Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
STA2014-1255

Parametric versus Semi-parametric Mixed Models for Panel Count Data

Vandna Jowaheer
Associate Professor
University of Mauritius
Mauritius

Brajendra Sutradhar
Professor
Memorial University of Newfoundland
Canada

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Jowaheer, V. and Sutradhar, B., (2014) **"Parametric versus Semi-parametric Mixed Models for Panel Count Data",** Athens: ATINER'S Conference Paper Series, No: **STA2014-1255.**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr
URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

ISSN: **2241-2891**
18/09/2014

Parametric versus Semi-parametric Mixed Models for Panel Count Data

Vandna Jowaheer
Associate Professor
University of Mauritius
Mauritius

Brajendra Sutradhar
Professor
Memorial University of Newfoundland
Canada

Abstract

Panel count data analysis is an important research topic in socio-economic and biomedical fields, among others. Recently, Sutradhar, Jowaheer and Rao (2014) have demonstrated that the so-called GQL (generalized quasilikelihood) approach is more efficient than other existing popular approaches in estimating the regression parameters involved in the fully specified regression function of the mixed model for panel count data. There are however situations, where fully specified regression function may not be enough to explain the mean response of the counts. This motivates the use of a semiparametric regression function in such a mixed model, but the estimation of the parameters would be cumbersome as compared to the estimation of the parameters for parametric mixed model. In this paper, we examine the model mis-specification effects on the estimation of the regression and over-dispersion parameters by using the GQL approach. Consistent estimation for the proposed semi-parametric model is also provided.

Keywords: Conditional auto-correlations; Kernel based semi-parametric generalized quasi-likelihood estimation; Non-parametric function; Random effects; Semi-parametric mixed effects regression model.

Introduction

When repeated counts such as patent counts awarded to an industry along with associated covariates such as type of firm, book value of the capital, and time dependent research and development (R & D) expenditures are collected over a small period of time, it may be of interest to study the influence of the covariates on the repeated count responses after taking the correlations of the repeated responses into account. Over the last three decades, this type of panel count data has been analyzed by many econometricians by fitting conditional random effects model with so-called conditional maximum likelihood (CML) approach (Wooldridge (1999, eqn. (2.6), p. 79)), and the instrumental variables based generalized method of moments (IVGMM) (Montalvao (1997, eqn. (32), p. 85)) approaches. See also Hausman, Hall and Griliches (1984), Blundell, Griffith and Windmeijer (1995) for similar inferences. These random effects based mixed models are developed by using the so-called specified regression function. When the random effects are assumed to have normal distribution (Breslow and Clayton (1993), Jiang (1998), Sutradhar (2004)), the specified regression function becomes fully parametric.

In some practical situations, Poisson-normal mixture based mixed model may, however, not satisfactorily fit the longitudinal count data. As a remedy, in this paper we generalize this mixed model to the so-called semi-parametric mixed model, where a smooth non-parametric function in time is added to the specified linear predictor, in order to explain the mean and the variance of the data well. This new semi-parametric dynamic mixed model (SDMM) along with its basic correlation properties are given in Section 2. Note that the proposed SDMM may also be treated as a generalization of the semiparametric dynamic fixed model (SDFM) in the longitudinal count data setup recently studied by Sutradhar and Warriyar (2014). For various ‘working’ correlations based inferences for the SDFM, we refer to Severini and Staniswalis (1994, Section 8), Lin and Carroll (2001), Fan et al. (2007), Fan and Wu (2008), and Hua (2010), among others. Because the non-parametric function involved in the regression function makes the estimation of the covariate effects and variance component of the random effects, in Section 3, we examine the effect of ignoring the non-parametric function on the estimation of these two parameters. In Section 4, we provide an outline for the estimation of the parameters of the SDMMs. The paper concludes in Section 5.

Auto-correlations for Semi-parametric Mixed Model

Auto-correlations for Prametric Mixed Model for Count Data

Let t_{ij} denote the time at which the jth (j = 1,...,n_i) count response is recorded from the ith (i = 1,...,I) individual, and y_{ij} denote this response. Next suppose that $y_i = (y_{i1},...,y_{ij},...,y_{ini})^0$ denotes the $n_i \times 1$ vector of repeated counts for the ith (i = 1,...,K) individual. Also suppose that these repeated count responses are influenced by a fixed and known $p \times n_i$

covariate matrix $X_i^0 = (x_{i1}(t_{i1}),...,x_{ij}(t_{ij}),...,x_{in_i}(t_{in_i}))$, $x_{ij}(t_{ij})$ being the p−dimensional covariate vector for the ith individual at time point t_{ij}, as well as by an individual random effect γ_i^*. Further because the individual random effect remains the same during the data collection period, it is reasonable to assume that conditional on this individual random effect γ_i^*, any two responses of the ith individual collected at two different times, say y_{ij} and y_{ik} for $j < k$, will be serially correlated. We assume that conditional on the random effects, the repeated count responses follow an AR(1) (auto-regressive order 1) Poisson model [see Sutradhar (2003, Section 3) for details on this model, also see Sutradhar (2011, Chapter 8)]. The conditional mean, variance and correlations are given as follows:

$$E[Y_{ij}|\gamma_i^*] = \text{var}[Y_{ij}|\gamma_i] = m_{ij}^* = \exp(x'_{ij}(t_{ij})\beta + \gamma_i^*), \quad (1)$$

and for $j < k$, the lag $(k-j)$ correlation conditional on γ_i^* has the formula

$$\text{Corr } [Y_{ij}, Y_{ik}|\gamma_i] = \rho^{k-j}\sqrt{\frac{m_{ij}^*}{m_{ik}^*}}, \quad (2)$$

which is free from γ_i^*.
Next because $x^0_{ij}(t_{ij})\beta + \gamma_i^*$ is an additive function in a marginal log linear model, it is reasonable to assume that $\gamma_i^* \overset{iid}{\sim} N(0, \sigma_\gamma^2)$ [Breslow and Clayton (1993), Jiang (1998), Sutradhar (2004)]. One may then show that the unconditional mean and variance of y_{ij}, and the correlation between y_{ij} and y_{ik} have
the formulas

$$E[Y_{ij}] = m_{ij} = \exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2]$$

$$\text{var}[Y_{ij}] = v_{i,jj}(\beta, \sigma_\gamma^2) = m_{ij} + [\exp(\sigma_\gamma^2) - 1]m_{ij}^2, \quad (3)$$

and

$$\text{Corr } (Y_{ij}, Y_{ik}) = \frac{\rho^{j-k}(\frac{1}{m_{ik}}) + [\exp(\sigma_\gamma^2) - 1]}{\left[\left\{[\exp(\sigma_\gamma^2) - 1] + \frac{1}{m_{ij}}\right\}\left\{[\exp(\sigma_\gamma^2) - 1] + \frac{1}{m_{ik}}\right\}\right]^{\frac{1}{2}}}, \quad (4)$$

respectively.

Auto-correlations for Semi-parametric Mixed Model for Count Data

As mentioned in the last section, in some situations in practice, the specified linear mixed function $x'_{ij}(t_{ij})\beta + \gamma_i^*$ may not be enough to explain the data, particularly the mean and variance of the data. To address this issue, it may be appropriate to add a non-parametric function $\psi(t_{ij})$ to the linear mixed function. Consequently, one may easily generalize the Poisson AR(1) mixed model to the semi-parametric mixed model, simply by substituting the linear mixed predictor $x^0_{ij}(t_{ij})\beta + \gamma_i^*$ with $x^0_{ij}(t_{ij})\beta + \gamma_i^* + \psi(t_{ij})$. Thus, as opposed to the formula in (1), under the proposed semi-parametric mixed model, the conditional mean and variance have the formulas

$$E[Y_{ij}|\gamma_i^*] = \mu_{ij}^*(x_{ij}\beta; \gamma_i^*, \psi(t_{ij})) = \exp(x'_{ij}(t_{ij})\beta + \gamma_i^* + \psi(t_{ij}))$$

$$= \text{var}[Y_{ij}|\gamma_i] = \sigma_{i,jj}^*(x_{ij}\beta; \gamma_i^*, \psi(t_{ij})), \quad (5)$$

and for $j < k$, as opposed to (2), the lag $(k - j)$ correlation conditional on γ_i^* has the formula

$$\text{Corr}\,[Y_{ij}, Y_{ik}|\gamma_i] = \rho^{k-j}\sqrt{\frac{\mu_{ij}^*(x_{ij}\beta\;;\gamma_i^*, \psi(t_{ij}))}{\mu_{ik}^*(x_{ik}\beta\;;\gamma_i^*, \psi(t_{ik}))}}, \quad (6)$$

which is free from γ_i^*.

Next by assuming that $\gamma_i^* \overset{iid}{\sim} N(0, \sigma_\gamma^2)$, by similar calculations as in (3)(4), one may show that the unconditional mean and variance of y_{ij}, and the correlation between y_{ij} and y_{ik} have the formulas

$$E[Y_{ij}] = \mu_{ij}(x_{ij}\beta\;, \sigma_\gamma^2; \psi(t_{ij})) = \exp[x'_{ij}(t_{ij}\beta\; + \sigma_\gamma^2/2 + \psi(t_{ij})]$$
$$\text{var}[Y_{ij}] = \sigma_{i,jj}(x_{ij}\beta\;, \sigma_\gamma^2; \psi(t_{ij}))$$
$$= \mu_{ij}(x_{ij}\beta\;, \sigma_\gamma^2; \psi(t_{ij})) + [\exp(\sigma_\gamma^2) - 1]\mu_{ij}^2(x_{ij}\beta\;, \sigma_\gamma^2; \psi(t_{ij})), \quad (7)$$

and

$$\text{Corr}\;(Y_{ij}, Y_{ik}) = \frac{\rho^{j-k}(\frac{1}{\mu_{ik}(\cdot)}) + [\exp(\sigma_\gamma^2) - 1]}{\left[\left\{[\exp(\sigma_\gamma^2) - 1] + \frac{1}{\mu_{ij}(\cdot)}\right\}\left\{[\exp(\sigma_\gamma^2) - 1] + \frac{1}{\mu_{ik}(\cdot)}\right\}\right]^{\frac{1}{2}}}, \quad (8)$$

respectively.

Naive GQL (NGQL) Estimation for Regression and Over-dispersion Parameters

When non-parametric function $\psi(t_{ij})$ is ignored from the model (5), one goes back to the model (1)-(2) and may estimate the regression parameter β and over-dispersion parameter σ_γ^2, by solving the GQL estimating equations developed based on the unconditional moments given by (3) and (4). Let $y_i = (y_{i1},...,y_{ij},...,y_{ini})^0$ be the $n_i \times 1$ vector of repeated counts, and $m_i(\beta,\sigma_\gamma^2) = (m_{i1},...,m_{ij},...,m_{ini})^0$ and $V_{i,N}(\beta,\sigma_\gamma^2,\rho) = (v_{i,jk}) : n_i \times n_i$ be the mean vector and covariance matrix of the response vector y_i under the naive model (1)-(2). Here m_{ij} and $v_{i,jj}$ are given in (3), and for $j < k$, the naive covariance has the formula

$$v_{i,jk} = \rho^{k-j}m_{ij} + [\exp(\sigma_\gamma^2) - 1]m_{ij}m_{ik}.$$

Asymptotic Bias in β Estimation

By using the above notations, the naive GQL (NGQL) estimating equation for β has the form

$$\sum_{i=1}^{K}\frac{\partial m_i'(\beta\;, \sigma_\gamma^2)}{\partial\beta}V_{i,N}^{-1}(\beta\;, \sigma_\gamma^2, \rho)(y_i - m_i(\beta\;, \sigma_\gamma^2)) = 0, \quad (9)$$

which is, however, not an unbiased equation. This is because, under the true semi-parametric model (5)-(6),

$$E[Y_i - m_i(\beta,\sigma_\gamma^2))] = \mu_i(\beta,\sigma_\gamma^2,\psi(\cdot)) - m_i(\beta,\sigma_\gamma^2) \quad (10)$$
$$= [m_{i1}\{\exp(\psi(t_{i1}) - 1\},...,m_{ij}\{\exp(\psi(t_{ij}) - 1\},...,m_{ini}\{\exp(\psi(t_{ini}) - 1\}]^0\; 6= 01_{ni},$$

unless $\psi(t_{ij}) = 0$ for all $j = 1,...,n_i$. Consequently, the NGQL estimating equation (9) is bound to produce biased estimate for β.

To compute the bias, we first write the iterative equation to solve (9). This equation has the form

$$\hat{\beta}_{NGQL}(r+1) = \hat{\beta}_{NGQL}(r) + \left[\left\{\sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho) \frac{\partial m_i(\beta, \sigma_\gamma^2)}{\partial \beta'}\right\}^{-1} \times \sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho)(y_i - m_i(\beta, \sigma_\gamma^2))\right]_{\beta = \hat{\beta}_{NGQL}(r)} \quad (11)$$

yielding the asymptotic bias as

$$\lim_{K\to\infty} [\hat{\beta}_{NGQL} - \beta] = \lim_{K\to\infty} \left[\left\{\sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho) \frac{\partial m_i(\beta, \sigma_\gamma^2)}{\partial \beta'}\right\}^{-1} \times \sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho)(y_i - m_i(\beta, \sigma_\gamma^2))\right] \longrightarrow$$

$$E\left[\left\{\sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho) \frac{\partial m_i(\beta, \sigma_\gamma^2)}{\partial \beta'}\right\}^{-1} \times \sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho)(Y_i - m_i(\beta, \sigma_\gamma^2))\right]$$

$$= \left[\left\{\sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho) \frac{\partial m_i(\beta, \sigma_\gamma^2)}{\partial \beta'}\right\}^{-1} \times \sum_{i=1}^{K} \frac{\partial m_i'(\beta, \sigma_\gamma^2)}{\partial \beta} V_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho)(\mu_i(\beta, \sigma_\gamma^2, \psi(\cdot)) - m_i(\beta, \sigma_\gamma^2))\right]. \quad (12)$$

Asymptotic Bias in σ_γ^2 Estimation

It is standard to estimate this over-dispersion parameter by exploiting the second order responses. Let

$$u_i = [y_{i1}^2, \ldots, y_{ij}^2, \ldots, y_{in_i}^2, y_{i1}y_{i2}, \ldots, y_{ij}y_{ik}, \ldots, y_{i,n_i-1}y_{in_i}]' \quad (13)$$

be the $n_i(n_i+1)/2 \times 1$ vector of second order responses, and
$\xi_i = [\xi_{i,11},\ldots,\xi_{i,jj},\ldots,\xi_{i,n_in_i},\xi_{i,12},\ldots,\xi_{i,jk},\ldots,\xi_{i,(n_i-1)n_i}]' = E[U_i]$,
be its expectation under the naive model (1)-(4). Note that it follows from (3)-(4) that

$$\begin{aligned} \xi_{i,jj} &= E[Y_{ij}^2] = m_{ij} + [\exp(\sigma_\gamma^2)]m_{ij}^2 \\ \xi_{i,jk} &= E[Y_{ij}Y_{ik}] = \rho^{k-j}m_{ij} + [\exp(\sigma_\gamma^2)]m_{ij}m_{ik}, \text{ for } j < k. \end{aligned} \quad (14)$$

Also, let

$$\text{cov}[U_i] = \Omega_{i,N}(\beta, \sigma_\gamma^2, \rho)$$

under the naive model (1)-(4). One may then construct the NGQL estimating equation for σ_γ^2 as

$$\sum_{i=1}^{K} \frac{\partial \xi_i'(\beta, \sigma_\gamma^2, \rho)}{\partial \sigma_\gamma^2} \Omega_{i,N}^{-1}(\beta, \sigma_\gamma^2, \rho)[u_i - \xi_i(\beta, \sigma_\gamma^2, \rho)] = 0, \quad (15)$$

which is solved iteratively by using the iterative formula

$$\hat{\sigma}^2_{\gamma,NGQL}(r+1) = \hat{\sigma}^2_{\gamma,NGQL}(r) + \left[\left\{\sum_{i=1}^{K}\frac{\partial\xi_i'(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\Omega^{-1}_{i,N}(\beta,\sigma^2_\gamma,\rho)\frac{\partial\xi_i(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\right\}^{-1} \times \sum_{i=1}^{K}\frac{\partial\xi_i'(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\Omega^{-1}_{i,N}(\beta,\sigma^2_\gamma,\rho)[u_i-\xi_i(\beta,\sigma^2_\gamma,\rho)]\right]_{\sigma^2_\gamma=\hat{\sigma}^2_{\gamma,NGQL}(r)} \quad (16)$$

By similar calculations as in Section 3.1, one may compute the asymptotic bias for the naive estimator of σ^2_γ as

$$\lim_{K\to\infty}[\hat{\sigma}^2_{\gamma,NGQL}-\sigma^2_\gamma] = \left[\left\{\sum_{i=1}^{K}\frac{\partial\xi_i'(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\Omega^{-1}_{i,N}(\beta,\sigma^2_\gamma,\rho)\frac{\partial\xi_i(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\right\}^{-1} \times \sum_{i=1}^{K}\frac{\partial\xi_i'(\beta,\sigma^2_\gamma,\rho)}{\partial\sigma^2_\gamma}\Omega^{-1}_{i,N}(\beta,\sigma^2_\gamma,\rho)[E[U_i|\text{model (5)-(8)}]-\xi_i(\beta,\sigma^2_\gamma,\rho)]\right], \quad (17)$$

where

$$E[U_i|\text{semiparametric model (5)-(8)}] = \lambda_i = [\lambda_{i,11},\ldots,\lambda_{i,jj},\ldots,\lambda_{i,n_in_i},\lambda_{i,12},\ldots,\lambda_{i,jk},\ldots,\lambda_{i,(n_i-1)n_i}]', \quad (18)$$

with

$$\lambda_{i,jj} = E[Y_{ij}^2] = \mu_{ij}(\beta,\sigma^2_\gamma,\psi(\cdot)) + [\exp(\sigma^2_\gamma)]\mu^2_{ij}(\beta,\sigma^2_\gamma,\psi(\cdot)) \text{ by (7)}$$

$$\lambda_{i,jk} = E[Y_{ij}Y_{ik}] = \rho^{k-j}\mu_{ij}(\beta,\sigma^2_\gamma,\psi(\cdot)) + [\exp(\sigma^2_\gamma)]\mu_{ij}(\beta,\sigma^2_\gamma,\psi(\cdot))\mu_{ik}(\beta,\sigma^2_\gamma,\psi(\cdot)), \text{ for } j<k, \text{ by (8)}. \quad (19)$$

Consistent Estimation for the Semiparametric Model

The NGQL estimators of β and σ^2_γ obtained by (11) and (16) are biased and hence inconsistent. The bias was caused due to ignoring the non-parametric function $\psi(\cdot)$ from the model. Thus to obtain consistent estimators for these parameters, one has to obtain first a consistent estimator for $\psi(\cdot)$ for known β and σ^2_γ, and then estimate these later parameters.

Consistent Estimation of Non-parametric Function ψ(.)

For notational convenience, in what follows, we use μ_{ij} for $\mu_{ij}(x_{ij}\beta,\sigma^2_\gamma;\psi(\cdot))$.

At time point t_{ij} we now estimate the non-parametric function $\psi(t_{ij})$ by using a semi-parametric quasi-likelihood (SQL) approach. More specifically, similar to Severini and Staniswalis (1994), we write the SQL estimating equation for $\psi(t_{ij})|_{t_{ij}=t_0}$ as

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\,\frac{\partial\mu_{ij}}{\partial\psi(t_0)}\left(\frac{y_{ij}-\mu_{ij}}{\sigma_{i,jj}}\right)=0, \quad (20)$$

where $\mu_{ij}\equiv\exp(x'_{ij}(t_{ij})\beta+\sigma^2_\gamma/2+\psi(t_{ij}))$ and $\sigma_{i,jj}\equiv\mu_{ij}+[\exp(\sigma^2_\gamma)-1]\mu^2_{ij}$ by (7), and

$$w_{ij}(t_0) = \frac{p_{ij}(\frac{t_0 - t_{ij}}{b})}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij}(\frac{t_0 - t_{ij}}{b})} \quad (21)$$

is a kernel weight with p_{ij} as the kernel density. Note that when $w_{ij}(t_0) = 1$ for all i and j, the SQL estimating equation reduces to the well known QL estimating equation (Wedderburn (1974)).

Next because

$$\frac{\partial \mu_{ij}(\beta, x_{ij}, \sigma_\gamma^2; \psi(t_0))}{\partial \psi(t_0)} = \frac{\partial\left[\exp(x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \psi(t_0))\right]}{\partial \psi(t_0)} = \exp(x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \psi(t_0)) = \mu_{ij},$$

the SQL estimating equation (20) reduces to

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\mu_{ij}[\sigma_{i,jj}]^{-1}[y_{ij} - \mu_{ij}] = 0. \quad (22)$$

Note that because $\sigma_{i,jj}$ is a function of μ_{ij}, by assuming $\psi(\cdot)$ in $\sigma_{i,jj}$ is known from a previous iteration or so, one may compute an improved model weight by using

$$\tilde{w}_{ij}(t_0) = \frac{\mu_{ij}(\beta, \sigma_\gamma^2, \hat{\psi}_0(t_0))}{\sigma_{i,jj}(\beta, \sigma_\gamma^2, \hat{\psi}_0(t_0))}, \quad (23)$$

where β and σ_γ^2 are still assumed to be known. When these weights are used in (22), the SQL estimating equation for $\psi(\cdot)$ takes the form

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\tilde{w}_{ij}(t_0)[y_{ij} - \exp(x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \psi(t_0))] = 0, \quad (24)$$

yielding the improved estimate for $\psi(\cdot)$ as

$$\hat{\psi}(t_0) = log\left(\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\tilde{w}_{ij}(t_0)y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\tilde{w}_{ij}(t_0)exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2]}\right). \quad (25)$$

Consistent Estimation of β

In the last section, the non-parametric function $\psi(t_{ij})$ is estimated by (25), as a function of β and σ_γ^2. To be precise and clear, we express this estimated function obtained from (25) as $\hat{\psi}(t_{ij},\beta,\sigma_\gamma^2)$, and write the mean response as a function of $\hat{\psi}(.)$ as

$$E[Y_{ij}|x_{ij},\hat{\psi}(\cdot)] = \tilde{\mu}_{ij} = exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \hat{\psi}(t_{ij},\beta,\sigma_\gamma^2)] \quad (26)$$

Suppose that $\Sigma_i^{(ns)}(\beta, \sigma_\gamma^2, \rho, \hat{\psi}(\beta, \sigma_\gamma^2))$ is the non-stationary covariance matrix of y_i computed by (7) and (8). One may then obtain the semi-parametric GQL (SGQL) estimate of β by solving the GQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial(\tilde{\mu}_i)'}{\partial\beta} [\Sigma_i^{(ns)}(\beta, \sigma_\gamma^2, \rho, \hat{\psi}(\beta, \sigma_\gamma^2))]^{-1} (y_i - \tilde{\mu}_i) = 0, \quad (27)$$

[Sutradhar (2003)] for β, where

$$\frac{\partial(\tilde{\mu}_i)'}{\partial\beta} = \frac{\partial(\tilde{\mu}_{i1}, \ldots, \tilde{\mu}_{ij}, \ldots, \tilde{\mu}_{in_i})}{\partial\beta},$$

with

$$\frac{\partial \tilde{\mu}_{ij}}{\beta} = \frac{\partial}{\beta}\left(exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \hat{\psi}(t_{ij},\beta, \sigma_\gamma^2)]\right)$$
$$= \tilde{\mu}_{ij}\left(x_{ij}(t_{ij}) + \frac{\partial}{\beta}\hat{\psi}(t_{ij},\beta, \sigma_\gamma^2)\right)$$
$$= \tilde{\mu}_{ij}\left[x_{ij}(t_{ij}) - \frac{\sum_{\ell=1}^{K}\sum_{u=1}^{n_\ell} w_{\ell u}(t_{ij})\tilde{w}_{\ell u}(t_{ij}) exp[x'_{\ell u}(t_{\ell u})\beta + \sigma_\gamma^2/2] x_{\ell u}(t_{\ell u})}{\sum_{\ell=1}^{K}\sum_{u=1}^{n_\ell} w_{\ell u}(t_{ij})\tilde{w}_{\ell u}(t_{ij}) exp[x'_{\ell u}(t_{\ell u})\beta + \sigma_\gamma^2/2]}\right]. \quad (28)$$

Because $E[Y_i - \tilde{\mu}_i] = 0$, the SGQL estimating equation (27) is unbiased and hence it will produce consistent estimator for β. For known σ_γ^2 and ρ, this SSGQL estimating equation may be solved for β using the well-known Newton-Raphson iterative procedure. To be specific, first write

$$f(\beta) = \sum_{i=1}^{K} \frac{\partial(\tilde{\mu}_i)'}{\beta}\ [\Sigma_i^{(ns)}(\cdot)]^{-1}\ (y_i - \tilde{\mu}_i).$$

Now starting with an initial value for β, each step of the following iterative equation

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} - [(f'(\beta))^{-1} f(\beta)]_{\beta=\beta_{(r)}} \quad (29)$$

updates the value of β until convergence. The derivative function, $f'(\beta)$ at $\beta = \beta_{(r)}$ in (29) is calculated as

$$f'(\beta) = -\sum_{i=1}^{K} \frac{\partial(\tilde{\mu}_i)'}{\beta}\ [\Sigma_i^{(ns)}(\cdot)]^{-1}\ \frac{\partial \tilde{\mu}_i}{\beta'}.$$

Consistent Estimation of σ_γ^2

Recall from Section 3.2 that u_i (13) has its expectation λ_i (18) under the semi-parametric model (5)-(8). Suppose that $\Omega_i(\beta, \sigma_\gamma^2, \rho, \hat{\psi}(\cdot))$ is the covariance matrix of u_i. By similar calculations as for the β estimation from the last section, one may obtain consistent estimator of σ_γ^2 by solving the SGQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial \lambda'_i}{\partial \sigma_\gamma^2} \Omega_i^{-1}(\beta, \sigma_\gamma^2, \rho, \hat{\psi}(\cdot))(u_i - \lambda_i) = 0, \quad (30)$$

with

$$\frac{\partial \lambda'_i}{\partial \sigma_\gamma^2} = [\frac{\partial \lambda_{i,11}}{\partial \sigma_\gamma^2}, \ldots, \frac{\partial \lambda_{i,jj}}{\partial \sigma_\gamma^2}, \ldots, \frac{\partial \lambda_{i,n_i n_i}}{\partial \sigma_\gamma^2}, \frac{\partial \lambda_{i,12}}{\partial \sigma_\gamma^2}, \ldots, \frac{\partial \lambda_{i,12}}{\partial \sigma_\gamma^2}, \ldots, \frac{\partial \lambda_{i,(n_i-1)n_i}}{\partial \sigma_\gamma^2}],$$

where, by (19), one writes

$$\frac{\partial \lambda_{i,jj}}{\partial \sigma_\gamma^2} = \frac{\partial \tilde{\mu}_{ij}(\cdot)}{\partial \sigma_\gamma^2}\left[1 + 2\exp(\sigma_\gamma^2)\tilde{\mu}_{ij}(\cdot)\right] + [\exp(\sigma_\gamma^2)]\{\tilde{\mu}_{ij}(\cdot)\}^2, \text{for } j = 1, \ldots, n_i$$
$$\frac{\partial \lambda_{i,jk}}{\partial \sigma_\gamma^2} = \rho^{k-j}\frac{\partial \tilde{\mu}_{ij}(\cdot)}{\partial \sigma_\gamma^2} + [\exp(\sigma_\gamma^2)]\tilde{\mu}_{ij}(\cdot)\tilde{\mu}_{ik}(\cdot) + \exp(\sigma_\gamma^2)\left[\tilde{\mu}_{ij}(\cdot)\frac{\partial \tilde{\mu}_{ik}(\cdot)}{\partial \sigma_\gamma^2} + \tilde{\mu}_{ik}(\cdot)\frac{\partial \tilde{\mu}_{ij}(\cdot)}{\partial \sigma_\gamma^2}\right], \text{ for } j < k, \quad (31)$$

with

$$\begin{aligned}\frac{\partial \tilde{\mu}_{ij}}{\partial} &= \frac{\partial}{\partial \sigma_\gamma^2}\left(exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \hat{\psi}(t_{ij}\beta, \sigma_\gamma^2)]\right)\\ &= \tilde{\mu}_{ij}\left(\frac{1}{2} + \frac{\partial}{\partial \sigma_\gamma^2}\hat{\psi}(t_{ij}\beta, \sigma_\gamma^2)\right)\\ &= \frac{\tilde{\mu}_{ij}}{2}\left[1 - \frac{\sum_{\ell=1}^{K}\sum_{u=1}^{n_\ell} w_{\ell u}(t_{ij})\tilde{w}_{\ell u}(t_{ij})exp[x'_{\ell u}(t_{\ell u})\beta + \sigma_\gamma^2/2]}{\sum_{\ell=1}^{K}\sum_{u=1}^{n_\ell} w_{\ell u}(t_{ij})\tilde{w}_{\ell u}(t_{ij})exp[x'_{\ell u}(t_{\ell u})\beta + \sigma_\gamma^2/2]}\right]\\ &= 0, \end{aligned} \tag{32}$$

yielding

$$\frac{\partial \lambda_{i,jj}}{\partial \sigma_\gamma^2} = [\exp(\sigma_\gamma^2)]\{\tilde{\mu}_{ij}(\cdot)\}^2, \text{for } j = 1, \ldots, n_i$$

$$\frac{\partial \lambda_{i,jk}}{\partial \sigma_\gamma^2} = [\exp(\sigma_\gamma^2)]\tilde{\mu}_{ij}(\cdot)\tilde{\mu}_{ik}(\cdot), \text{ for } j < k, \tag{33}$$

Estimation of ρ

Note that the SGQL estimating equation (27) for β and (30) for σ_γ^2, require ρ to be known. We obtain a consistent estimator for this parameter by using a method of moments. To be specific, following Sutradhar (2010), we use the sample lag 1 auto-correlation and equate that to its population counterpart to construct the desired moment estimating equation for ρ. That is , for the moment estimate for ρ, we solve

$$A_1 = \frac{\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^* y_{i,j-1}^*}{\sum_{i=1}^{K}(n_i-1)}}{\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}^*)^2}{\sum_{i=1}^{K} n_i}} \approx \frac{E\left[\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^* y_{i,j-1}^*}{\sum_{i=1}^{K}(n_i-1)}\right]}{E\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}^*)^2}{\sum_{i=1}^{K} n_i}\right]}, \tag{34}$$

where

$$y_{ij}^* = \frac{y_{ij} - \mu_{ij}}{\sqrt{\sigma_{i,jj}}},$$

with

$$\begin{aligned}\mu_{ij}(x_{ij}\beta, \sigma_\gamma^2; \psi(t_{ij})) &= \exp[x'_{ij}(t_{ij})\beta + \sigma_\gamma^2/2 + \psi(t_{ij})]\\ \sigma_{i,jj}(x_{ij}\beta, \sigma_\gamma^2; \psi(t_{ij})) &= \mu_{ij}(x_{ij}\beta, \sigma_\gamma^2; \psi(t_{ij})) + [\exp(\sigma_\gamma^2) - 1]\mu_{ij}^2(x_{ij}\beta, \sigma_\gamma^2; \psi(t_{ij}))\end{aligned}$$

Next because, by (7)-(8),

$$\begin{aligned}E[Y_{ij}^{*2}] &= 1\\ E[Y_{ij}^* Y_{i,j-1}^*] &= \frac{1}{\{\sigma_{i,jj}\sigma_{i,j-1,j-1}^*\}^{\frac{1}{2}}}\left[\rho\mu_{i,j-1} + \{\exp(\sigma_\gamma^2) - 1\}\mu_{i,j-1}\mu_{ij}\right],\end{aligned}$$

it then follows from (34) that

$$A_1 = \rho G_1 + B_1, \tag{35}$$

yielding

$$\rho = \frac{A_1 - B_1}{G_1},$$

where

$$B_1 = \frac{\{\exp(\sigma_\gamma^2) - 1\}}{\sum_{i=1}^K (n_i - 1)} \sum_{i=1}^K \sum_{j=2}^{n_i} \frac{\mu_{i,j-1}\mu_{ij}}{\{\sigma_{i,jj}\sigma^*_{i,j-1,j-1}\}^{\frac{1}{2}}}$$

$$G_1 = \frac{1}{\sum_{i=1}^K (n_i - 1)} \sum_{i=1}^K \sum_{j=2}^{n_i} \frac{\mu_{i,j-1}}{\{\sigma_{i,jj}\sigma^*_{i,j-1,j-1}\}^{\frac{1}{2}}}.$$

Concluding Remarks

Panel count data model involves the regression parameter *β,* overdispersion parameterσ_γ^2 and a correlation index parameter *ρ,* whereas a semi-parametric model involves these parameter and an additional non-parametric function $\psi(\cdot)$ in time. It is shown in the paper that when the main parameters β andσ_γ^2 are estimated ignoring $\psi(\cdot)$, their estimators become biased and hence inconsistent. As a remedy, the non-parametric function is first estimated consistently for known β andσ_γ^2, and then the later parameters are estimated consistently.

The correlation index parameter is also estimated consistently.

References

Blundell, R., Griffith, R. & Windmeijer, F. (1995). Individual effects and dynamics in count data. *Discussion paper* 95-03. Department of Economics, University College London.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association* 88, 925.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, 102, 632-641.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association*, 103, 1520-1533.

Hausman, J.A., Hall, B.H. & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R and D relationship. *Econometrica* 52, 908-938.

Hua, L. (2010). *Spline-based sieve semiparametric generalized estimating equation for panel count data.* Unpublished PhD thesis. Department of Biostatistics, University of Iowa.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association* 93, 720-729.

Lin, X. and Carroll, R. J. (2001). Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. *Journal of the American Statistical Association*, 96, 1045-1056.

Montalvo, J. G. (1997). GMM estimation of count-panel-data models with fixed effects and predetermined instruments. *Journal of Business and Economic Statistics* 15, 82-89.

Severini, T. A., and Staniswalis, J. G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, 89, 501-511.

Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, 18, 377-393.

Sutradhar, B. C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Sankhya B* 66, 261-289.

Sutradhar, B. C. (2010). Inferences in generalized linear longitudinal mixed models. *Canadian Journal of Statistics, Special issue*, 38, 174-196.

Sutradhar, B. C., Jowaheer, V. & Rao, R. Prabhakar (2014). Remarks on asymptotic efficient estimation for regression effects in stationary and non-stationary models for panel count data. *Brazilian Journal of Probability and Statistics.* 28, 241-254.

Sutradhar, B. C. & Warriyar, K. V. Vineetha (2014). Semi-parametric models for longitudinal count data. Submitted.

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

Wooldridge, J. (1999). Distribution-free estimation of some non-linear panel data models. *Journal of Econometrics* 90, 77-97.