Regression models for mixed Poisson and continuous longitudinal data

Ying Yang*, †, Jian Kang, Kai Mao and Jie Zhang

Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People’s Republic of China

SUMMARY

In this article we develop flexible regression models in two respects to evaluate the influence of the covariate variables on the mixed Poisson and continuous responses and to evaluate how the correlation between Poisson response and continuous response changes over time. A scenario for dealing with regression models of mixed continuous and Poisson responses when the heterogeneous variance and correlation changing over time exist is proposed. Our general approach is first to jointly build marginal model and to check whether the variance and correlation change over time via likelihood ratio test. If the variance and correlation change over time, we will do a suitable data transformation to properly evaluate the influence of the covariates on the mixed responses. The proposed methods are applied to the interstitial cystitis data base (ICDB) cohort study, and we find that the positive correlations significantly change over time, which suggests heterogeneous variances should not be ignored in modelling and inference. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: Poisson responses; generalized estimating equation; longitudinal data; correlation changing over time; likelihood ratio test

1. INTRODUCTION

The interstitial cystitis data base (ICDB) cohort study [1] motivated this study. Interstitial cystitis (IC) is a chronic illness characterized by symptoms in at least one of three facts: pain in the pelvic or bladder area, urgency (pressure to urinate) and frequency of urination. Not all subjects diagnosed with IC have all three symptoms, and there are theories that the diagnosis includes subgroups with different underlying disease mechanisms and different sets of symptoms. One of the aims

*Correspondence to: Ying Yang, Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People’s Republic of China.
† E-mail: yyang@math.tsinghua.edu.cn

Contract/grant sponsor: National Natural Science Foundation of China; contract/grant number: 10671106

Received 17 October 2006
Accepted 22 October 2006
evaluating this theory would be to determine whether symptoms tend to co-fluctuate together (suggesting a single underlying aetiology) or vary independently (suggesting multiple mechanisms at work). The ICDB contains longitudinal data on pain, urgency and urinary frequency from a prevalent cohort of subjects with IC. For each individual, pain score \( p \), urinary urgency \( u \) and urinary frequency \( f \) are associated with covariates, such as the demographic and clinic characteristics of patients, which were observed at different times, where variables \( p \) and \( u \) are continuous variables and \( f \) is discrete. It is reasonable to regard \( f \) as a Poisson variable. Our objective of this article is to develop regression models in two respects to evaluate the influence of the covariate variables on the responses and to investigate the correlation changing over time between the continuous response (e.g. \( p \) or \( u \)) and Poisson response (e.g. \( f \)).

For the bivariate binary response and continuous response with clustering, Fitzmaurice and Laird [2] discussed regression methods for jointly analysing bivariate binary and continuous responses that arise from development toxicity studies. They modelled marginal expectations of both variables and whereas the association between the binary and continuous response is regarded as a nuisance parameter of the data. Based on the general location model of Olkin and Tate [3], they developed a likelihood-based approach to estimate the marginal mean parameters. In our ICDB data set, the discrete variable is Poisson instead of binary. Obviously, the methods developed by Fitzmaurice and Laird [2] could not be applied directly to our problem. On the other hand, we are interested in the correlation between Poisson response and continuous response. Therefore, we have to estimate the correlation over time.

In this article we develop methods to jointly model marginal expectation of the Poisson response and continuous response and to explore how the correlation between the above two responses changes over time. The generalized estimating equation (GEE) methodology is used to estimate the regression parameters and the correlations over time. The likelihood ratio test is employed to check whether the variances change over time. We find that the variance and correlation change over time in our motivated ICDB data set. Further the data transformation technique and likelihood ratio test approach are proposed to deal with the case of variances and correlation changing over time. Our general approach is first to jointly build marginal model and to check whether the variance and correlation change over time via likelihood ratio test. If the variance and correlation change over time, we will do a suitable data transformation to properly evaluate the influence of the covariates on the mixed responses. Details are discussed in Sections 2.2 and 3.

We will describe the joint distribution as the product of the marginal distribution of one response and the conditional distribution of the other. Following this idea, two models are available, depending on which response is studied firstly. There are some literature on both of these methods: (1) describe the marginal distribution of discrete response first, then define the conditional distribution of the continuous one (for more details, see [3–5] and the references therein); (2) describe the marginal distribution of continuous response first, then give the conditional distribution of the discrete one (e.g. [6]).

The remainder of the article is organized as follows. In Section 2, we describe the regression model and estimation methods for bivariate Poisson and continuous responses. In Section 3 we investigate the correlation between Poisson and continuous responses, a regression model directly on the correlation coefficient is proposed in Section 3.1 and we provide another approach to study the correlation through the covariance structure of random errors over time in Section 3.2. The likelihood ratio test approach is used to test whether the variances and correlation change over time.
In the context of the variances changing over time, a data transformation technique is proposed to shake the impact of heterogeneous variances upon the evaluation of the varying coefficient function in the regression model of the mixed responses. In Section 4 we conduct Monte Carlo study to evaluate the performance of the proposed methods. Section 5 applies the approaches on the ICDB diastases. Section 6 contains a discussion.

2. MODEL AND PARAMETER ESTIMATION

In this section we describe regression model for mixed Poisson and continuous responses and provide methods to estimate the regression parameters.

2.1. Likelihood function

Let \( X_i \) and \( Y_i \) be Poisson and continuous response, respectively, \( Z_{1i} \) and \( Z_{2i} \) be \( p \times 1 \) and \( q \times 1 \) covariate vectors, respectively, \( i = 1, \ldots, n \). Assume the marginal distribution of \( X_i \) is Poisson

\[
 f(x_i | Z_i) = \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}, \quad 1 \leq i \leq n
\]

where \( \log \lambda_i = \theta_i = Z_{1i}^T \beta_1 \), and \( \lambda_i \) is the parameter in Poisson distribution and \( \beta_1 \) is a \( p \times 1 \) unknown parameter vector to be estimated, the superscript T denotes the transpose of a matrix or vector. Obviously, \( E(x_i) = \lambda_i \), \( \text{var}(x_i) = \lambda_i \).

Given \( X_i \), we can assume that the conditional distribution of \( Y_i \) is normal

\[
 f_{Y_i | X_i}(y_i | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i - \gamma(x_i - \lambda_i))^2}{2\sigma^2}\right\}
\]

where \( \mu_i = Z_{2i}^T \beta_2 \), \( \gamma \) is a regression parameter of \( Y_i \) on \( X_i \) describing their correlation and \( \beta_2 \) is an unknown \( q \times 1 \) vector. For convenience of notations, let \( z = (\beta_2^T, \gamma)^T \) and \( W_i = (Z_{2i}^T, x_i - \lambda_i)^T \). Then

\[
 E(Y_i | X_i) = W_i^T z
\]

So, the joint distribution of \((X_i, Y_i)\) can be written as

\[
 f_{X_i, Y_i}(x_i, y_i) = f_{X_i}(x_i) f_{Y_i | X_i}(y_i | x_i)
\]

where \( E(Y_i) = Z_{2i}^T \beta_2 \), \( \beta_1 \) and \( \beta_2 \) are regression parameters, which have marginal interpretations.

To the log mean of Poisson response and the mean of continuous response, we have following linear regression models:

\[
 \log \lambda_i = Z_{1i}^T \beta_1 \\
 \mu_i = Z_{2i}^T \beta_2
\]

2.2. Likelihood equations

For convenience we assume \( Z_{1i} = Z_{2i} = Z_i \), otherwise we can define \( Z_i \) is the combination of \( Z_{1i} \) and \( Z_{2i} \) if \( Z_{1i} \neq Z_{2i} \).
Then the likelihood equations can be written as:

\[
\sum_{i=1}^{n} \frac{\partial l_i(y_i, x_i)}{\partial \beta_1} = \sum_{i=1}^{n} (Z_i(x_i - \lambda_i) - Z_i \Delta_i(y_i - W_i^T \beta) \sigma^{-2})
\]  
(1)

\[
\sum_{i=1}^{n} \frac{\partial l_i(y_i, x_i)}{\partial \mu} = \sum_{i=1}^{n} W_i(y_i - W_i^T \mu) \sigma^{-2}
\]  
(2)

\[
\sum_{i=1}^{n} \frac{\partial l_i(y_i, x_i)}{\partial \sigma^2} = \Phi^{-1} \sum_{i=1}^{n} (c_i - \sigma^2)
\]  
(3)

where \( \Delta_i = \text{var}(X_i) = \lambda_i, C_i = (Y_i - W_i^T \mu)^2 \) and \( \Phi = \text{var}(C_i) \).

Note that the likelihood equations for \( \beta_1, \mu, \sigma^2 \) given by (1)–(3) can be also written as

\[
\sum_{i=1}^{n} \left( \frac{\partial l_i}{\partial \beta_1} \right) = \sum_{i=1}^{n} \begin{pmatrix} \Delta_i Z_i & -\gamma \Delta_i Z_i & 0 \\ 0 & W_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta_i^{-1} & 0 & 0 \\ 0 & \sigma^{-2} & 0 \\ 0 & 0 & \Phi^{-1} \end{pmatrix} \begin{pmatrix} x_i - \lambda_i \\ y_i - W_i^T \beta \\ c_i - \sigma^2 \end{pmatrix}
\]  

\[
\sum_{i=1}^{n} \left( \frac{\partial E(X_i)}{\partial \beta_1} \frac{\partial E(X_i)}{\partial \mu} \frac{\partial E(X_i)}{\partial \sigma^2} \right)^T \mathbf{cov}^{-1} \begin{pmatrix} X_i \\ Y_i | X_i \\ C_i | X_i \end{pmatrix} \begin{pmatrix} x_i - E(X_i) \\ y_i - E(Y_i | X_i) \\ c_i - E(C_i | X_i) \end{pmatrix}
\]

The likelihood equations above can be solved iteratively. Apply iteratively reweighed least squares (IRS) to (2) and (3) we can get maximum likelihood estimates (MLE) of \( \mu \) and \( \sigma^2 \)

\[
\hat{\mu} = \left( \sum_{i=1}^{n} W_i W_i^T \right)^{-1} \sum_{i=1}^{n} W_i y_i
\]  
(4)

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - W_i^T \hat{\mu} \right)^2
\]  
(5)

Then we can put those estimates into (1) and use Fisher Scoring algorithm to get MLE of \( \beta_1 \).
The covariance of \( \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}, \hat{\sigma}^2 \) can be approximated by the inverse of the Fisher information matrix, that is

\[
\text{cov}^{-1}(\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}, \hat{\sigma}^2) \approx \sum_{i=1}^{n} \begin{bmatrix}
(\Delta_i + \Delta_i^2 \gamma^2 \sigma^{-2}) Z_i Z_i^T & (\sigma^{-2} \gamma \Delta_i)^2 Z_i Z_i^T & 0 & 0 \\
(\sigma^{-2} \gamma \Delta_i) Z_i Z_i^T & \sigma^{-2} Z_i Z_i^T & 0 & 0 \\
0 & 0 & \Delta_i \sigma^{-2} & 0 \\
0 & 0 & 0 & \frac{1}{2} \sigma^{-4}
\end{bmatrix}
\]

Note that the elements in the position \( \hat{\beta}_1, \hat{\sigma}^2 \), \( \hat{\beta}_2, \hat{\sigma}^2 \), \( \hat{\gamma}, \hat{\sigma}^2 \) in the above covariance matrix are all zeroes, indicating that the three groups of parameters are orthogonal, which is consistent with the conclusion of binary data [2, 7, 8]. The conclusion of this section is similar to those in [2].

### 2.3. GEE methodology to handle longitudinal data

In this section we consider the case of longitudinal data, which means there are multiple observations at different times for each unit. Here we assume that each of \( N \) independent individuals has observations on \( n_i \) mixed Poisson and continuous response vectors \( (X_i, Y_i) \), where \( X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})^T, Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T \) and the covariate variables can be written as \( Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in_i}) \), a \( p \times n_i \) matrix, where \( Z_{ik} \) is a \( p \times 1 \) vector.

In this case, as pointed out by Liang et al. [9], it is unavailable and too complicated to use maximum likelihood method. Here we employ GEE methodology to handle longitudinal mixed Poisson and continuous data. First, we build regression model for Poisson and continuous responses as follows:

\[
\log(E(X_i)) = Z_i^T \beta_1
\]

\[
E(Y_{ik}|X_i) = Z_{ik}^T \beta_2 + \gamma_1(X_{ik} - \lambda_i) + \gamma_2 S_i
\]

where \( S_i = \sum_{j=1}^{n_i} (X_{ij} - \lambda_{ij}) \), and \( \gamma_1, \gamma_2 \) denote the correlation between \( X_i \) and \( Y_i \) with \( \gamma_1 \) representing the influence of \( X_i \) on \( Y_i \) within the observation and \( \gamma_2 \) the influence of other observations within that unit. To simplify the expression, let \( W_{ik} = (Z_{ik}^T, X_{ik} - \lambda_{ik}, S_i)^T \) and \( z = (\beta_2^T, \gamma_1, \gamma_2)^T \). Then the second formula above can be simplified as

\[
E(Y_{ik}|X_i) = W_{ik}^T z
\]

Following the same idea of Fitzmaurice and Laird [2], we assume that within correlation are \( \rho_X \) and \( \rho_Y \) for the Poisson and continuous responses, respectively, and we anticipate observations within a unit to be positively correlated. Therefore, we get the following approximately covariance matrices:

\[
\text{cov}(X_i) = V_{1i} \approx \Delta_i^{1/2}[(1 - \rho_X) I_i + \rho_X J_i] \Delta_i^{1/2}
\]

\[
\text{cov}(Y_i|X_i) = V_{2i} \approx \sigma^2[(1 - \rho_Y) I_i + \rho_Y J_i]
\]

where \( \Delta_i \) is diagonal matrix with elements \( \text{var}(X_i) = \lambda_i, I_i \) is an \( n_i \times n_i \) identity matrix, and \( J_i \) is an \( n_i \times n_i \) matrix of 1’s.
The specific definitions of $\rho_X, \rho_Y, \sigma^2$ are as follows:

\[ r_{X_{ij}} = \frac{X_{ij} - \hat{\lambda}_{ij}(\hat{\beta}_1)}{\sqrt{\hat{\lambda}_{ij}(\hat{\beta}_1)}} \quad (10) \]

\[ r_{Y_{ij}} = Y_{ij} - W_{ij}^T \hat{\alpha} \quad (11) \]

\[ \hat{\Phi}^{-1} = \left( \sum_{i=1}^{n} \sum_{j=1}^{n} r_{X_{ij}}^2 \right) \left( \sum_{i=1}^{n} n_i - p \right) \quad (12) \]

\[ \hat{\rho}_X = \hat{\Phi} \left( \sum_{i=1}^{n} \sum_{j \geq j'} r_{X_{ij} r_{X_{ij}'}} \right) \left( \sum_{i=1}^{n} \frac{1}{2} n_i(n_i - 1) - p \right) \quad (13) \]

\[ \hat{\sigma}^2 = \left( \sum_{i=1}^{n} \sum_{j=1}^{n} r_{Y_{ij}}^2 \right) \left( \sum_{i=1}^{n} n_i - p - 1 \right) \quad (14) \]

\[ \hat{\rho}_Y = \hat{\sigma}^{-2} \left( \sum_{i=1}^{n} \sum_{j \geq j'} r_{Y_{ij} r_{Y_{ij}'}} \right) \left( \sum_{i=1}^{n} \frac{1}{2} n_i(n_i - 1) - p - 1 \right) \quad (15) \]

Similar to the likelihood equations in Section 2.2, we have

\[ \sum_{i=1}^{n} \left( \frac{\partial l_i}{\partial \beta_1} \right) = \sum_{i=1}^{n} \left( \frac{\partial E(X_i)}{\partial \beta_1} \frac{\partial E(X_i)}{\partial \alpha} \right)^T \text{cov}^{-1} \left( \begin{array}{c} X_i \\ Y_i | X_i \end{array} \right) \left( \begin{array}{c} x_i - E(X_i) \\ y_i - E(Y_i | X_i) \end{array} \right) \]

This yields the following equation system of GEE’s for $\beta_1$ and $\alpha$:

\[ \sum_{i=1}^{n} \left( \begin{array}{cc} \Delta_i Z_i & (\gamma_1 + \gamma_2) \Delta_i Z_i \\ 0 & W_i \end{array} \right) \left( \begin{array}{c} V_{1i}^{-1} \\ 0 \end{array} \right) \left( \begin{array}{c} X_i - \hat{\lambda}_i \\ Y_i - W_i^T \hat{\alpha} \end{array} \right) = 0 \quad (16) \]

Given a group of initial values of ($\beta_1, \alpha$), put them into (10)–(15) and then we can obtain estimates of ($\rho_X, \rho_Y, \sigma^2$), denoted by ($\hat{\rho}_X, \hat{\rho}_Y, \hat{\sigma}^2$), and then substitute them into (8) and (9) to obtain estimates of ($V_{1i}, V_{2i}$), finally into (16) and solve this equation to the updated values of $\hat{\beta}_1$ and $\hat{\alpha}$, the estimate of $\beta_1$ and $\alpha$. So we can obtain these estimates iteratively. Furthermore, it can be showed that the estimate ($\hat{\beta}_1, \hat{\alpha}$) is a consistent estimate and has an asymptotic multivariate normal distribution.

In addition, ‘model-based’ covariance matrix of parameters can be estimated by

\[ \Sigma_m = I_0^{-1} \quad (17) \]
where
\[
I_0 = \sum_{i=1}^{n} \left( \Delta_i Z_i \quad -(\gamma_1 + \gamma_2) \Delta_i Z_i \right) \left( \begin{array}{cc} V_{1i}^{-1} & 0 \\ 0 & V_{2i}^{-1} \end{array} \right) \left( \Delta_i Z_i \quad -(\gamma_1 + \gamma_2) \Delta_i Z_i \right)^T
\]

When \( \text{cov}(X_i) \neq V_{1i} \) or \( \text{cov}(Y_i|X_i) \neq V_{2i} \), this means the variance–covariance matrix is ‘misspecified’, we could use another robust method to estimate covariance
\[
\Sigma_c = I^{-1}_0 I_1 I^{-1}_0
\]

where
\[
I_1 = \sum_{i=1}^{n} \left( \begin{array}{cc} Z_i^T \Delta_i & -(\gamma_1 + \gamma_2) Z_i^T \Delta_i \\ 0 & W_i^T \end{array} \right) \left( \begin{array}{cc} V_{1i}^{-1} & 0 \\ 0 & V_{2i}^{-1} \end{array} \right) \text{cov}(X_i, Y_i)
\]

\[
\times \left( \begin{array}{cc} V_{1i}^{-1} & 0 \\ 0 & V_{2i}^{-1} \end{array} \right) \left( \begin{array}{cc} Z_i^T \Delta_i & -(\gamma_1 + \gamma_2) Z_i^T \Delta_i \\ 0 & W_i^T \end{array} \right)^T
\]

and \( \text{cov}(X_i, Y_i) \) can be replaced by \([X_i - \hat{\lambda}_i(\hat{\beta}_1), Y_i - W_i \hat{\hat{\lambda}}][X_i - \hat{\lambda}_i(\hat{\beta}_1), Y_i - W_i \hat{\hat{\lambda}}]^T\) in calculation.

Finally, consider testing problem
\[
H_0 : L\Theta = 0 \quad \text{versus} \quad H_1 : L\Theta \neq 0
\]

as noted by Zhang and Boos [10] and Rotnitzky and Jewell [11] and combining estimates (17) and (18), we can construct a statistic for test (19)
\[
T = S(\hat{\Theta})\Sigma_m L^T(L\Sigma_m L^T)^{-1} L\Sigma_m S(\hat{\Theta})^T
\]

where \( L \) is a user-defined \( k \times (2p+2) \) matrix, \( \Theta = (\beta_1, \beta_2, \gamma_1, \gamma_2) \), \( S(\hat{\Theta}) \) is the value of generalized estimating equation at \( \hat{\Theta} \), which is the regression parameter resulting from GEE under the restricted model \( L\Theta = 0 \). This statistic approximately follows the chi-square distribution \( \chi^2_k \) with \( k \) degree of freedom.

3. CORRELATION OVER TIME

One of the important problems in our study is concerned with the correlation of the two responses. We will answer question of how the correlation between Poisson and continuous response change over time and how covariates impact on these correlations. On the other hand, we should remove the impact of heterogenous variances upon the evaluation of the varying coefficient function in the regression model of the mixed responses. For these issues arising from inference of longitudinal mixed continuous and Poisson responses, we provide a rounded scenario to deal with them in this section.
3.1. Regression model for correlation

In this section we try to construct a regression model directly on the correlation coefficient. Since we construct regression models, we need to get a group of sample correlation coefficients at the points of observed times and a group of covariate variables. There are two possible ways: first, we obtain sample correlation coefficients based on the response at the observed time points of each unit, and then we obtain a group of correlation coefficients and the covariate variable group consisting exactly of the covariate variables of each unit. Although we can obtain an estimate of the population sample correlation, we would fail to observe the correlation change over time, which is the focus of our current work. So we adopt the second way: to obtain a sample correlation coefficient based on the response of all the units on each time point and form a group of the correlation coefficient at the observed time points. However, another issue may arise: how to choose the appropriate covariate variable group? In this situation at each time point we need to construct a covariate variable which apparently should contain the information of all the covariate variables of the units at this time point. A natural way is to take the average of the covariate variables at each time point and thus form our covariate variable group. For simplicity, we suppose the observed time points are regular, that is, we observe all the individuals at the given time points, which implies $n_1 = \ldots = n_N \equiv n$ and we set the time points are $t_1, \ldots, t_n$.

So we consider the following log-linear model:

$$
\log \frac{1 + r_{t_j}}{1 - r_{t_j}} = \bar{Z}_{t_j}^T b_1 + t_j b_2 + \epsilon_{t_j}, \quad 1 \leq j \leq n
$$

(21)

where $b_1$ and $b_2$ are the regression coefficients to be estimated, $r_{t_j}$ is the observed correlation obtained based on the responses $X_{it_j}, Y_{it_j}$ of all the units at a certain measure time (say $t_j$), that is

$$
r_{t_j} = \frac{\sum_{i=1}^N (X_{it_j} - \bar{X}_{t_j})(Y_{it_j} - \bar{Y}_{t_j})}{\sqrt{\sum_{i=1}^N (X_{it_j} - \bar{X}_{t_j})^2} \sqrt{\sum_{i=1}^N (Y_{it_j} - \bar{Y}_{t_j})^2}}
$$

$\bar{Z}_{t_j}$ is the average of the continuous covariates variables on time point $t_j$, i.e. $\bar{Z}_{t_j} = (1/N) \sum_{i=1}^N Z_{it_j}$ and $\epsilon_{t_1}, \ldots, \epsilon_{t_n}$ are independent random errors with mean zero and finite variances. The estimates of $b_1$ and $b_2$ in (21) can be obtained through the common least square method.

We can also fit the following non-linear correlation regression model:

$$
r_{t_j} = \frac{e^{\bar{Z}_{t_j}^T b_1 + t_j b_2}}{e^{\bar{Z}_{t_j}^T b_1 + t_j b_2} + 1} + \epsilon_{t_j}, \quad 1 \leq j \leq n
$$

(22)

Although non-linear least square iterative method will be used to find the estimate of $b_1$ and $b_2$ and will take more time than direct least square estimate in (21), our simulation results show that the former is more effective than the latter provided we choose the least square estimate in (21) as the initial values in (22) at the cost of longer computing time.

We further discuss model (21) for the correlation. Obviously, the model (21) is not applicable to the classification variables. If covariates $Z_{ij}$ is continuous, it seems reasonable to take average of all the covariates of individuals at any given time. However, this could cause two major problems: first, the averaging is perhaps not the best method to summarize the information of all the units. Alternative choice of covariates in (21) is possible, we take some sufficient statistics as...
the covariates of model (21). The further research is expected. Second, even more importantly, averaging method choosing covariates in model (21) may not be applicable if covariates include discrete, binary or categorical data and we will investigate this problem in the future.

3.2. Covariance structure of the random error

Since the correlation of the two responses is related to the covariance structure of the random errors, it may be helpful to focus on the covariance structure changing over time. So we change the joint model (7) into varying-coefficient model [12]

\[
\begin{align*}
\log (x_{ij} + 1) &= Z_{ij1}\beta_1(t_j) + \epsilon_{ij1} \\
y_{ij} &= Z_{ij2}\beta_2(t_j) + \epsilon_{ij2}
\end{align*}
\]

(23)

where independent random errors \( \epsilon_{ij} = (\epsilon_{ij1}, \epsilon_{ij2})^T \sim N(0, \Sigma(t_j)) \)

\[
\Sigma(t_j) = \begin{pmatrix}
\sigma_1^2(t_j) & \rho(t_j)\sigma_1(t_j)\sigma_2(t_j) \\
\rho(t_j)\sigma_1(t_j)\sigma_2(t_j) & \sigma_2^2(t_j)
\end{pmatrix}
\]

is the population covariance structure at time point \( t_j, i = 1, 2, \ldots, m_j, j = 1, 2, \ldots, M, m_j \) denotes the number of the individuals at the time point \( t_j \), and \( M \) denotes the number of the observed reforming time points; \( Z_{ijk} = (Z_{ij1k}, \ldots, Z_{ijk}) \), \( \beta_k(t_j) = (\beta_{k1}(t_j), \beta_{k1}(t_j), \ldots, \beta_{kp}(t_j))^T, k = 1, 2 \). The correlation function \( \rho(\cdot) \) could reflect how correlation between Poisson and continuous responses change over time. At the same time, we can check whether the variance functions \( \sigma_1^2(\cdot) \) and \( \sigma_2^2(\cdot) \) change over time.

The covariance matrix function \( \Sigma(\cdot) \) can be estimated using two-step methods.

1. First step: for any fixed time point \( t_j \), we use ordinary least square methods to estimate \( \hat{\beta}_1(t_j) \) and \( \hat{\beta}_2(t_j) \) based on the observations of all individuals at the given time, and then compute the residuals \( \hat{\epsilon}_{ij}, i = 1, \ldots, m_j \). The estimate of \( \Sigma(t) \) at \( t = t_j \) is defined by

\[
\hat{\Sigma}(t_j) = \frac{1}{N - p - 1} \sum_{i=1}^{m_j} \hat{\epsilon}_{ij}\hat{\epsilon}_{ij}^T
\]

2. Second step: using the standard smoothing technique, e.g. kernel estimate, smoothing spline or local polynomial, we can smooth all the components \( \sigma_1(t), \sigma_2(t) \) and \( \rho(t) \) of \( \Sigma(t) \), respectively.

In order to check whether the variance function \( \Sigma(t) \) changes over time, we want to test the null hypothesis

\[
H_0 : \Sigma(t_j) = \Sigma_0 \quad \text{for all} \quad j = 1, \ldots, M
\]

(24)

where

\[
\Sigma_0 = \begin{pmatrix}
\sigma_{01}^2 & \rho_0\sigma_{01}\sigma_{02} \\
\rho_0\sigma_{01}\sigma_{02} & \sigma_{02}^2
\end{pmatrix}
\]
is a constant positive matrix, against the alternative

\[ H_1 : \Sigma(t_{j_1}) \neq \Sigma(t_{j_2}) \quad \exists \ j_1 \neq j_2 \in \{1, \ldots, M\} \]

For this testing problem, the likelihood ratio test is employed. The likelihood ratio statistic is defined by

\[
LR = \sup_{\theta \in \Theta_0} L(\theta|X, Y, Z) / \sup_{\theta \in \Theta} L(\theta|X, Y, Z) = \frac{[\hat{\sigma}_{01}^2 \hat{\sigma}_{02}^2 (1 - \hat{\rho}_0^2)]_{j=1}^M j_{1/2}}{\prod_{j=1}^M [\hat{\sigma}_{1j}^2(t_j) \hat{\sigma}_{2j}^2(t_j) (1 - \hat{\rho}_j^2(t_j))]^{-m_{j/2}}} \tag{25}
\]

where \( L(\theta|X, Y, Z) \) denotes the likelihood function, \( \theta = (\theta(t_1), \theta(t_2), \ldots, \theta(t_M)) \), \( \hat{\theta}(t_j) = (\beta_1(t_j)^T, \beta_2(t_j)^T, \sigma_1(t_j), \sigma_2(t_j), \rho(t_j)) \), \( X = (x_1, x_2, \ldots, x_M), \ x_j = (x_{1j}, \ldots, x_{mj}), \ Y = (y_1, y_2, \ldots, y_M), \ y_j = (y_{1j}, y_{2j}, \ldots, y_{mj}), \ Z = (Z_1, Z_2, \ldots, Z_M), \ Z_j = (Z_{1j}, Z_{2j}), \ Z_{j1} = (Z_{1j1}, \ldots, Z_{mj1}), \ Z_{j2} = (Z_{1j2}, \ldots, Z_{mj2}); \ \Theta \) denotes the whole parameter space and \( \Theta_0 \) denotes the parameter space under \( H_0 \). The formulae for calculating MLE \( \hat{\Sigma}(t_j), j = 1, 2, \ldots, M \), and \( \hat{\Sigma}_0 \) are given in Appendix A.

Under \( H_0 \), the statistic \(-2 \log LR\) asymptotically follows \([13]\) the chi-square distribution \( \chi_r^2 \) with \( r = (3M + P_1 + P_2) - (3 + P_1 + P_2) = 3(M - 1) \) degree of freedom, where \( 3M + P_1 + P_2 \) is dimension of \( \Theta \) and \( 3 + P_1 + P_2 \) is dimension of \( \Theta_0 \), \( P_i \), \( i = 1, 2 \) is the dimension of \( \beta_i(t_j) \).

Once the null hypothesis \( H_0 \) is rejected, which means that the variance function \( \Sigma(t) \) does change over time, we have to consider how to get rid of its impact on the estimate of \( \beta_i(t_j), i = 1, 2; j = 1, 2, \ldots, M \) and on the significant test of covariates. Data transformation is a simple and effective method to deal with this problem. Under the framework of model (23), we propose to carry out linear transformation using matrix \( \Sigma(t_j)^{-1/2} \) for both two responses and their correspondent covariates at the time point \( t_j \), which can transform the varying variance structures of random errors at different time points into a uniform variance structure unchanging over time. Let \( x_{ij}^*, y_{ij}^*, Z_{ij1}^*, Z_{ij2}^* \) denote transformed response variables and covariates of the \( i \)th individual, respectively, and the relationship between the transformed and original data are

\[
\begin{pmatrix}
  x_{ij}^* \\
  y_{ij}^*
\end{pmatrix} = \Sigma(t_j)^{-1/2} \begin{pmatrix}
  \log(x_{ij} + 1) \\
  y_{ij}
\end{pmatrix}, \quad \begin{pmatrix}
  Z_{ij1}^* \\
  Z_{ij2}^*
\end{pmatrix} = \Sigma(t_j)^{-1/2} \begin{pmatrix}
  Z_{ij1} \\
  Z_{ij2}
\end{pmatrix} \tag{26}
\]

After applying the transformation (26) model (23), we have the transformed model

\[
x_{ij}^* = Z_{ij1}^* \beta_1(t_j) + \epsilon_{ij1}^* \quad \text{and} \quad y_{ij}^* = Z_{ij2}^* \beta_2(t_j) + \epsilon_{ij2}^* \tag{27}
\]

where independent random errors \( \epsilon_{ij}^* = (\epsilon_{ij1}^*, \epsilon_{ij2}^*)^T \sim N(0, I_2) \), \( I_2 \) is a \( 2 \times 2 \) identity matrix.

Now we can test the significance of each covariate based on model (27). For given \( k = 1, 2, \ p = 1, 2, \ldots, P \), consider the null hypothesis

\[ H_0 : \beta_{kp}(t_j) = 0 \quad \text{for all} \ j = 1, \ldots, M \tag{28} \]

versus

\[ H_1 : \beta_{kp}(t_j) \neq 0 \quad \exists \ j \in \{1, \ldots, M\} \]

where \( \beta_k(t) = (\beta_{k1}(t), \beta_{k2}(t), \ldots, \beta_{kp}(t))^T, k = 1, 2 \).
Construct a likelihood ratio test statistic as follows:

\[
LR_{k,p} = \begin{cases} 
\frac{\exp \left( -\sum_{j=1}^{M} \frac{1}{2} \| x_j^* - \hat{Z}_{j1[-p]}^* \hat{\beta}_{1[-p]}(t_j) \|^2 \right)}{\exp \left( -\sum_{j=1}^{M} \frac{1}{2} \| x_j^* - \hat{Z}_{j1} \hat{\beta}_1(t_j) \|^2 \right)}, & k = 1 \\
\frac{\exp \left( -\sum_{j=1}^{M} \frac{1}{2} \| y_j^* - \hat{Z}_{2[-p]}^* \hat{\beta}_{2[-p]}(t_j) \|^2 \right)}{\exp \left( -\sum_{j=1}^{M} \frac{1}{2} \| y_j^* - \hat{Z}_{j2} \hat{\beta}_2(t_j) \|^2 \right)}, & k = 2
\end{cases}
\] (29)

where the MLE of \( \beta_{k[-p]}(t_j) = (\beta_{k1}, \ldots, \beta_{k(p-1)}, \beta_{k(p+1)}, \ldots, \beta_{kp})^T \) can be obtained by

\[
\hat{\beta}_{1[-p]}(t_j) = (Z_{j1[-p]}^* Z_{j1[-p]}^*)^{-1} Z_{j1[-p]}^* x_j^*
\]
\[
\hat{\beta}_{2[-p]}(t_j) = (Z_{j2[-p]}^* Z_{j2[-p]}^*)^{-1} Z_{j2[-p]}^* y_j^*
\]

which are the estimates of \( \beta_1(t_j) \) and \( \beta_2(t_j) \) under the null hypothesis and \( Z_{jk[-p]}^* = (Z_{jk1}, \ldots, Z_{jk(p-1)}^*, Z_{j(k+1)}^*, \ldots, Z_{jkp})^T \), the dot in the norm is summation for all \( i \). The definitions of \( x_j^*, y_j^* \), etc. are similar to those of \( x_j, y_j \), e.g. \( x_j^* = (x_{j1}^*, \ldots, x_{jm_j}^*) \), etc.

Similar to the previous asymptotic distribution of the likelihood ratio test statistic for checking variance whether changes over time, \(-2\log LR_{k,p}\) asymptotically follows the chi-square distribution \( \chi^2_r \) with \( r = M \) degree of freedom for all given \( k \) and \( p \).

4. SIMULATION

In order to demonstrate the performance of the proposed procedures in Sections 2 and 3, small sample simulation experiments were done under a variety of parameter combinations.

In model (7) we take the following test parameters:

\[
\beta_1 = (0.5, 0.5, -0.8, 0, 0.2, -0.5, 0.7, 0.9, -0.1, -0.5)'
\]
\[
\beta_2 = (0.9, -0.5, 1.2, -0.5, 2, -1.5, 0.7, 0.9, -1, -0.5)'
\]
\[
\gamma_1 = 2, \quad \gamma_2 = 0.1, \quad m_i = 50, \quad M = 20
\]

and the design matrix \( Z \) with dimension \( 10 \times 50 \) reshaped by 500 normal random variable with zero mean and unit variance. We repeat 200 replications and show the bias of the estimates of \( \beta_1, \beta_2, \gamma_1, \gamma_2 \) in Figure 1.

For joint model (23) with time-varying correlation and variances, a Monte Carlo study was done under a variety of regression parameter, design matrix \( Z \), variance functions and correlation function changing over time. The results indicate good performance of the proposed methods. We only show a part of the fitted results in Figure 2. Here we take 50 equally spaced time points \( t = \{1, 50, \frac{1}{50}, \ldots, 1\} \), the variance functions \( \sigma_1(t) = (1 + t)^2, \sigma_2(t) = \sqrt{2 - t^2} \) and correlation function \( \rho(t) = \frac{1}{2}(t + \sin(2\pi t)) \), true \( \beta_1, \beta_2 \) and the design matrix \( Z \) as the beginning of this section.
We also conduct the simulation to evaluate the power of the proposed test (24) for checking the variability of the variance. We, respectively, set variance function $\Sigma(t)$ as the following four cases: case 1. $\sigma_1(t) = \sigma_2(t) = 1$, $\rho(t) = 0$; case 2. $\sigma_1(t) = (1+t)^2$, $\sigma_2(t) = \sqrt{2 - t^2}$, $\rho(t) = 0$; case 3. $\sigma_1(t) = \sigma_2(t) = 1$, $\rho(t) = \frac{1}{2}(t + \sin(2\pi t))$; case 4. $\sigma_1(t) = (1 + t)^2$, $\sigma_2(t) = \sqrt{2 - t^2}$, $\rho(t) = \frac{1}{2}(t + \sin(2\pi t))$. And repeat 150, 50, 50, 50 replications for case 1, case 2, case 3, case 4, respectively. The test power we obtain is 0.9430, which indicates that the test has good performance to check the variability of variance.
Similarly, we also do the simulation to validate the power of the test (28) and we obtain the power as 0.9820, which is good to validate the significance of covariates.

5. APPLICATION TO THE ICDB DATA

In this section we apply the models and estimation methods described in Sections 2 and 3 to the ICDB data [1] mentioned in the Introduction. A total of 637 eligible patients were entered into the study and followed for symptoms of pain \( (p) \), urgency \( (u) \) and urinary frequency \( (f) \). In those responses, it is reasonable to regard urinary frequency as Poisson random variable. Some of the research results are reported in [1, 14, 15]. After deleting missing data from the data set, 611 patients with 15 covariates were entered into our data analysis. For the baseline demographic characteristics of 611 patients, see Table I, which are almost similar to Table II in [1]. The definitions of the covariates are given in Table II.

First, we use the method in Section 2.3 to estimate parameters \( \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_1, \hat{\gamma}_2 \). The Poisson response is urinary frequency, and the continuous ones are pain score or urinary urgency. We have obtained two groups of results for \( (p, f) \) and \( (u, f) \). Here we only report the results for \( (u, f) \) because of the similarity of the results of two groups. The covariate ‘severity’ is a three-level categorical variable, so we introduce dummy variables to deal with it. The estimates of the parameters \( \gamma_1 \) and \( \gamma_2 \) are 0.258 and 0 and their corresponding \( p \)-values are 0.07 and 0.06, respectively. The estimates of the parameters \( \beta_1 \) and \( \beta_2 \) and their corresponding \( p \)-values are listed in Table I.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No. patients (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>54(8.84)</td>
</tr>
<tr>
<td>F</td>
<td>557(91.16)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>569(93.13)</td>
</tr>
<tr>
<td>Other</td>
<td>42(6.87)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Partnered</td>
<td>430(70.38)</td>
</tr>
<tr>
<td>Alone</td>
<td>181(29.62)</td>
</tr>
<tr>
<td><strong>Employment</strong></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>372(60.88)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>68(11.13)</td>
</tr>
<tr>
<td>Home/Retired</td>
<td>171(27.99)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>High School or less</td>
<td>260(42.55)</td>
</tr>
<tr>
<td>College or Advanced</td>
<td>351(57.45)</td>
</tr>
<tr>
<td><strong>Annual household income ($)</strong></td>
<td></td>
</tr>
<tr>
<td>Less than 30,000</td>
<td>175(28.64)</td>
</tr>
<tr>
<td>30,000 or Greater</td>
<td>436(71.36)</td>
</tr>
<tr>
<td><strong>Previous interstitial cystitis diagnosis by physician</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>413(67.59)</td>
</tr>
<tr>
<td>No</td>
<td>198(32.41)</td>
</tr>
</tbody>
</table>

Table I. Some baseline demographic characteristics of 611 patients.
Table II. The definition and type of parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>Continuous</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex: M/F</td>
<td>Binary</td>
</tr>
<tr>
<td>Income</td>
<td>Annual household income: Less than/more than 30000$</td>
<td>Binary</td>
</tr>
<tr>
<td>shx_2</td>
<td>Previous interstitial cystitis diagnosis by physician: yes/no</td>
<td>Binary</td>
</tr>
<tr>
<td>urod_7</td>
<td>Volume at first sensation</td>
<td>Continuous</td>
</tr>
<tr>
<td>urod_9</td>
<td>Volume at maximal capacity</td>
<td>Continuous</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
<td>Continuous</td>
</tr>
<tr>
<td>Severity</td>
<td>Severity of symptoms</td>
<td>Categorical:3</td>
</tr>
</tbody>
</table>

Table III. Parameter estimates of \((u, f)\).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(\hat{\beta}_1)</th>
<th>(p)</th>
<th>(\hat{\beta}_2)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>-0.438</td>
<td>0.00</td>
<td>-0.380</td>
<td>0.07</td>
</tr>
<tr>
<td>Income</td>
<td>0.003</td>
<td>0.96</td>
<td>-0.453</td>
<td>0.00</td>
</tr>
<tr>
<td>shx_2</td>
<td>0.217</td>
<td>0.00</td>
<td>0.161</td>
<td>0.29</td>
</tr>
<tr>
<td>urod_7</td>
<td>-0.001</td>
<td>0.00</td>
<td>-0.002</td>
<td>0.26</td>
</tr>
<tr>
<td>urod_9</td>
<td>-0.002</td>
<td>0.00</td>
<td>-0.001</td>
<td>0.20</td>
</tr>
<tr>
<td>Age</td>
<td>0.021</td>
<td>0.00</td>
<td>0.003</td>
<td>0.56</td>
</tr>
<tr>
<td>Severity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sev_1</td>
<td>-0.045</td>
<td>0.86</td>
<td>3.725</td>
<td>0.00</td>
</tr>
<tr>
<td>sev_2</td>
<td>0.436</td>
<td>0.07</td>
<td>5.406</td>
<td>0.00</td>
</tr>
<tr>
<td>sev_3</td>
<td>0.934</td>
<td>0.00</td>
<td>6.656</td>
<td>0.00</td>
</tr>
</tbody>
</table>

in Table III. The \(p\) values of these variables are the results of hypothesis tests whether each level term is zero.

From Table III we can see that covariates ‘sex’, ‘shx_2’, ‘urod_7’, ‘urod_9’, ‘age’ and ‘severity’ are significant under the same size in the estimation of \(\hat{\beta}_1\), and covariates ‘sex’, ‘income’ and ‘severity’ except the first level are significant under the same size in the estimation of \(\hat{\beta}_2\). Thus, we can conclude that above selected covariates are more important to the origin of the disease.

As what we did in Section 3.2, we can estimate the correlation over time of the pair \((u, f)\). Using the algorithm given in Section 3.2, the estimates of variance functions and correlation function are shown in Figure 3.

From these figures we can see that the correlation between response pair \((u, f)\) obviously increase over time. And we employ the test (24) to validate the significance of the covariance structure changing over time and obtain \(p\) value as \(1.7529 \times 10^{-7}\), which is powerful to reject the \(H_0\) in (24) and gives a statistical interpretation for the changing covariance structure. These findings suggest us to take account of covariance structure changing over time into the practical modelling. The solid lines in Figure 3(a) were obtained by kernel estimate method in which the bandwidth was chosen based on least square cross-validation technique. From Figure 3(b) and (c),
Figure 3. (a) The correlation change over time for \((u, f)\). The circles denote the observed correlation coefficient, the real line denotes kernel estimate, dashed line denotes the estimated \(\rho\), and the dotted line denotes the sample correlation coefficient after taking log on frequency. (b) and (c) are the estimates of variance functions \(\sigma_1(t)\) and \(\sigma_2(t)\), respectively. The dashed line denotes the estimate of variance; the dotted line denotes the smoothed variance based on the least square cross-validated kernel estimates.

Table IV. The significance levels of covariates \(\beta_1(t)\) and \(\beta_2(t)\) based on test (28).

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\beta}_1(t))</th>
<th>(\hat{\beta}_2(t))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>0.00</td>
<td>0.61</td>
</tr>
<tr>
<td>Income</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>shx.2</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>urod.7</td>
<td>0.00</td>
<td>0.48</td>
</tr>
<tr>
<td>urod.9</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>Age</td>
<td>0.02</td>
<td>0.86</td>
</tr>
<tr>
<td>Severity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sev.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>sev.2</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>sev.3</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

we can see that the variances \(\sigma_1(t)\) and \(\sigma_2(t)\) decrease with time. Intuitively, we cannot ignore the change of variance over time when we make inference for significance of the covariates.

Furthermore, the model in Section 3.1 could be used to investigate the correlation changing over time. However, as we mentioned the way in that section fails to provide a satisfactory statistical interpretation if there exist discrete variables in model. So it is probably not desirable for our data set and we will not employ it to analyse our data.

Hence, we use the data transformation technique proposed in Section 3.2 and based on the new data to make inference for significance of the covariates under the framework of model (23), which can eliminate the impact of the covariance changing over time.
Based on test (28), we also can give the significance of covariate over time. Like Table III, we can see from Table IV that covariates ‘sex’, ‘shx_2’, ‘urod_7’, ‘urod_9’, ‘age’ and ‘severity’ are significant under the same size in the estimation of $\beta_1$, and covariates ‘income’, ‘shx_2’ and ‘severity’ are significant under the same size in the estimation of $\beta_2$. Compared with Table III, we find the results based on test (28) are slightly different from the results of test (19), which does not consider the covariance changing over time. First, the covariate ‘sex’ of $\beta_2$ is not significant in test (28) but significant in test (19). Second, the covariate ‘shx_2’ of $\beta_2$ is significant in test (28) but not in test (19). The differences we list above indicate that the covariance changing over time to some extent has impact on the significance of the covariates. The estimates of varying-coefficient function component of $\beta_1(t)$ and $\beta_2(t)$, respectively, for the covariates sex, income, shx_2, urod_7, urod_9, age, severity with three levels sev_1, sev_2 and sev_3 are shown in Figure 4. From Figure 4 we can also observe the trends of varying-coefficient function component of $\beta_1(t)$ and $\beta_2(t)$, respectively, for the covariates sex, income, shx_2, urod_7, urod_9, age, severity with three levels sev_1, sev_2 and sev_3.
\( \beta_2(t) \) changing over time. The model and test in Section 3.2 is an effective method to complement the model in Section 2.

6. DISCUSSION

In this article we build regression models for mixed Poisson and continuous responses. We focus on the regression models and modelling correlation over time through joint linear regression with time-varying variance–covariance structure. The regression parameter and variance–covariance matrix are estimated by the GEE methodology and the two step methods. We first assume the marginal distribution of the discrete variable and then consider the conditional distribution of the continuous one. The Poisson response is treated as a log-linear model \[16\], which possesses a desirable advantage that regression parameters have marginal interpretation. Furthermore, this method is robust, that is, even the link function of \( X \) and \( Y \) is misspecified, the other parameters can still be well estimated. In our simulation we have shown that even the parameters \( \gamma_1 \) and \( \gamma_2 \) in model (7) are mistaken, the estimation of \( \beta_1 \) and \( \beta_2 \) can still work well. From the results in Section 5, it is easy to see that the variances change over time and we should be careful to make inference based on the models in Section 2. When we emphasize particularly on building relationship between the responses and covariates, we should eradicate the changing variance. The methods given in Section 3.2 accommodate changing variance and correlation. To this end, we firstly employ the likelihood ratio test to check whether the covariance changes over time and we use data transformation technique further such that the covariance structure remains unchanging if we reject the null hypothesis covariance keeps unchanging. Fieuws and Verbeke [17] discussed the problem how the associations between the hearing thresholds for the frequencies evolve over time. They did not fit the observed marginal correlation in their Figure 2 very well, the most possible reason is that they did not take into account the heterogeneous variance and covariance. If we could get their data used in their Figure 2, we might give a satisfactory conclusion using our methods developed in this article.

Although the maximum likelihood method is an effective way to analyse data, it could not be applied to longitudinal data, the main reason is that the independence assumption does not hold in this situation. So the GEE methodology, which only requires the moments of the variables, is employed. Compared with the maximum likelihood method, this GEE methodology has even higher efficiency and can guarantee the robustness of the result.

As to the problem arising from research of IC, we apply our model to the real data. From the results we can determine which covariates variables have heavy influence on the disease, and further we find the increasing correlation changes over time.

APPENDIX A: LIKELIHOOD RATIO TEST FOR VARIANCE CHANGING OVER TIME

The likelihood function of model (23) is

\[
L(\theta|X, Y, Z) = \prod_{j=1}^{M} f(\log(X_j + 1), Y_j|\theta, Z_j)
\]
Table A1. Maximum likelihood estimate of $\theta$.

<table>
<thead>
<tr>
<th>$\theta \in \Theta$</th>
<th>$\theta \in \Theta_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\alpha}_1(t_j) = \sqrt{\frac{| \log(X_j+1) - Z_j \hat{\beta}_1(t_j) |^2}{m_j}}$</td>
<td>$\hat{\alpha}<em>{01} = \sqrt{\frac{\sum</em>{j=1}^{M} | \log(X_j+1) - Z_j \hat{\beta}<em>1(t_j) |^2}{\sum</em>{j=1}^{M} m_j}}$</td>
</tr>
<tr>
<td>$\hat{\alpha}_2(t_j) = \sqrt{\frac{| Y_j - Z_j \hat{\beta}_2(t_j) |^2}{m_j}}$</td>
<td>$\hat{\alpha}<em>{02} = \sqrt{\frac{\sum</em>{j=1}^{M} | Y_j - Z_j \hat{\beta}<em>2(t_j) |^2}{\sum</em>{j=1}^{M} m_j}}$</td>
</tr>
<tr>
<td>$\hat{\rho}(t_j) = \frac{(\log(X_j+1) - Z_j \hat{\beta}_1(t_j))^T(Y_j - Z_j \hat{\beta}_2(t_j))}{m_j \sigma_1(t_j) \sigma_2(t_j)}$</td>
<td>$\hat{\rho}<em>0 = \frac{\sum</em>{j=1}^{M} (\log(X_j+1) - Z_j \hat{\beta}<em>1(t_j))^T(Y_j - Z_j \hat{\beta}<em>2(t_j))}{\sum</em>{j=1}^{M} m_j \hat{\alpha}</em>{01} \hat{\alpha}_{02}}$</td>
</tr>
</tbody>
</table>

\[
\hat{\beta}_1(t_j) = (Z_j^T Z_j)^{-1} Z_j^T \log(X_j+1)
\]

\[
\hat{\beta}_2(t_j) = (Z_j^T Z_j)^{-1} Z_j^T Y_j
\]

\[
= \prod_{j=1}^{M} \left( 2\pi \sigma_1^2(t_j) \sigma_2^2(t_j) (1 - \rho^2(t_j)) \right)^{-m_j/2} \exp \left\{ \frac{\sum_{j=1}^{M}}{2(1 - \rho^2(t_j)) \sigma_1^2(t_j)} \left[ \frac{\| \log(X_j+1) - Z_j \beta_1(t_j) \|^2}{m_j} - \frac{\| Y_j - Z_j \beta_2(t_j) \|^2}{m_j} \right] \right\}
\]

where $f(\cdot; \cdot; \theta, Z_j)$ denote the joint density function of random vector $(\log(X_j+1), Y_j)$ at time point $t_j$.

Based on likelihood function (A1), we can calculate the MLE when $\theta \in \Theta$ and $\theta \in \Theta_0$, respectively, (Table A1).

According to Table A1, we can obtain the likelihood ratio statistic in (25).

ACKNOWLEDGEMENTS

The first author wishes to thank Prof. Kathleen Joy Propert at University of Pennsylvania School of Medicine for providing the data form the interstitial cystitis data base. The first author is also grateful to Prof. Wensheng Guo’s discussion. The authors thanks Prof. Michael Campbell, Editor, anonymous referees for their insightful suggestions and careful reading that greatly improved this article.

REFERENCES


DOI: 10.1002/sim