributions can also be derived in a similar way.

Remark 0.2. From Fact 4 of Efron (1986), as $Y_{ij} \sim DoublePoisson(\mu_{ij}, \phi_{ij})$, the first two moments of Y_{ij} are obtained by assuming $Y_{ij} \sim \phi_{ij} Z_{ij}$ where $Z_{ij} \sim Poi(\mu_{ij}/\phi_{ij})$, while the higher moments can also be calculated by the moment generating functions, so we have $E(Z_{ij}^2) = (\mu_{ij}/\phi_{ij})^2 + \mu_{ij}/\phi_{ij}$, $E(Y_{ij}^2) = \mu_{ij}^2 + \mu_{ij}\phi_{ij}$, $E(Z_{ij}^4) = \mu_{ij}/\phi_{ij} + 7(\mu_{ij}/\phi_{ij})^2 + 6(\mu_{ij}/\phi_{ij})^3 + (\mu_{ij}/\phi_{ij})^4$ and $E(Y_{ij}^4) = \mu_{ij}\phi_{ij}^3 + 7\mu_{ij}^2\phi_{ij}^2 + 6\mu_{ij}^3\phi_{ij} + \mu_{ij}^4$. Then $Var(Y_{ij}^2) = \mu_{ij}\phi_{ij}^3 + 6\mu_{ij}^2\phi_{ij}^2 + 4\mu_{ij}^3\phi_{ij}$.

Therefore, the regression parameter γ on the dispersion level, can be obtained by solving the following estimating equations:

$$S(\gamma) = \sum_{i=1}^{N} D_{i2}^T \tilde{V}_{i2}^{-1} r_{i2} = 0_q \qquad (0.8)$$

where $r_{i2} = y_i^2 - m_i$ denotes the residual vector on the variance level and m_i is the 2nd order moments of Y_i. Recall that $m_{ij} = E(Y_{ij}^2) = \mu_{ij} + (\phi_{ij} + 1)\mu_{ij}^2$ and $D_{i2} = \partial m_i / \partial \gamma^T$. Analogous to (0.7), $\tilde{V}_{i2}$ is assumed to be the working covariance matrix instead of the true structure, defined as $H_i^{1/2} R_2(\rho_2) H_i^{1/2}$ where $H_i = diag[Var(Y_i^2)]$. R_2 and ρ_2 have the same meanings as R_1 and ρ_1 but are at the dispersion level.

The estimates $\hat{\beta}$ and $\hat{\gamma}$ can be obtained by solving equations (0.7) and (0.8). Note that with known β and ρ_2, m_i and $\tilde{V}_{i2}$ are considered as the functions of γ only and actually, β and ρ_2 can be substituted by $\hat{\beta}$ and $\hat{\rho}_2$.

As in the standard GEE, the working correlation $R_1(\rho_1)$ is taken an *exchangeable* structure without loss of generality, so is $R_2(\rho_2)$. That is,

$$R_i(\rho_i) = \begin{pmatrix} 1 & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \cdots & \rho_i \\ \vdots & & \ddots & \vdots \\ \rho_i & \rho_i & \cdots & 1 \end{pmatrix}, \quad i = 1, 2 \tag{0.9}$$

where the off-diagonal elements are all equal to ρ_i, $\rho_i \in [-1, 1]$. Other typical specifications of working correlation could be considered. The alternative choice of R_i may be $AR(1)$ in which $\rho_{tt^\star} = \rho_i^{|j-j^\star|}$, or independence where there is no correlation within each subject.

As mentioned in Liang & Zeger (1986) and modified by Jowaheer & Sutradhar (2002), the correlation parameter ρ_i can be separately estimated by the moment method. Hence using the same way as in aforementioned works, we obtained the estimators of ρ_i under *exchangeable* working correlation via the formulas

$$\hat{\rho}_1 = \frac{\sum_{i=1}^{N} \sum_{j>j^\star} \hat{r}_{ij}\hat{r}_{ij^\star}}{\frac{N}{2}(T-1)T - p}, \tag{0.10}$$

$$\hat{\rho}_2 = \frac{\sum_{i=1}^{N} \sum_{j>j^\star} \hat{s}_{ij}\hat{s}_{ij^\star}}{\frac{N}{2}(T-1)T - p - q}, \tag{0.11}$$

where the standardized residual

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}\hat{\phi}_{ij}}}$$

and

$$\hat{s}_{ij} = \frac{y_{ij}^2 - m_{ij}(\hat{\beta}, \hat{\gamma})}{\sqrt{\hat{\mu}_{ij} + (6 + 7\hat{\phi}_{ij})\hat{\mu}_{ij}^2 + (4 + 14\hat{\phi}_{ij} + 12\hat{\phi}_{ij}^2)\hat{\mu}_{ij}^3 + (4\hat{\phi}_{ij} + 10\hat{\phi}_{ij}^2 + 6\hat{\phi}_{ij}^3)\hat{\mu}_{ij}^4}}$$

with $\hat{\mu}_{ij} = \mu_{ij}(\hat{\beta})$ and $\hat{\phi}_{ij} = \phi_{ij}(\hat{\gamma})$. In fact $\hat{\rho}_1$ is a consistent estimator of ρ_1, due to the moment estimator properties.

It can also be evaluated simultaneously with regression parameter β by GEE2 proposed by Prentice & Zhao (1991).

Now, with unknown ρ_i replaced by $\hat{\rho}_i$, β and γ can be estimated simultaneously. Combining (0.7) with (0.8), we define the following joint estimating equations,

$$S(\hat{\beta}, \hat{\gamma}) = \sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} r_i = 0_{p+q} \tag{0.12}$$

where

$$D_i = \begin{pmatrix} D_{i1} & 0 \\ D_{i21} & D_{i2} \end{pmatrix}_{2T\times(p+q)} = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta'} & 0 \\ \frac{\partial m_i}{\partial \beta'} & \frac{\partial m_i}{\partial \gamma'} \end{pmatrix}_{2T\times(p+q)}$$

and $\tilde{V}_i = \begin{pmatrix} \tilde{V}_{i1} & cov(Y_i, Y_i^2) \\ cov(Y_i^2, Y_i) & \tilde{V}_{i2} \end{pmatrix}_{2T\times 2T}$, $r_i = \begin{pmatrix} y_i - \mu_i \\ y_i^2 - m_i \end{pmatrix}_{2T\times 1}$.

The equations above require the correct specification of the first two moments μ_i and m_i but yield the consistent estimator of mean parameter β even when the covariance is incorrectly constructed (Prentice & Zhao 1991). This robust property of this estimating equations motivates us to strengthen the estimation of β by choosing the appropriate covariance specification. That is, the off-diagonal part of $\tilde{V}_i$, denoted by $cov(Y_i, Y_i^2)$ is substituted by zero matrix. The same reason could also be applied on $D_{i21} = \partial m_i/\partial\beta'$ in D_i so that D_i becomes diagonal matrix.

0.2 Algorithm

Before implementing Newton-Raphson method to solve (0.12), we define the information matrix $\Sigma(\beta, \gamma) = \sum_{i=1}^N D_i^T \tilde{V}_i^{-1} D_i$. Then β, γ could be estimated by the following iterative process:

1. Giving the initial values $\hat{\beta}^{(0)}$ and $\hat{\gamma}^{(0)}$, compute $\hat{\rho}_1$ and $\hat{\rho}_2$ by (0.10)-(0.11) above, denoted by $\hat{\rho}_1^{(0)}$ and $\hat{\rho}_2^{(0)}$.

2. Calculate Σ with D_i, $\tilde{V}_i$ and r_i defined in (0.12) at $\hat{\beta}^{(0)}$, $\hat{\gamma}^{(0)}$, $\hat{\rho}_1^{(0)}$ and $\hat{\rho}_2^{(0)}$.

3. Let $\hat{\beta}^{(j)}$ and $\hat{\gamma}^{(j)}$ denote the values of the estimator in jth iteration step, then $\hat{\beta}^{(j+1)}$ and $\hat{\gamma}^{(j+1)}$ are given by

$$\begin{pmatrix} \hat{\beta}^{(j+1)} \\ \hat{\gamma}^{(j+1)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{(j)} \\ \hat{\gamma}^{(j)} \end{pmatrix} + \left(\sum_{i=1}^N D_i^T \tilde{V}_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^N D_i^T \tilde{V}_i^{-1} r_i \right), \tag{0.13}$$

   where (0.13) the second term on the right side is evaluated at $\hat{\beta}^{(j)}$ and $\hat{\gamma}^{(j)}$.

4. Update $\hat{\rho}_1$ and $\hat{\rho}_2$ with $\hat{\beta}^{(j+1)}$ and $\hat{\gamma}^{(j+1)}$, and denote them as $\hat{\rho}_1^{(j+1)}$ and $\hat{\rho}_2^{(j+1)}$.

5. If some convergence criteria is attained, say

$$\left\| \begin{pmatrix} \hat{\beta}^{(j+1)} \\ \hat{\gamma}^{(j+1)} \end{pmatrix} - \begin{pmatrix} \hat{\beta}^{(j)} \\ \hat{\gamma}^{(j)} \end{pmatrix} \right\| \left\| \begin{pmatrix} \hat{\beta}^{(j)} \\ \hat{\gamma}^{(j)} \end{pmatrix} \right\|^{-1} < \epsilon$$

where ϵ is sufficiently small, the iteration stops then $\hat{\beta} = \hat{\beta}^{(j+1)}$, $\hat{\gamma} = \hat{\gamma}^{(j+1)}$, $\hat{\rho}_1 = \hat{\rho}_1^{(j+1)}$ and $\hat{\rho}_2 = \hat{\rho}_2^{(j+1)}$. Otherwise, go back to step (3) until the result converges.

4.Asymptotic inference

We discuss the performance of both estimators $\hat{\beta}$ and $\hat{\gamma}$ in this section when N becomes sufficiently large. For convenience we just use θ to represent $(p+q)$ dimensional vector (β, γ). For simplicity, the following items Σ, D_i, r_i and $\tilde{V}_i$ denotes the true value, respectively. Following the outstanding work of Fahrmeir & Kaufmann (1985), Liang & Zeger (1986) and later Prentice & Zhao (1991), we conclude two important asymptotic properties.

Comparing (0.12) with (0.7), it is easy to learn that our joint model mimics the structure of the estimating equations at the mean level only. Hence we concentrate on the joint equations (0.12), and assume that the following mild regularity conditions hold:

(A1) $S(\theta)$ and $\Sigma(\theta)$ are continuously thrice differentiable.

(A2) $\Sigma(\theta)$ is positive definite with probability one as N tends to ∞, and $\|\Sigma(\theta) - \Sigma_0(\theta)\| \to 0$ uniformly where Σ_0 is positive function for all θ.

(A3) The derivatives of $\Sigma_0(\theta)$ are $O_p(1)$ for all θ.

(A4) $S_i(\theta) = D_i^T \tilde{V}_i^{-1} r_i$ has finite covariance for all θ.

Theorem 0.3. *($\sqrt{N}$ consistency) Under mild regularity conditions, we have that the solution of (0.12), $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta_0\| = O_p(\sqrt{1/N}). \tag{0.14}$$

θ_0 denote the true mean and dispersion parameters respectively.

Theorem 0.4. *(Asymptotic normality) Under mild regularity conditions, as $N \longrightarrow \infty$, $N^{1/2}(\hat{\theta} - \theta_0)$ asymptotically converges to normal distribution with mean 0_{p+q} and covariance matrix estimated by*

$$N\Sigma^{-1} \left(\sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} r_i r_i^T \tilde{V}_i^{-1} D_i \right) \Sigma^{-1}, \tag{0.15}$$

which is a consistent estimator of the true asymptotic covariance given by

$$N\Sigma^{-1} \left(\sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} Cov(Y_i, Y_i^2) \tilde{V}_i^{-1} D_i \right) \Sigma^{-1}, \tag{0.16}$$

where usually $Cov(Y_i, Y_i^2)$ is unknown.

Recall that under the GEE approach the consistency of estimator does not depend on the choice of working correlation structure, which may however has negative effect on the estimation efficiency.

To ensure the consistency and asymptotic normality of the estimator, the random vector Y_i, $i = 1, \ldots, N$ is supposed to have finite 2nd moment. Thus for the dispersion part of our model, it is required that the 4th moment of Y_i must exist and better to derive its explicit form. We put the proofs of Theorem 0.3 and Theorem 0.4 in Appendix.

5. Simulation study

In the simulation, we assumed that the longitudinal responses $\{Y_{ij},\ 1 \le i \le N,\ 1 \le j \le T\}$ follows the negative binomial distribution (0.1). The correlation between y_{ij} and $y_{ij^\star}$, $j \neq j^\star$, is approximated by preassumed "working" structure but independence is applied between different subject i. To simplify our problem but without loss of generality, we considered there are two coefficients on the mean level and the dispersion is only determined by two covariates in our regression model. Following the construction process derived by McKenzie (1986), we designed matrices $\mathbf{X}_i$ and $\mathbf{W}_i$ and generated the corresponding response Y_i as follows. First of all, for each subject i, the covariates $x_{ij} = (x_{ij0}, x_{ij1})^T$ on the mean were assumed to be

$$x_{ij0} = \begin{cases} -1 & : \quad (i = 1, 2, \ldots, N/4) \\ 0 & : \quad (i = N/4+1, \ldots, 3N/4) \\ 1 & : \quad (i = 3N/4+1, \ldots, N) \end{cases} , \quad x_{ij1} \sim \text{Uniform}(0,1).$$

Similarly letting $w_{ij} = (w_{ij0}, w_{ij1})^T$, for $j = 1, \ldots, T$, we again set $w_{ij1} \equiv 1$ and generated $w_{ij1} \equiv b_i$ where b_i was drawn from uniform interval $(0,1)$ as well. According to (0.5), we had $\mu_{ij} = \exp(x_{ij}^T\beta)$, $\beta = (\beta_0, \beta_1)^T$, and $\phi_{ij} = \exp(w_{ij}^T\gamma) + 1$, $\gamma = (\gamma_0, \gamma_1)^T$.

For comparison, we considered number of subject N to be 80, 160 and 240, and repeated times T to be 8 and 20, respectively. We set the true values of the regression parameters to be $\beta = (1, 1)$ and $\gamma = (1, 1)$. The *exchangeable* correlation matrix was assumed in the estimating equations with correlation parameter $\rho_1 = 0.7, 0.5, 0.3$ and 0.1, ranging from a high degree to a low degree of correlation.

Given the values of $\mathbf{X}_i$ and $\mathbf{W}_i$ with the true regression coefficients and correlation assumption, we thereafter generated the corresponding response Y_i using the method developed by McKenzie (1986). For $j = 1, \ldots, T$, the following relation among the observations of Y_i must be satisfied:

$$y_{ij} = \varphi_{ij} y_{i0} + e_{ij}.$$

Details of the generation of Y_i above are provided in Appendix. Applying the Newton-Raphson algorithm (0.13), we can estimate β and γ consistently according to (0.15). Meanwhile ρ_1 and ρ_2 were estimated via (0.10) and (0.11).

To evaluate the proposed inference procedure, we repeated 1000 simulations and we summarized resulting estimates and their standard errors in Table 1. It is evident that the estimates of two correlation parameters generally perform satisfactorily, even when the number of subject and the subject size are quite small are good enough under both $N = 80$ and $T = 8$. It is noticed that under same ρ_1 and N, the estimate of ρ_1 is slightly improved when sample size T increases. This is because in the longitudinal data the large number in each subject may ease the influence of the extra variation between each repeated time and lead to lower sum of Pearson residuals.

Regarding the regression coefficients, the estimates $\hat{\beta}$ and $\hat{\gamma}$ are close to their true values, especially for large N and T, and low degree of correlation. As expected, when the number of subject and sample size increase, the standard error of the coefficient estimates reduce, providing more efficient and less biased estimates $\hat{\beta}$ and $\hat{\gamma}$. In conclusion,

Table 1: The simulation results for regression parameters on mean and dispersion levels based on *exchangeable* correlation structure after 1000 repeated runs.

ρ_1	N	T	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
0.7	240	20	0.679 [0.058]	0.430 [0.146]	0.995 (0.167)	0.931 (0.203)	0.977 (0.611)	0.998 (1.084)
		8	0.695 [0.074]	0.439 [0.156]	0.980 (0.174)	0.897 (0.211)	1.058 (0.647)	0.970 (1.151)
	160	20	0.657 [0.077]	0.363 [0.140]	0.981 (0.201)	0.876 (0.243)	1.038 (0.719)	0.860 (1.274)
		8	0.678 [0.089]	0.383 [0.147]	0.989 (0.208)	0.858 (0.254)	1.067 (0.769)	0.817 (1.381)
	80	20	0.606 [0.109]	0.267 [0.130]	0.957 (0.263)	0.774 (0.328)	1.084 (0.961)	0.700 (1.685)
		8	0.639 [0.119]	0.290 [0.145]	0.968 (0.272)	0.761 (0.339)	1.071 (1.046)	0.657 (1.864)
0.5	240	20	0.479 [0.062]	0.298 [0.122]	0.997 (0.145)	0.927 (0.174)	0.985 (0.509)	0.975 (0.898)
		8	0.511 [0.073]	0.314 [0.129]	0.991 (0.152)	0.930 (0.184)	0.995 (0.552)	0.972 (0.988)
	160	20	0.463 [0.071]	0.254 [0.111]	0.980 (0.171)	0.900 (0.208)	1.011 (0.599)	0.916 (1.046)
		8	0.502 [0.086]	0.276 [0.120]	0.993 (0.186)	0.898 (0.224)	1.040 (0.675)	0.857 (1.190)
	80	20	0.425 [0.093]	0.176 [0.096]	0.979 (0.226)	0.804 (0.280)	1.041 (0.786)	0.885 (1.382)
		8	0.455 [0.108]	0.193 [0.107]	0.975 (0.237)	0.784 (0.300)	1.045 (0.921)	0.857 (1.633)
0.3	240	20	0.301 [0.052]	0.188 [0.093]	0.994 (0.117)	0.970 (0.142)	0.990 (0.419)	0.965 (0.742)
		8	0.330 [0.064]	0.196 [0.097]	1.000 (0.129)	0.944 (0.157)	0.999 (0.470)	0.915 (0.832)
	160	20	0.287 [0.063]	0.155 [0.081]	0.990 (0.139)	0.942 (0.169)	1.001 (0.485)	0.939 (0.846)
		8	0.319 [0.074]	0.172 [0.093]	0.992 (0.156)	0.918 (0.190)	0.983 (0.560)	0.991 (0.995)
	80	20	0.259 [0.075]	0.106 [0.066]	0.985 (0.188)	0.880 (0.228)	1.014 (0.614)	0.941 (1.089)
		8	0.288 [0.092]	0.116 [0.074]	0.971 (0.211)	0.837 (0.257)	0.988 (0.741)	0.993 (1.338)
0.1	240	20	0.112 [0.030]	0.072 [0.047]	0.998 (0.081)	0.985 (0.097)	1.006 (0.300)	0.956 (0.537)
		8	0.142 [0.045]	0.091 [0.067]	0.997 (0.107)	0.972 (0.137)	1.000 (0.387)	0.972 (0.793)
	160	20	0.110 [0.035]	0.064 [0.047]	0.991 (0.097)	0.974 (0.117)	0.960 (0.354)	1.042 (0.631)
		8	0.134 [0.050]	0.075 [0.061]	1.001 (0.122)	0.955 (0.147)	0.970 (0.453)	1.016 (0.804)
	80	20	0.101 [0.045]	0.046 [0.039]	0.987 (0.134)	0.945 (0.162)	0.965 (0.460)	1.053 (0.823)
		8	0.122 [0.061]	0.052 [0.051]	0.986 (0.165)	0.906 (0.202)	0.981 (0.619)	0.985 (1.110)

† [ ] denotes sample standard deviation and ( ) estimated standard error.

our proposed method is able to handle the log-linear regression both on the mean and dispersion, particularly for the case with large sample and weak correlation.

There are still some limitations found in the simulation study. When the correlation within each subject is quite high, $\rho_1 = 0.7$, in general the divergence rate is greater than the one under smaller correlation assumptions. As the number of subject and subject size go down, less simulation outcomes diverge. This implies that the proposed method performs well when the sample size is large enough but may have some drawback in the small sample case.

6. Application to CUPDAY data

0.3 Overview

According to the formulation of the milk product provided, the children attending in the study were randomly allocated to 2 groups, namely 'intervention' and 'control'. Thus to fit a regression model for the CUPDAY data, we first defined an indicating variables (**Group**) where 1 represents the trial group and 0 the control group. Besides, we took the number of days with four or more stools (**FOURSTOOL**) as the responses, which measures whether or not the added ingredients make positive effects on diarrhoea. The other 3 covariates included in our specified model are the adjusted age (**Age**), average milk consumption of each child during the study (**Milkcons**) and number of days other family members in illness (**Familyill**). Note that the number of days enrolled in the study is treated as an exposure variable in the linear predictor of model. 29 child-care centres are considered. We chose the link functions of both mean and dispersion level to be *log*, which is known as the canonical link for count data, and we used the *exchangeable* assumption to simulate the correlation within each cluster.

0.4 Result analysis

First, we fitted the multivariate Poisson model as Binns et al. (2007) and found great overdispersion among the data. According to our past study and experience, in the estimation procedure, the constant dispersion in the Poisson count data does not affect the result since it can always be canceled out in the iterative algorithm. This motivates us to investigate the target data more deeply. After reexamining the CUPDAY data, we found that there exists an extra outlier in center 29 thus we doubted that it may significantly give rise to the dispersion and cause huge bias in yielding estimates. The preliminary analysis result is shown in Table 2. Compared with the same Poisson model for the complete data, the reduced data model still exhibits strong overdispersion, though the dispersion drops to be nearly half of the former one. The **Group** turns from negative effect to be positive effect on the response, which means the trial product may not reduce the incidence of diarrhoea, while effects of other covariates do not change much.

This motivates us to consider that **Group** may have some association with the dispersion, which can not be revealed by the model with constant dispersion. Therefore, we chose the covariate **Group** of interest in the regression on the dispersion level.

Next we modeled the data by using the proposed joint regressions under negative binomial distribution on both mean and dispersion. For the purpose of comparison, we also applied the multivariate Poisson regression used in Binns et al. (2007) and the GEE method proposed by Jowaheer & Sutradhar (2002) for these repeated overdispersed data. From the results summarized in Table 3, under both the GEE negative binomial models, variable **Group** shows positive influence on the outcome of **FOURSTOOL** which is contrary to the corresponding effects in the Poisson regression. This implies that compared with the control product, the trial milk has limited effects on reducing the diarrhoeal rate. On the dispersion level, as we expected, the coefficient estimate of **Group** is -0.245, which indicates that the control group tends to have relatively high dispersion among the data, while the incidence of diarrhoea in the intervention group exhibits less fluctuation. In particular, the dispersion in the proposed joint GEE model is estimated to be 12.280 for the control group while it is only 9.708 in the intervention group.

Recall that the outlier of CUPDAY data is observed in the control group. The negative effect of **GROUP** now can explain the reason why we obtained extra large dispersion in the previous multivariate Poisson model. The estimates of correlation given in Table 3 show that the association which affects the response **FOURSTOOL** is fairly weak among children in the same center.

Table 2: The estimation results of Poisson regression models based on full data set and reduced data set.

Model	Estimates					
			Mean			Log-dispersion
	Int	Group	Age	Milkcons	Familyill	Int
Poisson	-3.805	-0.190	0.104	0.000	0.021	3.152
Complete data	(0.857)	(0.375)	(0.178)	(0.001)	(0.010)	/
GEE.pois	-4.035	0.128	0.138	0.000	0.030	2.622
Reduced data	(0.719)	(0.248)	(0.189)	(0.001)	(0.016)	/

† ( ) denotes estimated standard error.

Table 3: The estimation results of our proposed joint GEE modeling on both mean and dispersion levels for CUPDAY data (JGEE.NB), and the comparison with multivariate Poisson regression (Reg.Pois) and GEE with constant overdispersion under negative binomial (GEE.NB).

	Estimates					
	Reg.Pois		GEE.NB		JGEE.NB	
Mean						
Int	-3.805	(0.857)	-3.409	(0.843)	-3.428	(0.868)
Group	-0.190	(0.375)	0.279	(0.331)	0.251	(0.335)
Age	0.104	(0.178)	-0.277	(0.276)	-0.224	(0.287)
Milkcons	0.000	(0.001)	0.001	(0.001)	0.001	(0.001)
Familyill	0.021	(0.010)	0.005	(0.019)	0.004	(0.019)
Dispersion	ϕ		ϕ		Int	Group
Estimate	23.39		11.04		2.508	-0.245
Std.err	/		(4.496)		(0.574)	(0.730)
Correlation						
ρ_1	/		-0.008		-0.007	
ρ_2	/		-0.006		-0.005	

† ( ) denotes standard error.

To further explore the intervention effects, we also compare the 95% confidence intervals for the population mean μ under the multivariate Poisson regression and the proposed joint GEE method, respectively. Let $\bar{\hat{\mu}}$ denote the average estimated mean and $sd(\hat{\mu})$ the sample standard error for $\hat{\mu}$, which measures the variation of all $\hat{\mu}_{ij} = exp(x_{ij}^T\hat{\beta})$. By the law of large number, the $1-\kappa$ confidence interval for μ is constructed approximately by

$$P\left(\left|\frac{\bar{\hat{\mu}}-\mu}{sd(\hat{\mu})/\sqrt{NT}}\right| < Z_{1-\kappa/2}\right) = 1-\kappa, \tag{0.17}$$

where $Z_{1-\kappa/2}$ is the $1-\kappa/2$ normal quantile. The data set is then split into two subsets, one for the intervention group and the other for the control group. Table 4 reports the 95% confidence intervals for μ under the two models with different group. It is shown that the intervention has no positive effect on **FOURSTOOL** for children considering the variation on the dispersion level.

Table 4: 95% confidence intervals for multivariate Poisson and the proposed negative binomial GEE models.

	Intervention	Control
Poisson	(1.587, 1.961)	(2.100, 2.610)
JGEE.NB	(2.783, 3.450)	(2.307, 2.838)

7. Conclusion In this paper, we propose a joint GEE method for analyzing longitudinal or clustered count data with overdispersion. The method extends the standard GEE model for population average to both mean and dispersion levels and allows for the relationship between the covariates and the extra variation existing in the model. In order to handle the possible overdispersion that occurs in the count data, we used the negative binomial assumption to model the link between mean and variance structure. Our method provides a natural extension to the GEE approach for overdispersed longitudinal data developed by Jowaheer & Sutradhar (2002). Our simulation results show that the proposed method performs satisfactorily in estimating model parameters provided that the sample size is relatively large. To set up the target count data, an *exchangeable* process was derived for simulation studies.

Next we applied our method to a set of overdispersed data by modeling the dispersion part with the known covariates. Compared with the early analysis of CUPDAY research, it was seen that the result could be strongly impacted by the large dispersion, which is often ignored and treated as nuisance parameter in the standard GEE analysis. Moreover, the overdispersion was integrated by the covariates of interest. The analysis results support the fact that the overdispersion must be affected by some factors contributing to the construction the whole statistical model.

Appendix

$\langle 1 \rangle$ ***Variance derivation.***

To simplify the computation in the following steps, we use different notations from Section to define the negative binomial distribution, assuming the pmf of a negative binomial count y to be

$$p(y) = \frac{\Gamma(y+r)}{\Gamma(r)y!} p^r (1-p)^y.$$

Then the moment generating function for y, is known as

$$M(t) = p^r [1 - qe^t]^{-r},$$

where $q = 1 - p$. Note that the parameters in the definition of $p(y)$ above are easily associated with those in (0.1) as $r = \phi^{-1}$, $p = (1 + \phi\mu)^{-1}$ and $\mu = rq/p$. Thus, by the fact that the raw moment $E(Y^n) = M^{(n)}(0)$, we differentiate $M(t)$ with respect to t and obtain that

$$
\begin{aligned}
M'(t) &= p^r qr(1-qe^t)^{-r-1}e^t, \\
M''(t) &= p^r qr(1-qe^t)^{-r-1}e^t + p^r qre^t(1-qe^t)^{-r-2}(r+1)qe^t \\
&= p^r qre^t(1-qe^t)^{-r-2}(1+qre^t), \\
M^{(3)}(t) &= p^r qre^t(1-qe^t)^{-r-2}(1+qre^t) + p^r q^2r^2e^{2t}(1-qe^t)^{-r-2} \\
&\quad + p^r qre^t(1-qe^t)^{-r-3}(1+qre^t)(r+2)qe^t \\
&= p^r qre^t(1-qe^t)^{-r-3}[(1-qe^t)(1+qre^t) + qre^t(1-qe^t) + (1+qre^t)(r+2)qe^t] \\
&= p^r qre^t(1-qe^t)^{-r-3}[1+(3qr+q)e^t + q^2r^2e^{2t}], \\
M^{(4)}(t) &= p^r qre^t(1-qe^t)^{-r-3}[1+(3qr+q)e^t + q^2r^2e^{2t}] \\
&\quad + p^r qre^t(1-qe^t)^{-r-3}[(3qr+q)e^t + 2q^2r^2e^{2t}] \\
&\quad + p^r q^2r(r+3)e^{2t}(1-qe^t)^{-r-4}[1+(3qr+q)e^t + q^2r^2e^{2t}] \\
&= p^r qre^{2t}(1-qe^t)^{-r-4}[1+(7qr+4q)e^t + (6q^2r^2+4q^2r+q^2)e^{2t} + q^3r^3e^{3t}].
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
E(Y^4) = M^{(4)}(0) &= \frac{rq}{p^4}(1+4q+7qr+q^2+4q^2r+6q^2r^2+q^3r^3) \\
&= (1+\phi\mu)^3\mu + 4(1+\phi\mu)^2\phi\mu^2 + 7(1+\phi\mu)^2\mu^2 + (1+\phi\mu)\phi^2\mu^3 + \\
&\quad 4(1+\phi\mu)\phi\mu^3 + 6(1+\phi\mu)\mu^3 + \mu^4 \\
&= \mu + (7+7\phi)\mu^2 + (6+18\phi+12\phi^2)\mu^3 + (1+6\phi+11\phi^2+6\phi^3)\mu^4,
\end{aligned}
$$

then it follows that $Var(Y^2) = E(Y^4) - [E(Y^2)]^2 = \mu + (6+7\phi)\mu^2 + (4+16\phi+12\phi^2)\mu^3 +$

$(4\phi + 10\phi^2 + 6\phi^3)\mu^4$. $\square$

$\langle 2 \rangle$ ***Proof of Theorem 0.3***.

To get $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{1/N})$, it is sufficient to prove that for any $\epsilon > 0$, there exists some large constant $\mathcal{C}_\epsilon$ such that a local minimizer $\hat{\theta}$ lies in the interior of the ball $\left\{\theta_0 + N^{-\frac{1}{2}}u,\ \|u\| \leq \mathcal{C}_\epsilon\right\}$ (Fan & Li, 2001).

Next let $Q(\theta)$ be some quasi-likelihood function, such that

$$\dot{Q}(\theta) = \partial Q(\theta)/\partial\theta = S(\theta).$$

We also known the second derivative of $Q(\theta)$ exists given (1), denoted by $\ddot{Q}(\theta) = -\Sigma(\theta)$. Further we define

$$H(u) = Q(\theta_0 + N^{-\frac{1}{2}}u) - Q(\theta_0).$$

By (A1),

$$\begin{aligned} Q(\theta_0 + N^{-\frac{1}{2}}u) = Q(\theta_0) + N^{-\frac{1}{2}}u^T\dot{Q}(\theta_0) + \frac{1}{2}N^{-1}u^T\ddot{Q}(\theta_0)u \\ + \frac{1}{6}N^{-\frac{3}{2}}\left(\frac{\partial}{\partial\theta}(u^T\ddot{Q}(\theta^*)u)\right)^T u, \end{aligned}$$

where θ^* is between θ_0 and $\theta_0 + N^{-\frac{1}{2}}u$. We denote the last three terms on the right side by (a), (b) and (c) in order, then by (A4) and the law of large numbers, $S(\theta)$ is $O_p(\sqrt{N})$

for any fixed θ. Therefore,

$$\|(a)\| = N^{-\frac{1}{2}} u^T S(\theta_0) = O_p(1)\|u\|.$$

By (A2),

$$\|(b)\| = \frac{1}{2} N^{-1} u^T \ddot{Q}(\theta_0) u = \frac{1}{2} N^{-1} u^T \Sigma_0(\theta_0) u + o_p(1)\|u\|^2.$$

And by (A3),

$$\|(c)\| = \frac{1}{6} N^{-\frac{3}{2}} \left(\frac{\partial}{\partial \theta} (u^T \ddot{Q}(\theta^*) u) \right)^T u = O_p(\sqrt{1/N})\|u\|^3.$$

Finally, we have

$$\begin{aligned} H(u) &= (a) + (b) + (c) \\ &= O_p(1)\|u\| + \frac{1}{2} N^{-1} u^T \Sigma_0(\theta_0) u + o_p(1)\|u\|^2 + O_p(\sqrt{1/N})\|u\|^3. \end{aligned}$$

$H(u)$ is therefore dominated by the positive term $N^{-1} u^T \Sigma_0(\theta_0) u$. As $N \to \infty$, we conclude that there exists at least a local minimizer $\hat{\theta}$ of (0.12), which lies in the ball. □

⟨3⟩ ***Proof of Theorem 0.4***.

Using the Taylor expansion of $S\hat{\theta}$ about point θ_0, we obtain

$$
\begin{aligned}
S(\hat{\theta}) &= \dot{Q}(\theta_0) + \ddot{Q}(\theta_0)(\hat{\theta} - \theta 0) \\
&\quad + \frac{1}{6} N^{-\frac{3}{2}} \left(\frac{\partial}{\partial \theta} \left((\hat{\theta} - \theta_0)^T \ddot{Q}(\theta^*)(\hat{\theta} - \theta_0) \right) \right)^T ((\hat{\beta}, \hat{\gamma}) - (\beta_0, \gamma_0)),
\end{aligned}
$$

where θ^* is between $\hat{\theta}$ and θ_0. As it has been proved that $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{1/N})$ and from (c) in the proof of Theorem 0.3, we know that the last term on the right side of equation above is $o_p(1)$. Thus

$$
\hat{\theta} - \theta_0 = \Gamma(\theta_0)^{-1} S(\theta_0) + o_p(1).
$$

From (A4) and the Central Limit Theorem, it follows that $S(\theta_0)$ is normally distributed with 0 mean. Further by Stutsky's Theorem,

$$
\sqrt{N}(\hat{\theta} - \theta_0) \to \mathcal{N}\left(0, N\Sigma^{-1}\left(\sum_{i=1}^{N} D_i^T \tilde{V}_i^{-1} Cov(r_i) \tilde{V}_i^{-1} D_i\right)\Sigma^{-1}\right),
$$

in distribution, which coincides with the conclusion of (0.16). □

$\langle 4 \rangle$ ***Correlated responses generating process***.

In order to construct the model for negative binomial outcomes with exchangeable correlation, we develop the method in the same manner of McKenzie (1986). First, learning from the structures of X_i and W_i built in Section 0.2, we know that Y_{it}'s in each subject i share the same mean and dispersion values. We denote them as μ_i and ϕ_i. Under *exchangeable* assumption, y_{ij} can be generated through the following process:

$$y_{ij} = \varphi_{ij} * y_{i0} + e_{ij},$$

where the initial value $y_{i0} \sim NB(\mu_i, \phi_i)$, and $e_{ij} \sim NB((1-\sqrt{\rho_1})\mu_i,\ \phi_i/(1-\sqrt{\rho_1}))$, *iid* for all i and j. Similar in McKenzie (1986), $\varphi_{ij} * y_{i,j-1}$ denotes some binomial variable, conditional on y_{i0} and given φ_{ij}. For convenience, we define

$$b_{ij} = \varphi_{ij} * y_{i0}, \quad b_{ij}|y_{i0}, \varphi_{ij} \sim Binomial(y_{i0}, \varphi_{ij}),$$

where φ_{ij} is assumed to be independent for all i and j with $Beta(\sqrt{\rho_1}/\phi_i, (1-\sqrt{\rho_1})/\phi_i)$. The pmf of b_{ij} conditional on y_{i0} can be obtained by

$$\begin{aligned} p(b_{ij}|y_{i0}) = & \frac{y_{i0}!\Gamma(1/\phi_i)}{b_{ij}!(y_{i0}-b_{ij})!\Gamma\{(1-\sqrt{\rho_1})/\phi_i\}\Gamma(\sqrt{\rho_1}/\phi_i)} \\ & \times \frac{\Gamma(\sqrt{\rho_1}/\phi_i + b_{ij})\Gamma\{(1-\sqrt{\rho_1})/\phi_i + y_{i0} - b_{ij}\}}{\Gamma(1/\phi_{ij} + y_{i0})}. \end{aligned}$$

Since the pmf of y_{i0} is known as

$$p(y_{i0}) = \frac{\Gamma(y_{i0}+1/\phi_i)}{\Gamma(1/\phi_i)y_{i0}!} \left(\frac{1}{1+\phi_i\mu_i}\right)^{1/\phi_i} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{y_{i0}},$$

set $x_{ij} = y_{i0} - b_{ij}$, the pmf of b_{ij} is given by

$$
\begin{aligned}
p(b_{ij}) &= \sum_{y_{i0}=b_{ij}} p(b_{ij}|y_{i0})p(y_{i0}) \\
&= \sum_{y_{i0}=b_{ij}} \frac{\Gamma(\sqrt{\rho_1}/\phi_i + b_{ij})\Gamma\{(1-\sqrt{\rho_1})/\phi_i + y_{i0} - b_{ij}\}}{\Gamma(\sqrt{\rho_1}/\phi_i)\Gamma\{(1-\sqrt{\rho_1})/\phi_i\}b_{ij}!(y_{i0}-b_{ij})!} \left(\frac{1}{1+\phi_i\mu_i}\right)^{\frac{1}{\phi_i}} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{y_{i0}} \\
&= \frac{\Gamma(\sqrt{\rho_1}/\phi_i + b_{ij})}{b_{ij}!\Gamma(\sqrt{\rho_1}/\phi_i)} \sum_{x_{ij}=0} \frac{\Gamma\{(1-\sqrt{\rho_1})/\phi_i + x_{ij}\}}{\Gamma\{(1-\sqrt{\rho_1})/\phi_i\}x_{ij}!} \left(\frac{1}{1+\phi_i\mu_i}\right)^{\frac{1}{\phi_i}} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{x_{ij}+b_{ij}} \\
&= \frac{\Gamma(\sqrt{\rho_1}/\phi_i + b_{ij})}{b_{ij}!\Gamma(\sqrt{\rho_1}/\phi_i)} \left(\frac{1}{1+\phi_i\mu_i}\right)^{\frac{\sqrt{\rho_1}}{\phi_i}} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{b_{ij}} \\
&\times \sum_{x_{ij}=0} \frac{\Gamma\{(1-\sqrt{\rho_1})/\phi_i + x_{ij}\}}{\Gamma\{(1-\sqrt{\rho_1})/\phi_i\}x_{ij}!} \left(\frac{1}{1+\phi_i\mu_i}\right)^{\frac{(1-\sqrt{\rho_1})}{\phi_i}} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{x_{ij}}.
\end{aligned}
$$

It's easy to see that in the above equation $x_{ij} \sim NB((1-\sqrt{\rho_1})\mu_i, \phi_i/(1-\sqrt{\rho_1}))$. We have

$$
\sum_{x_{ij}=0} \frac{\Gamma\{(1-\sqrt{\rho_1})/\phi_i + x_{ij}\}}{\Gamma\{(1-\sqrt{\rho_1})/\phi_i\}x_{ij}!} \left(\frac{1}{1+\phi_i\mu_i}\right)^{\frac{(1-\sqrt{\rho_1})}{\phi_i}} \left(\frac{\phi_i\mu_i}{1+\phi_i\mu_i}\right)^{x_{ij}} = 1.
$$

Then $b_{ij} \sim NB(\sqrt{\rho_1}\mu_i, \phi_i/\sqrt{\rho_1})$. Recalling that $e_{ij} \sim NB((1-\sqrt{\rho_1})\mu_i, \phi_i/(1-\sqrt{\rho_1}))$, we conclude that $y_{ij} \sim NB(\mu_i, \phi_i)$ by verifying its moment generating function.

It is clear that the correlation structure could be deduced from the above process. For $k = 1, \ldots, T-j$, we know $y_{ij}y_{i,j+k} = (\varphi_{ij} * y_{i0} + e_{ij})(\varphi_{i,j+k} * y_{i0} + e_{i,j+k})$, then $E(y_{ij}y_{i,j+k}) = \rho_1\mu_i + \rho_1(\phi_i+1)\mu_i^2 + (1-\rho_1)\mu^2 = \rho_1\mu_i + \rho_1\phi_i\mu_i^2 + \mu_i^2$ since all the variables are independent over all i and j. Therefore $corr(y_{ij}, y_{i,j+k}) = \rho_1$, leading to the exchangeable correlation structure. $\square$

References

[1] Binns, C., Lee, A., Harding, H., Gracey, M., and Barclay, D. The cupday study: prebiotic-probiotic milk product in 1-3-year-old children attending childcare centres. *Acta Padiatrica. 96* (2007), 1646–1650.

[2] Breslow, N. Extra-poisson variation in log-linear models. *Journal of Applied Statistics. 33* (1984), 38–44.

[3] Consul, P. C., and Jain, G. C. A generalization of the poisson distribution. *Technometrics 15*, 4 (1973), 791–799.

[4] Cox, D. R. Some remarks on overdispersion. *Biometrika 70*, 1 (1983), 269–274.

[5] Efron, B. Double exponential-families and their use in generalized linear-regression. *Journal of the American Statistical Association. 81* (1986), 709–721.

[6] Fahrmeir, L., and Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics. 13* (1985), 342–368.

[7] Fleiss, J. *Statistical methods for rates and proportions.* J. Wiley & Sons, New York, 1981.

[8] Gelfand, A., and Dalal, S. A note on overdispersed exponential families. *Biometrika. 77* (1990), 55–64.

[9] Jowaheer, V., and Sutradhar, B. Analysing longitudinal count data with overdispersion. *Biometrika. 89* (2002), 389–399.

[10] Lawless, J. Negative binomial and mixed poisson regression. *Canadian Journal of Statistics. 15* (1987), 209–225.

[11] Leng, C., Zhang, W., and Pan, J. Semiparametric mean covariance regression analysis for longitudinal data. *Journal of the American Statistical Association. 105* (2010), 181–193.

[12] Liang, K., and Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika. 73* (1986), 13–22.

[13] McKenzie, E. Autoregressive moving-average processes with negative binomial and geometric marginal distributions. *Advances in Applied Probability. 18* (1986), 679–705.

[14] Pourahmadi, M. Joint mean-covariance models with applications to longitudinal data: uncostrained paprameterisation. *Biometrika. 86* (1999), 677–690.

[15] Prentice, R., and Zhao, L. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics. 47* (1991), 825–839.

[16] Thall, P., and Vail, S. Some covariance models for longitudinal count data with overdispersion. *Biometrics. 46* (1990), 657–671.

[17] Ye, H., and Pan, J. Modelling of covariance structures in generalized estimating equations for longitudinal data. *Biometrika. 93* (2006), 927–941.