Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 1, 28–46
DOI: 10.1214/14-BJPS254

Inferences in median regression models for asymmetric longitudinal data: A quasi-likelihood approach

Varathan Nagarajah[a], Brajendra C. Sutradhar[a], Vandna Jowaheer[b] and Atanu Biswas[c]

[a]*Memorial University*
[b]*University of Mauritius*
[c]*Indian Statistical Institute*

Abstract. In the independence setup, when the responses exhibit high degree of asymmetry, the median regression model is preferred to the mean regression model to obtain consistent and efficient regression estimates. However, when this type of asymmetric data are collected repeatedly over time, developing median regression model for such correlated asymmetric data may not be easy. As a remedy, there exist some studies where the longitudinal correlations of this type of asymmetric data have been computed using the moment estimates for all pairwise correlations and these correlations of repeated (multi-dimensional) data used to develop a median based quasi-likelihood approach for estimation of the regression effects. By considering an autoregressive order 1 (AR(1)) model for longitudinal exponential responses, in this paper, it is however, demonstrated that the existing pairwise estimates of correlations under median regression model may yield inefficient estimates as compared to the simpler independence assumption based estimates. We illustrate the inference techniques discussed in the paper by reanalyzing the well-known labor pain data.

1 Introduction

Suppose that a scalar response y_{it} and a p-dimensional vector of covariates x_{it} are observed for cluster $i = 1, 2, \dots, K$ at a time point t $(t = 1, 2, \dots, T)$. For the ith cluster, let $y_i = (y_{i1}, \dots, y_{it}, \dots, y_{iT})'$ be the response vector and $X_i = (x_{i1}, \dots, x_{it}, \dots, x_{iT})'$ bet the $T \times p$ matrix of covariates. Let β denote the $p \times 1$ vector of regression parameters which measures the effects of x_{it} on y_{it} for all $t = 1, 2, \dots, T$ and for all $i = 1, 2, \dots, K$. When the responses are continuous and their distributions are symmetric, one fits the linear model

$$y_i = X_i\beta + \varepsilon_i \tag{1.1}$$

to estimate β. Suppose that $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$ in (1.1) has the mean vector 0 and covariance matrix $\Sigma_i = A_i^{1/2} C(\rho) A_i^{1/2}$ with $A_i = \mathrm{diag}[\mathrm{Var}(\varepsilon_{i1}), \dots, \mathrm{Var}(\varepsilon_{it}), \dots,$

Key words and phrases. Asymmetric data, bias in regression estimates, efficiency, generalized quasi-likelihood, longitudinal correlations for asymmetric data, mean and median regression, repeated responses over time.

Received February 2014; accepted July 2014.

$\mathrm{Var}(\varepsilon_{iT})]$ and $C(\rho)$ as the $T \times T$ correlation matrix. It then follows that the well-known generalized least square (GLS) estimator of β is a solution of the estimating equation

$$\sum_{i=1}^{K} X_i' \Sigma_i^{-1}(y_i - X_i\beta) = 0. \tag{1.2}$$

Note that if the responses are continuous but their distributions are asymmetric such as Gamma (McCullagh and Nelder (1989, p. 300)), β estimates of the mean regression model (1.1) can be inefficient (Bassett and Koenker (1978)). As a remedy, for the asymmetric data in the independence setup, that is when T repeated responses $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ are treated to be independent, one may follow Morgenthaler (1992), among others and model the median rather than the mean of the responses as a function of the covariates x. More specifically, let $m_{it} = (m_{i1}, \ldots, m_{iT})'$ be the median of y_{it} and $m_i = (m_{i1}, \ldots, m_{iT})'$. Further, for some link function $g(\cdot)$, let

$$g(m_{it}) = x_{it}'\beta. \tag{1.3}$$

For independent data, Morgenthaler (1992, eqn. (3.1)) suggested to solve an absolute deviations based quasi-likelihood (ADQL) estimating equation

$$\sum_{i=1}^{K} D_i' \{\mathrm{diag}[s_{i1}, \ldots, s_{it}, \ldots, s_{iT}]\}^{-1} \{\mathrm{sgn}(y_i - m_i)\} = 0, \tag{1.4}$$

where s_{it} is an user-supplied function that models the scatter of the responses as a function of the median m_{it}, D_i is the $T \times p$ partial derivatives of the mean vector with respect to β, that is $D_i' = \partial\mu_i'/\partial\beta$, where $\mu_i = [\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT}]'$ is the mean response vector of y_i. This ADQL approach appears to have several limitations. First, s_{it} is chosen as $s_{it} \propto m_{it}^2$ which may be an appropriate choice only if m_{it} holds a proportionality relation to μ_{it} so that $m_{it} = c\mu_{it}$ for a suitable constant c for all $t = 1, 2, \ldots, T$. Second, using a mean response based gradient matrix D_i is also dependent on such proportionality relation between means and medians, which may not hold for all t.

We now turn back to the longitudinal case where it is expected that the repeated responses $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ will be correlated. To accommodate this type of dependent observations, Jung (1996), for example, suggests to solve an indicator function based quasi-likelihood estimating equation, where indicator variable is defined as $I(y_{it} \geq m_{it})$, with m_{it} as the median of y_{it} as in (1.3). More specifically, in the cluster regression setup, Jung's quasi-likelihood (QL) estimating equation can be expressed as

$$\sum_{i=1}^{K} \frac{1}{\phi_i} B_i' \Gamma_i \Omega_i^{-1} \left\{ I(y_i \geq m_i) - \frac{1}{2} 1_T \right\} = 0, \tag{1.5}$$

where $I(y_i \geq m_i) = [I(y_{i1} \geq m_{i1}), \dots, I(y_{it} \geq m_{it}), \dots, I(y_{iT} \geq m_{iT})]'$ is the $T \times 1$ vector of indicator functions, 1_T is the $T \times 1$ unit vector, Ω_i is the $T \times T$ covariance matrix of $[I(y_i \geq m_i) - \frac{1}{2}1_T]$, B_i is the $T \times p$ first derivative matrix of m_i with respect to β, that is, $B_i = \partial m_i'/\partial\beta$, where $m_i = (m_{i1}, \dots, m_{iT})'$ with $g(m_{it}) = x_{it}'\beta$, and $\phi_i^{-1}\Gamma_i = \phi_i^{-1}\,\mathrm{diag}[\gamma(m_{i1}), \dots, \gamma(m_{it}), \dots, \gamma(m_{iT})]$, where $\phi_i^{-1}\gamma(m_{it})$ is the probability density function (pdf) of y_{it} evaluated at the median m_{it}. Jung (1996) refers to the solution of (1.5) for β as the maximum quasi-likelihood estimate. Note that the QL estimating equation (1.5) may be treated as a generalization of the ADQL estimating equation (1.4), from the independent setup to the longitudinal setup. However, one cannot compute Ω_i, the covariance matrix of the vector of indicator functions, as we cannot compute the pairwise bivariate distributions of the elements of the asymmetric response vector $y_i = (y_{i1}, \dots, y_{it}, \dots, y_{iT})'$. This is because the correlation structure or the joint distribution of the repeated responses may not be available. To resolve this computational issue, Jung (1996) has estimated the pairwise elements of Ω_i matrix by estimating the bivariate probability of any two indicator variables using a distribution free moment approach. There are, however, several limitations to this pairwise probability estimation by using such a moment approach. First, if the repeated responses follow an auto-correlation model, which is most likely in practice, using pairwise probabilities based on the concept of unstructured correlations for repeated data may yield inefficient estimates, as in this approach one is computing too many correlations whereas auto-correlation model contains only a few lag correlations. Next, there is no guaranty that this type of unstructured correlations based estimation can be more efficient than using simpler independence assumption based Ω_i for the estimation of β (Sutradhar (2011)) involved in the median regression function.

In this paper, we examine the aforementioned efficiency issue under an exponential auto-regressive of order 1 (EAR(1)) model, where y_{it} marginally follows an exponential distribution with a specified median regression function in β, and the repeated responses $y_{i1}, \dots, y_{it}, \dots, y_{iT}$, follow a Gausian type AR(1) correlation structure. This EAR(1) model with its basic properties is discussed in Section 2. The true model based GQL estimation (Sutradhar (2003)) and various semi-parametric GQL estimation approaches including the independence based approach, are discussed in the same section. In Section 3, the median regression estimation based on three correlation structures such as (a) Jung's (1996) unstructured correlations, (b) an auto-correlation structure (Sutradhar (2003)), and (c) 'working' independence structure, are compared with the true EAR(1) based estimation through a simulation study. The simulation results in this paper show that the simpler independence assumption based estimates, surprisingly are more efficient (in the sense mean squared error) than other two correlation structures based estimates. In the simulation study, we have also included the mean regression based QL estimation approach to examine mainly its relative performance for high degree of asymmetry in exponential data caused by certain outliers. In Section 4, we illustrate the application of the aforementioned QL estimation approaches for the analysis of a well-known labor pain data which was earlier analysed by Davis (1991), Jung (1996) and Geraci and Bottai (2007), among others.

2 Exponential AR(1) Model

Some authors such as Geraci and Bottai (2007) have modelled the asymmetric data at a given time point by a Laplace distribution, and modelled the correlations through the common individual random effects showed by the repeated responses. However, even though the random effects generate an equicorrelation structure for the repeated responses, they do not appear to address the time effects (Sutradhar (2011, Section 2.4)). This is because these individuals specific random effects may remain the same throughout the data collection period and hence cannot represent any time effects. In the longitudinal setup, some authors considered linear fixed or mixed models with errors unspecified for the purpose of quantile regression estimation. See, for example, Koenker (2004), Karlsson (2007), and Fu and Wang (2012). In their approaches, the longitudinal correlations are accommodated either through random effects or by using arbitrary 'working' correlations. Galvao Jr. (2011) considered a dynamic panel data model, that accommodates the longitudinal correlations but the empirical studies were confined to symmetric errors.

In this paper, as opposed to the aforementioned studies, we consider asymmetric longitudinal responses and study the median based regression effects. To be specific, following Hasan et al. (2007, Section 2.1, p. 552) we consider a class of non-stationary auto-correlations models for longitudinal exponential failure time data, AR(1) model is being an important special case. We consider this EAR(1) model and provide estimating equation for median based regression parameters. Suppose that the response y_{i1} follows an exponential distribution with parameter $\lambda_{i1} = h(x'_{i1}\beta)$, for a suitable known link function $h(\cdot)$. That is,

$$f(y_{i1}) = \lambda_{i1} \exp(-\lambda_{i1} y_{i1}). \tag{2.1}$$

Next for $t = 2, \dots, T$, following Hasan et al. (2007, eqn. (2.1)) [see also Gaver and Lewis (1980)], we write a dynamic model in exponential variables as,

$$y_{it} = \rho_i y_{i,t-1} + I_{it} a_{it}, \qquad t = 2, \dots, T; i = 1, \dots, K, \tag{2.2}$$

where, for a given i, $\{a_{it}, t = 1, \dots, T; i = 1, \dots, K\}$ is a sequence of exponential random variables with parameter $\lambda_{it} = \exp(-x'_{it}\beta)$ and $\rho_i = \rho\frac{\lambda_{i,t-1}}{\lambda_{it}}$ with ρ as a probability parameter or correlation parameter $(0 \le \rho \le 1)$. In (2.2), I_{it} is an indicator variable such that

$$I_{it} = \begin{cases} 0 & \text{with probability } \rho, \\ 1 & \text{with probability } 1 - \rho, \end{cases}$$

and I_{it} and a_{it} are assumed to be independent.

It is of interest to estimate β which is involved in both mean and median functions of the responses. Note that for the estimation of β from the mean function, one computes it from the mean regression based estimating equation, whereas the same β may be estimated from the median function by solving the median regression based estimating equation. The main objective of the paper is to develop a median regression based GQL estimating equation approach for β by accommodating the correlations of the longitudinal exponential data. This development is given in the next section. In the same section, we also provide various versions of this GQL approach by using suitable 'working' correlation structures. This we do in order to examine the correlation structure misspecification effect.

Note that the EAR(1) process (2.2) produces a auto-correlation structure given by

$$\mathrm{Corr}[Y_{it}, Y_{i,t+j}] = \rho^j$$

[see also Hassan et al. (2007, eqn. (2.2))] which, similar to Gaussian models, is known to be a stationary correlation structure. Thus, only stationary condition for the series $\{y_{it}, t = 1, \ldots, T\}$ is $0 < \rho < 1$, which does not have anything to do with the values of β in the mean function $\lambda_{it} = \exp(-x'_{it}\beta)$ of the exponential process. However, these means are non-stationary as they depend on time dependent covariates. For a smooth data with finite mean, one of course has to put condition on β which is however a different problem. For more details on the stationarity, for example, in the integer valued process one may refer to Jacobs and Lewis (1983), and for similar stationarity in exponential and gamma process one may refer to Sim (1986, 1990). Further note that for a stationarity of a linear process, it is often convenient to examine a condition whether the associated auto-covariance function has 'unit' roots (Box and Jenkins (1976, Sections 3.1.3, 3.1.4)). This alternative approach for stationarity is, however, not discussed adequately in the context of exponential and/or gamma dynamic models.

2.1 Median regression based GQL estimation

To develop the median based estimating equation, first we compute the median m_{it}, and the indicator variable $\delta(y_{it} \geq m_{it})$ of the responses for the EAR(1) model (2.1)–(2.2). As far as the computation for the median m_{it} under the dynamic model (2.2) is concerned, one requires to compute the marginal distribution of y_{it} for all $t = 1, \ldots, T$. For $t = 1$, the marginal distribution of y_{i1} is known by (2.1). Thus one may compute m_{i1}. However, for remaining $t = 2, \ldots, T$, one has to find the marginal distribution of y_{it} $(t = 2, \ldots, T)$ when it is known that $y_{i1}, \ldots, y_{iT}$ jointly follow the EAR(1) correlation model (2.2). This marginal distributional property for y_{it} under (2.2) has also been studied by some authors. For example, it has been shown in an unpublished Ph.D. thesis (Hasan (2004, Lemma 2.2)) [see also Gaver and Lewis (1980, Section 2) for similar properties with stationary parameters such as $\lambda_{it} = \lambda_i$] that y_{it} in general follows the exponential distribution with parameter λ_{it}.

Thus, it follows that, for all $t = 1, \dots, T$,

$$\int_0^{m_{it}} f(y_{it})\,dy_{it} = \int_0^{m_{it}} \frac{1}{\lambda_{it}} \exp(-\lambda_{it} y_{it})\,dy_{it} = \frac{1}{2} \quad \Longrightarrow \quad m_{it} = \frac{\log 2}{\lambda_{it}}. \tag{2.3}$$

Note that, in order to develop a median regression based estimating equation for β involved in the median $m_{it} = \frac{\log 2}{\lambda_{it}}$ with $\lambda_{it} = \exp(-x'_{it}\beta)$, one first defines an indicator variable relating y_{it} and m_{it}, as

$$\delta(y_{it} \ge m_{it}) = \begin{cases} 1 & \text{if } y_{it} \ge m_{it}, \\ 0 & \text{if } y_{it} < m_{it}. \end{cases} \tag{2.4}$$

It is clear that the expectation and variance of the indicator variable have formulas

$$\tilde{\mu}_{it} = E\big[\delta(y_{it} \ge m_{it})\big] = \frac{1}{2}, \qquad \tilde{\sigma}_{itt} = \mathrm{Var}\big[\delta(y_{it} \ge m_{it})\big] = \frac{1}{4}. \tag{2.5}$$

Next, the covariance between two indicator variables $\delta(y_{iv} \ge m_{iv})$ and $\delta(y_{it} \ge m_{it})$ is given by

$$\begin{aligned} \tilde{\sigma}_{ivt} &= \mathrm{Cov}\big[\delta(y_{iv} \ge m_{iv}), \delta(y_{it} \ge m_{it})\big] \\ &= E\big[\delta(y_{iv} \ge m_{iv})\delta(y_{it} \ge m_{it})\big] - \frac{1}{4} \\ &= \Pr(y_{iv} \ge m_{iv}, y_{it} \ge m_{it}) - \frac{1}{4}, \end{aligned} \tag{2.6}$$

where, for the present EAR(1) model, the bivariate probabilities $\Pr(y_{iv} \ge m_{iv}, y_{it} \ge m_{it})$ may be computed as

$$\begin{aligned} &\Pr(y_{iv} \ge m_{iv}, y_{it} \ge m_{it}) \\ &\quad = \begin{cases} e^{-\lambda_{iv} m_{iv}}; & \text{for } m_{it} \le \rho^{t-v} m_{iv}, \\ e^{-\lambda_{it} m_{it}} e^{-\lambda_{iv}(1-\rho^{t-v}) m_{iv}}; & \text{for } m_{it} > \rho^{t-v} m_{iv} \end{cases} \end{aligned} \tag{2.7}$$

[see Hasan (2004, Section 4.1.1, p. 59)] where $\rho^{t-v} = \mathrm{Corr}(y_{iv}, y_{it})$; for $v < t$.

Now by writing

$$\delta(y_i \ge m_i) = \big[\delta(y_{i1} \ge m_{i1}), \dots, \delta(y_{it} \ge m_{it}), \dots, \delta(y_{iT} \ge m_{iT})\big]' : T \times 1, \tag{2.8}$$

one obtains the mean and covariance of this T-dimensional variable $\delta(y_i \ge m_i)$ as

$$E\big[\delta(y_i \ge m_i)\big] = \tilde{\mu}_{i,\delta} = \frac{1}{2} 1_T \quad \text{and} \quad \tilde{\Sigma}_{i,\delta}(\beta, \rho) = (\tilde{\sigma}_{ivt}), \tag{2.9}$$

where $\tilde{\sigma}_{ivt}$ is given by (2.6).

One may then write the median regression based GQL estimating equation for β as

$$\sum_{i=1}^{K} \frac{\partial \delta(y_i \ge m_i)}{\partial \beta'} \tilde{\Sigma}_{i,\delta}^{-1} \left[\delta(y_i \ge m_i) - \frac{1}{2}\mathbf{1}_T\right] = 0 \tag{2.10}$$

[Jung (1996), Sutradhar (2003)] where

$$\frac{\partial \delta(y_i \geq m_i)}{\partial \beta'} = \frac{\partial}{\partial m_i}[2\tilde{F}_i(y_i - m_i) - 1]\frac{\partial m_i}{\partial \beta'} = -2\tilde{f}_i(m_i)\tilde{D}_i' \tag{2.11}$$

with

$$\tilde{D}_i' = \left[\frac{\partial m_{i1}}{\partial \beta}, \ldots, \frac{\partial m_{it}}{\partial \beta}, \ldots, \frac{\partial m_{iT}}{\partial \beta}\right] \quad \text{with } \frac{\partial m_{it}}{\partial \beta} = \frac{\partial(\log 2/\lambda_{it})}{\partial \beta},$$

$$\tilde{f}_i(m_i) = \mathrm{diag}[\tilde{f}_{i1}(m_{i1}), \ldots, \tilde{f}_{it}(m_{it}), \ldots, \tilde{f}_{iT}(m_{iT})] : T \times T, \tag{2.12}$$

$$\text{with } \tilde{f}_{it}(m_{it}) = f(y_{it})|_{y_{it}=m_{it}}.$$

Note that for known ρ, the estimating equation (2.10) may be solved iteratively using

$$\begin{aligned}\hat{\beta}(r+1) = \hat{\beta}(r) + &\left[\sum_{i=1}^{K} \frac{\partial \delta(y_i \geq m_i)}{\partial \beta'} \hat{\tilde{\Sigma}}_{i,\delta}^{-1} \frac{\partial \delta(y_i \geq m_i)'}{\partial \beta}\right]_r^{-1} \\ &\times \left[\sum_{i=1}^{K} \frac{\partial \delta(y_i \geq m_i)}{\partial \beta'} \hat{\tilde{\Sigma}}_{i,\delta}^{-1} \left\{\delta(y_i \geq m_i) - \frac{1}{2}\mathbf{1}_T\right\}\right]_r,\end{aligned} \tag{2.13}$$

where $[\,]_r$ is computed by evaluating the quantity in $[\,]$ using $\beta = \hat{\beta}(r)$.

2.1.1 *Semi-parametric GQL estimation.* Note that the construction of the median regression based GQL estimating equation (2.10) requires the knowledge of (a) marginal density of y_{it} to be evaluated at median m_{it}, and (b) the correlation structure for $\delta(y_i \geq m_i)$ (2.8) evaluated from pairwise bivariate probabilities (2.7) based on the correlation model such as EAR(1) structure (2.2) for the repeated responses. In this section, we relax the need for the correlation model indicated in (b). To be specific, we provide several semi-parametric versions of the median regression based GQL estimating equation (2.10), where assumption (a) is still used, but instead of (b), we consider three types of model free correlation structures as follows.

(*a*) *Using independence among repeated responses.* In this case, the correlation index parameter ρ is treated to be zero. Consequently, the pairwise bivariate probabilities for $\delta(y_{iv} \geq m_{iv})$ and $\delta(y_{it} \geq m_{it})$ given in (2.7), for example, reduce to

$$\Pr(y_{iv} \geq m_{iv}, y_{it} \geq m_{it}) = e^{-2\lambda_{it}m_{it}} = \frac{1}{4}, \tag{2.14}$$

because the marginal density $f(y_{it})$ is known, implying that the median $m_{it} = \frac{\log 2}{\lambda_{it}}$ is known. Applying (2.14) to (2.6), one obtains zero covariances or correlations, i.e., $\tilde{\sigma}_{ivt} = 0$. Thus, one may simply use $\tilde{\Sigma}_{i,\delta} = \mathrm{Cov}(\delta(y_i \geq m_i)) = A_{i,\delta}^{1/2}\tilde{C}_{i,\delta}(\rho)A_{i,\delta}^{1/2} = \frac{1}{4}\tilde{C}_{i,\delta} = \frac{1}{4}I_T$ in (2.10) for the estimation of β.

Note that this result for the indicator variables, namely $\tilde{\Sigma}_{i,\delta} = \mathrm{Cov}(\delta(y_i \geq m_i)) = \frac{1}{4} I_T$ also holds when exponential responses $y_{i1}, \ldots, y_{iT}$ follow an MA(1) (moving average of order 1) [EMA(1)] or equi-correlated (EQC) [EEQC] model (Hasan et al. (2007, eqns. (2.5) and (2.11))).

(*b*) *Jung's approach.* To apply the QL estimating equation (1.5), Jung (1996) has estimated the pairwise elements of $\tilde{\Sigma}_{i,\delta}$ matrix by estimating the bivariate probability of any two indicator variables using a distribution free moment approach. To be specific, the pairwise bivariate probabilities for $\delta(y_{iv} \geq m_{iv})$ and $\delta(y_{it} \geq m_{it})$ have been non-parametrically estimated by using the proportion as

$$\hat{\Pr}(y_{iv} \geq m_{iv}, y_{it} \geq m_{it}) = \frac{\sum_{i=1}^{K} \delta(y_{iv} \geq m_{iv})\delta(y_{it} \geq m_{it})}{K}. \tag{2.15}$$

Thus, to construct $\tilde{\Sigma}_{i,\delta}$ matrix, one writes $\tilde{\Sigma}_{i,\delta} = (\tilde{\sigma}_{ivt})$, where

$$\tilde{\sigma}_{ivt} = \frac{\sum_{i=1}^{K} \delta(y_{iv} \geq m_{iv})\delta(y_{it} \geq m_{it})}{K} - \frac{1}{4}. \tag{2.16}$$

The QL estimate of β is then obtained by solving (1.5) or equivalently using the iterative equation (2.13).

(*c*) *Lag-correlation approach.* Note that, when the repeated responses $y_{i1}, \ldots, y_{iT}$ follow EAR(1), EMA(1) or EEQC structures, their correlations become lag dependent as in the Gaussian case. Thus, $\mathrm{Corr}(y_{iv}, y_{it})$ depends on $|v - t|$ rather than individuals $v, t = 1, \ldots, T$. For example, in the EAR(1) case $\mathrm{Corr}(y_{iv}, y_{it}) = \rho^{|t-v|}$ which may be denoted by $\rho_{|t-v|}$. This shows that unlike in (2.16) one only needs to compute the correlation matrix

$$\begin{aligned} \tilde{C}_i(\rho) &= \left(\mathrm{Corr}[\delta(y_{iv} \geq m_{iv}), \delta(y_{it} \geq m_{it})]\right) \\ &= \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \cdot & \cdots & 1 \end{bmatrix}, \end{aligned} \tag{2.17}$$

which we do by estimating $\tilde{\rho}_\ell$ as

$$\hat{\tilde{\rho}}_\ell = \left[\sum_{i=1}^{K} \sum_{t=1}^{T-\ell} \delta(y_{it} \geq m_{it}, y_{i,t+\ell} \geq m_{i,t+\ell})/\big(K(T-\ell)\big) - \frac{1}{4}\right] / (1/4). \tag{2.18}$$

Consequently, $\tilde{\Sigma}_{i,\delta}$ is computed as $\tilde{\Sigma}_{i,\delta} = \frac{1}{4}\tilde{C}_i(\rho)$, and $\hat{\beta}$ is obtained by solving (1.5) or equivalently using the iterative equation (2.13).

3 A simulation study

To examine the relative performance of the median regression based approaches described in Section 2.1.1, in this section we conduct a simulation study using finite sample size $K = 100$ individuals each having $T = 4$ repeated exponential responses. For the regression parameter β involved in $E(Y_{it}) = \lambda_{it}^{-1} = \exp(x'_{it}\beta) = \mu_{it}(\beta)$ or equivalently in the median function $m_{it}(\beta) = \text{median}(Y_{it}) = (\log 2)\lambda_{it}^{-1} = (\log 2)\exp(x'_{it}\beta)$, we consider $p = 1$ for simplicity with $\beta_1 \equiv \beta = 0.5, 0.7,$ and 1.0. For the design covariate we choose, for example, a stationary covariate with $x_{it} = \tilde{x}_i \sim U(0, 1)$ for all $t = 1, \dots, 4$, where $U(0, 1)$ denotes the Uniform distribution in the interval 0 to 1. Note that when the effects of stationary covariate $(\tilde{x}_i)$ on the responses y_{it} are examined, the median based estimating equations for β should be constructed by taking the correlations of the repeated responses $y_{i1}, \dots, y_{it}, \dots, y_{iT}$ into account, so that the effect of correlation index parameter ρ can be understood for β estimation. Further note that we could also choose non-stationary x_{it} for β estimation. But for simplicity we have not considered such covariates in the present study. This is mainly because of the fact that both non-stationary and stationary covariates based longitudinal EAR(1), EMA(1) and EEQC models produce the same correlation structure. As far as the correlation index parameter ρ is concerned we choose a large positive value, namely $\rho = 0.7$.

In the present study, we generate $y_{i1}, \dots, y_{i4}$ for each $i = 1, \dots, 100$, following the EAR(1) model $y_{it} = \rho_i y_{i,t-1} + I_{it}a_{it}$ (2.2), with $\rho_i = \rho\frac{\lambda_{i,t-1}}{\lambda_{it}}$, which however reduces to ρ (i.e., $\rho_i = \rho$) in the stationary case because λ_{it} and $\lambda_{i,t-1}$ are the same in such cases. For the estimation of β, we use the median regression based GQL estimating equation (2.10) and denote the estimate as GQL(TC) estimate, where TC stands for the true model based correlation structure. We then consider three non-parametric correlation structures based estimating equations, namely by using independence assumption (IND) from Section 2.1.1(a); Jung's QL approach (JQL) from Section 2.1.1(b); and by using lag-correlation (LC) based GQL (GQL(LC)) from Section 2.1.1(c). The simulated means (SM), simulated standard errors (SSE), simulated mean square errors (SMSE), and percentage efficiency (E_1) among median regression based approaches for the estimates of β are reported in the Table 1.

It is clear from the results of Table 1 that all three approaches, namely IND, JQL, and GQL(LC) appear to produce unbiased regression estimates similar to that of median based GQL(TC) estimates. However, when the standard errors of these three approaches are compared to the median based GQL(TC) approach, Jung's QL (JQL) approach appears to be less efficient as compared to the IND and GQL(LC) approaches. Between the last two approaches, that is, IND and GQL(LC), IND appears to be slightly more efficient. Thus, for the AR(1) based exponential data, median regression based IND approach appears to be the best in producing efficient regression estimates and this approach is simpler as compared to the other approaches.

Table 1 *Comparison of mean regression and median regression approaches for the estimation of regression parameter* (β = 0.5, 0.7, 1.0) *involved in an* EAR(1) *model with a large correlation value* $\rho = 0.7$; *based on* 500 *simulations*

				Statistic				
ρ	β	Regression	Estimation approach	SM	SSE	SMSE	E_1	E_2
0.7	0.5	Median	GQL(TC)	0.504	0.114	0.013	100	51
			IND	0.505	0.115	0.013	99	50
			JQL	0.488	0.178	0.031	41	21
			GQL(LC)	0.500	0.119	0.014	94	48
		Mean	GQL(GAC)	0.505	0.081	0.006	–	100
	0.7	Median	GQL(TC)	0.704	0.114	0.013	100	51
			IND	0.705	0.115	0.013	99	50
			JQL	0.689	0.171	0.029	45	23
			GQL(LC)	0.705	0.124	0.015	86	44
		Mean	GQL(GAC)	0.705	0.081	0.006	–	100
	1.0	Median	GQL(TC)	1.004	0.114	0.013	100	51
			IND	1.005	0.115	0.013	99	50
			JQL	0.984	0.212	0.045	29	15
			GQL(LC)	1.001	0.124	0.015	85	43
		Mean	GQL(GAC)	1.005	0.081	0.006	–	100

3.1 Mean regression based GQL estimation

Note that the aforementioned comparison is made among various correlation structure based median regression approaches. However, for the sake of completion, one may also be interested to examine the performance of a general auto-correlation (GAC) structure based mean regression GQL approach when it is known that the repeated data follow the EAR(1) model. To develop the GQL (Sutradhar (2010)) estimating equation, we first provide the computational formula for the mean, variance and correlations of the responses for the EAR(1) model (2.1)–(2.2). To be specific, following Hasan et al. (2007), one may write

$$\begin{aligned}\mu_{it}(\beta) &= E(Y_{it}) = E_{y_{i,t-1}}E\big[\{\rho_i y_{i,t-1} + I_{it}a_{it}\}|y_{i,t-1}\big]\\ &= E\big[\rho_i y_{i,t-1} + E(I_{it})E(a_{it})\big]\\ &= \rho_i\left\{\frac{1}{\lambda_{i,t-1}}\right\} + (1-\rho)\left\{\frac{1}{\lambda_{it}}\right\} = \frac{1}{\lambda_{it}},\end{aligned} \tag{3.1}$$

$$\begin{aligned}\sigma_{itt}(\beta) &= \mathrm{Var}(Y_{it})\\ &= E_{y_{i,t-1}}V\big[\{\rho_i y_{i,t-1} + I_{it}a_{it}\}|y_{i,t-1}\big]\\ &\quad + V_{y_{i,t-1}}E\big[\{\rho_i y_{i,t-1} + I_{it}a_{it}\}|y_{i,t-1}\big]\\ &= \frac{1}{\lambda_{it}^2},\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}E(Y_{it}Y_{i,t-\ell}) &= E_{y_{i,t-\ell}}E_{y_{i,t-\ell+1}}\cdots E_{y_{i,t-1}}E[Y_{it}Y_{i,t-\ell}|y_{i,t-1},\ldots,y_{i,t-\ell}]\\ &= \rho^{\ell}\sqrt{\frac{1}{\lambda_{it}^2}\frac{1}{\lambda_{i,t-\ell}^2}}+\left\{\frac{1}{\lambda_{it}}\right\}\left\{\frac{1}{\lambda_{it-\ell}}\right\}\\ &= \left\{\frac{1}{\lambda_{it}}\right\}\left\{\frac{1}{\lambda_{it-\ell}}\right\}[\rho^{\ell}+1],\end{aligned}$$

yielding

$$c_{iut}(\rho)=\text{Corr}(Y_{iu},Y_{it})=\rho^{|t-u|}. \tag{3.3}$$

Now by writing

$$\begin{aligned}\mu_i(\beta) &= [\mu_{i1}(\beta),\ldots,\mu_{it}(\beta),\ldots,\mu_{iT}(\beta)]' \quad \text{and}\\ \Sigma_i(\beta,\rho) &= A_i^{1/2}C_i(\rho)A_i^{1/2},\end{aligned} \tag{3.4}$$

where $A_i = \text{diag}[\sigma_{i11}(\beta),\ldots,\sigma_{itt}(\beta),\ldots,\sigma_{iTT}(\beta)]$ and $C_i(\rho)=(c_{iut}(\rho))$, one derives the mean regression based GQL estimating equation for β, as

$$\sum_{i=1}^{K}\frac{\partial\mu_i'}{\partial\beta}\Sigma_i^{-1}(\beta,\rho)(y_i-\mu_i)=0. \tag{3.5}$$

Note that for known ρ, this equation (3.5) may be solved iteratively using

$$\hat{\beta}(r+1)=\hat{\beta}(r)+\left[\sum_{i=1}^{K}\frac{\partial\mu_i'}{\partial\beta}\Sigma_i^{-1}(\beta,\rho)\frac{\partial\mu_i}{\partial\beta'}\right]_r^{-1}\left[\sum_{i=1}^{K}\frac{\partial\mu_i'}{\partial\beta}\Sigma_i^{-1}(y_i-\mu_i)\right]_r, \tag{3.6}$$

where $[\,]_r$ is computed by evaluating the quantity within the square brackets $[\,]$ using $\beta=\hat{\beta}(r)$. Next, because ρ is unknown in practice, it must be estimated. By using lag 1 sample correlations one may estimate ρ involved in $c_{iut}(\rho)=\rho^{|t-u|}$ by solving the moment equation given by

$$\hat{\rho}=\frac{\sum_{i=1}^{K}\sum_{t=1}^{T-1}\tilde{y}_{it}\tilde{y}_{i,t+1}/(K(T-1))}{\sum_{i=1}^{K}\sum_{t=1}^{T}\tilde{y}_{it}^2/(KT)}, \tag{3.7}$$

where $\tilde{y}_{it}$ is the standardized residual, defined as, $\tilde{y}_{it}=\frac{(y_{it}-\mu_{it})}{\{\sigma_{itt}\}^{1/2}}$.

3.1.1 *Performance of the linear regression based GQL approach.* For the same designs used above under median regression based approaches, we obtain estimate of β by solving the mean regression based GQL estimating equation (3.5). The SM, SSE and SMSE of this estimates are reported in Table 1 along with similar results under the median regression based approach. The overall percentage efficiency (E2) as compared to this mean regression based approach for the estimates of β are reported in the last column of Table 1.

The results of Table 1 show that the median regression based estimating equation produces less efficient (in the sense of MSE) estimates as compared to the mean regression based estimating equation. This is because the median based approaches when compared to the mean based approach produce estimates with $E_2 < 100$, where the efficiency E_2 for a selected method (M), is defined as $E_2(M) = \{\text{SMSE(Mean Based)}\}/\{\text{SMSE}(M)\} \times 100$. This result is not surprising because of the degree of asymmetry in the present exponential data which not so strong. However, to understand the effect of large asymmetry in the data, we have also generated asymmetric exponential data as in Table 1, but forced a small percentage (1%) of observations to be mean shifted outliers, such that for these observations $\tilde{x}_i$ was first generated from $U(0,1)$ and then for 1% of them $(\tilde{x}_i)$ was shifted to $\tilde{x}_i + 1.5$. The mean and median regression based GQL estimates for these outliers oriented data are shown in Table 2.

These results show that the mean regression based GQL estimates are now biased when compared to the corresponding estimates obtained in the outliers free case as in Table 1, whereas the median regression based new estimates do not appear to be affected by outliers. This prompted us to compare the relative bias (RB) as opposed to MSE, for the mean and median regression based estimates, where RB of an estimate, say $\hat{\beta}$, is defined as

$$\text{RB}(\hat{\beta}) = \frac{\hat{\beta} - \beta}{\text{s.e.}(\hat{\beta})} \times 100 \equiv \frac{SM - \beta}{SSE} \times 100.$$

Table 2 *Comparison of mean regression and median regression approaches for the estimation of regression parameter* (β = 0.5, 0.7, 1.0) *involved in an* $EAR(1)$ *model with a large correlation value* $\rho = 0.7$, *in the presence of* 1% *outliers through shifted covariate values*; *based on* 500 *simulations*

				Statistic			
ρ	β	Regression	Estimation approach	SM	SSE	SMSE	RB
0.7	0.5	Median	GQL(TC)	0.517	0.115	0.013	
			IND	0.518	0.116	0.014	
			JQL	0.508	0.142	0.020	5.630
			GQL(LC)	0.510	0.125	0.015	
		Mean	GQL(GAC)	0.523	0.081		28.770
	0.7	Median	GQL(TC)	0.720	0.116	0.013	
			IND	0.721	0.117	0.014	
			JQL	0.701	0.152	0.023	0.515
			GQL(LC)	0.714	0.130	0.017	
		Mean	GQL(GAC)	0.734	0.080		41.451
	1.0	Median	GQL(TC)	1.032	0.116	0.015	
			IND	1.0307	0.1171	0.014	
			JQL	1.015	0.152	0.023	3.845
			GQL(LC)	1.031	0.118	0.015	
		Mean	GQL(GAC)	1.056	0.084		66.853

It is clear from the last column of Table 2 that mean regression based GQL estimates have much larger relative bias, for example 66.85% when $\beta = 1.0$ and $\rho = 0.7$, as compared to 3.85% relative bias for the median based regression estimates. Thus, if the degree of asymmetry is high which is caused here due to added outliers, the median regression based approach appears to work better than the mean regression based approach.

4 Data analysis: Labor pain data

The labor pain data reported by Davis (1991) consists of repeated measurements of self-reported amount of pain on $K = 83$ women in labor, of which 43 were randomly assigned to a pain medication (treatment) group and 40 to a placebo group. At 30-minute intervals, the amount of pain was marked on a 100 mm line, where $0 =$ no pain and $100 =$ extreme pain. The maximum number of measurements for each woman was 6, but there are some missing values at later measurement times. The observed data under treatment and placebo groups are displayed in Figures 1 and 2, respectively.

It appears from Figure 1 that under the treatment group, the labor pain at any given time ($t = 1, \ldots, 6$) have an exponential form, whereas the marginal observed distributions at different times under the placebo group do not tend to follow the same distribution.

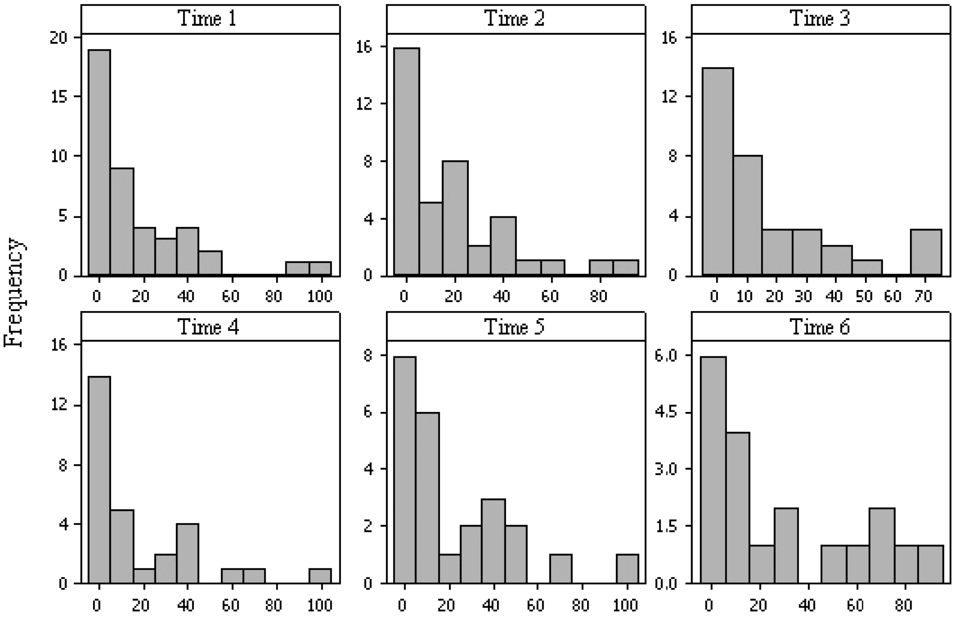

Figure 1 *Labor pain observed data for treatment group.*

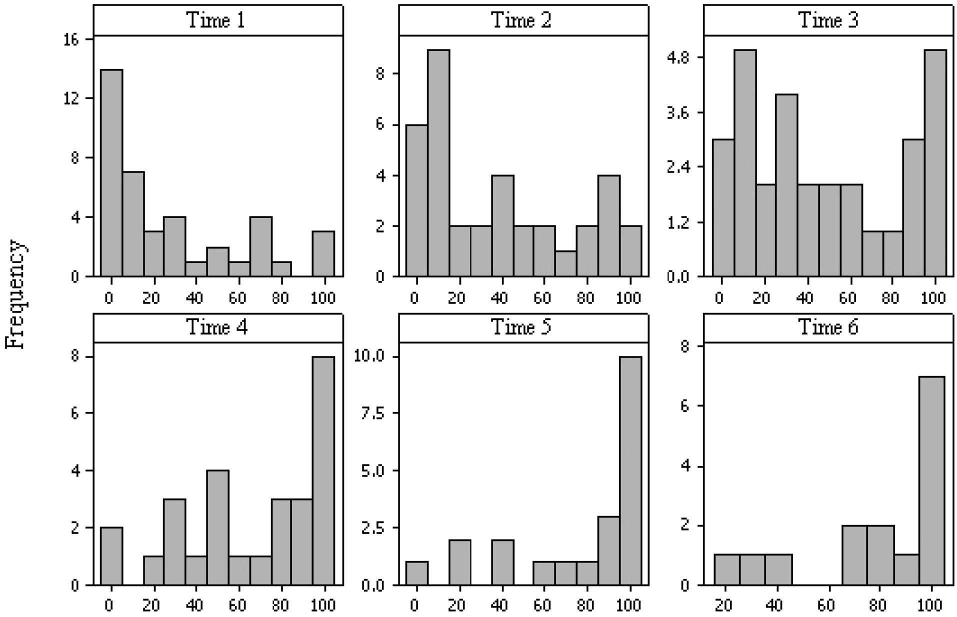

Figure 2 *Labor pain observed data for placebo group.*

Note that to understand the effect of times on the labor pain, Jung (1996, Section 6), for example, fitted a linear median regression model with errors having zero median. To be specific, Jung (1996) has fitted a model $y_{it} = \beta_0 + \beta_2 t + \epsilon_{it}$ for the treatment group, and obtained $\hat{\beta}_0 = 4.36$ and $\hat{\beta}_2 = 1.37$ by using pairwise correlation estimates based QL approach (JQL). In order to see how these estimates or model fit the observed data in Figure 1, we have generated $\hat{\epsilon}_{it}$ from uniform distribution $U(-\frac{1}{2}, \frac{1}{2})$ [to keep the distribution at median to be uniform as suggested by Jung (1996)] and estimated y_{it} as $\hat{y}_{it} = \hat{\beta}_0 + \hat{\beta}_2 t + \hat{\epsilon}_{it}$. The fitted data for this treatment group are displayed in Figure 3.

It is however clear that the histogram in Figure 3 do not exhibit the exponential form exhibited by Figure 1. Thus, even though JQL approach fits the median well, one may be compared for overall fitting. Note that if the inference procedure fits the original distribution well, one may estimate other quantiles as well if needed.

For the aforementioned reason, we have re-analysed the data set using correlated exponential model given in Section 2. Note that when we have compared the IND, JQL and GQL(LC) approaches to the true model based GQL(TC) through a simulation study in Section 3, it was found that IND followed by GQL(LC) produce more efficient regression estimates. As shown in Table 1, among all approaches, Jung's (1996) QL approach was the worst as it produces more bias estimates along with large standard errors. For this reason, we have fitted IND, GQL(LC) and GQL(TC) approaches to this data set. Because our main concern is to see the effect of times in treatment group, we have fitted the exponential model,

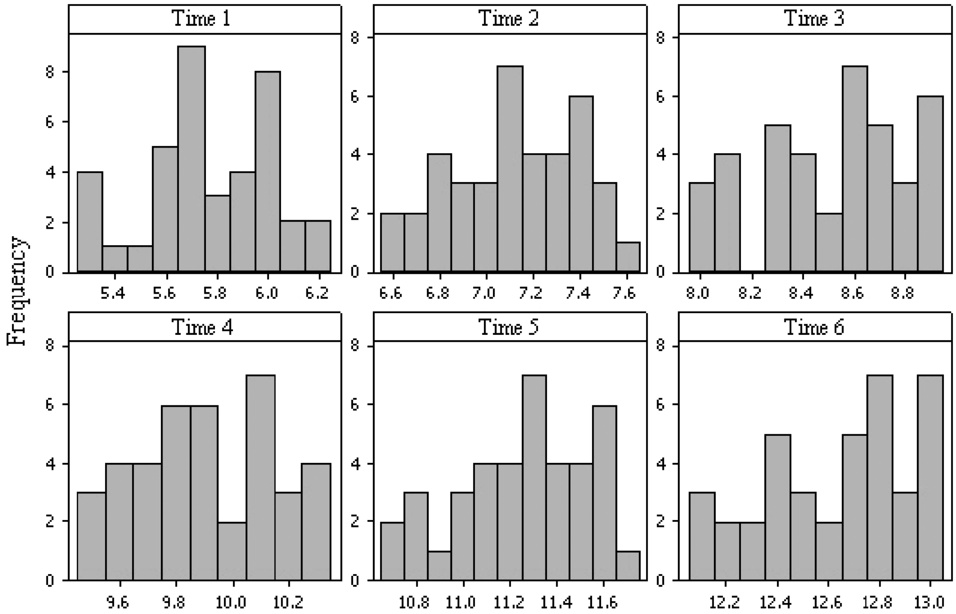

Figure 3 *Linear median regression based fitted labor pain data for the treatment group with* $U(-\frac{1}{2}, \frac{1}{2})$ *error.*

$y_{it} = \rho_i y_{i,t-1} + I_{it}a_{it}$ following (2.2) with median $m_{it} = (\log 2)\exp(\beta_0 + \beta_1 t)$. The parameter estimates along with their estimated standard errors (shown in parenthesis) under these three approaches were found to be

	β_0	β_1	ρ
GQL(TC)	2.161 (0.136)	0.138 (0.034)	0.746
IND	2.208 (0.114)	0.109 (0.032)	–
GQL(LC)	1.799 (0.118)	0.179 (0.019)	0.785

and as displayed in Figure 4, the fitted medians by these three approaches appear to agree well with the medians of the observed data (OBS).

These approaches also appear to fit the over all data distributions well. For example, using above mentioned $\hat{\beta}_0$, $\hat{\beta}_1$ in $m_{it} = (\log 2)\exp(\beta_0 + \beta_1 t)$ and $\hat{\rho}$ under both GQL(TC) and GQL(LC) approaches, when y_{it} were generated following the exponential distribution with median m_{it}, they produce the distributions as in Figures 5 and 6, respectively. These distributions appear to agree well with seemingly exponential distribution for the observed data displayed in Figure 1.

We also have estimated the parameters of the exponential model (2.2) using the mean regression based GQL(TC) approach. The parameter estimates along with the estimated standard errors under this approach were found to be $\hat{\beta}_0 =$ 2.549 (0.194), $\hat{\beta}_1 = 0.129$ (0.051) and $\hat{\rho} = 0.741$. These estimates including cor-

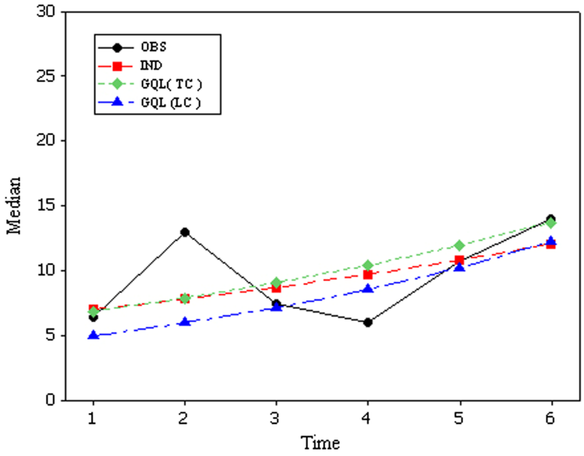

Figure 4 *Observed versus various model based fitted medians for the treatment group.*

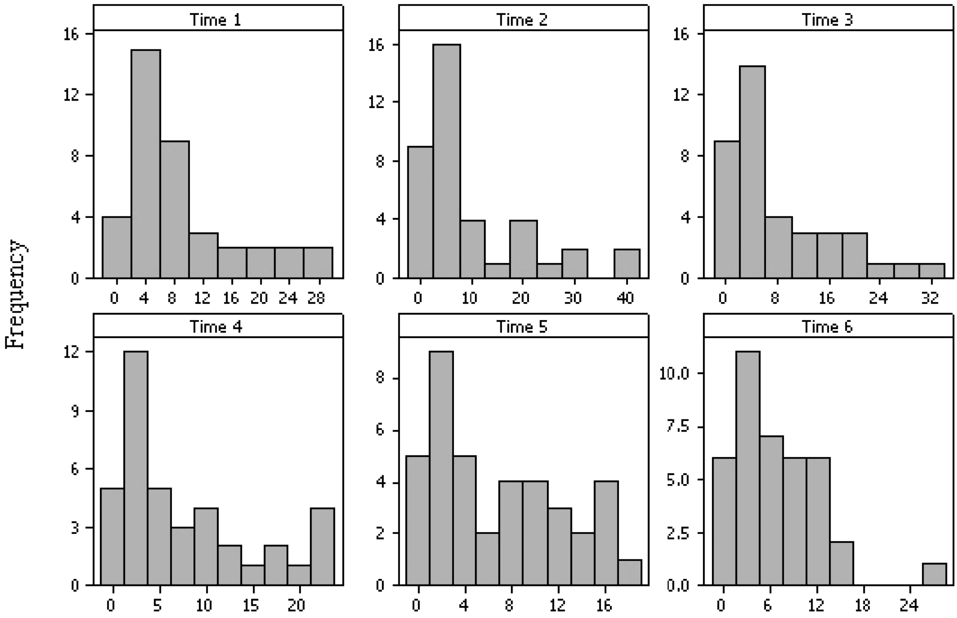

Figure 5 *Exponential median regression based fitted labor pain data for the treatment group under GQL(TC).*

relation estimate appear to be closer to the true model based GQL(TC) estimates but with larger standard errors for the regression estimates. The pattern of these standard errors to be different for this data set when compared to the simulation results reported in Table 1. However, in view of the simulation results reported in

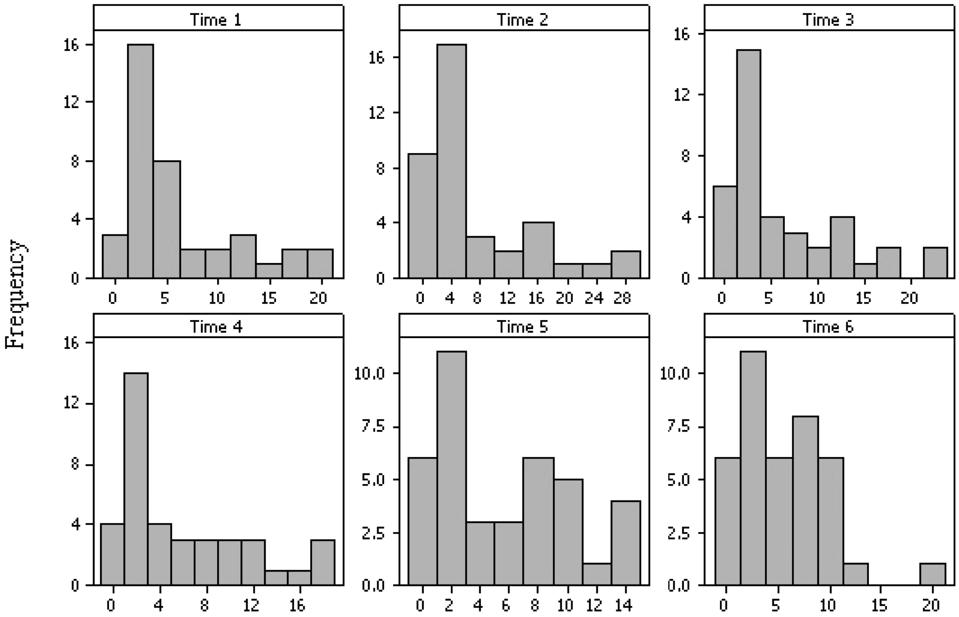

Figure 6 *Exponential median regression based fitted labor pain data for the treatment group under GQL(LC).*

Table 2 and because the observed data are highly asymmetric, the median based estimates are preferable to the mean based estimates.

5 Concluding remarks

In a regression setup for repeated asymmetric data such as exponential data, there exists a pair-wise correlation structure (semi-parametric) based median regression QL approach (Jung (1996)) for the estimation of the regression effects. In this paper, by using an AR(1) type correlation model for repeated exponential data, we have examined the relative performance of various median based GQL approaches. When median regression based approaches were compared among themselves (as opposed to mean regression based approach), it was found that independence assumption based QL approach performs better than the other competitive median based GQL approaches. The empirical results of this paper show that the mean regression based GQL approach may perform the same or better as compared to the median regression based GQL estimates, when the degree of asymmetry in the data is small. When asymmetry in the data will increase such as by introducing outliers, the median regression approaches would perform better than the mean regression based approach, as expected.

Acknowledgments

The authors thank a referee for helpful comments and suggestions.

References

Bassett, G. W. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* **73**, 618–622. MR0514166

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control*. London: Holden-Day. MR0436499

Davis, C. S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine* **10**, 1959–1980.

Fu, L. and Wang, Y. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis* **56**, 2526–2538. MR2910067

Galvao Jr., A. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics* **164**, 142–157. MR2821799

Gaver, D. P. and Lewis, P. A. W. (1980). First order auto regressive gamma sequences and point processes. *Advances in Applied Probability* **12**, 727–745. MR0578846

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics* **8**, 140–154.

Hasan, M. T., Sutradhar, B. C. and Sneddon, G. (2007). On correlation models for longitudinal failure time data. *Sankhyā* **69**, 548–580. MR2460008

Hasan, M. T. (2004). Analysis of longitudinal failure time data. Ph.D. thesis, Memorial Univ. Newfoundland, Canada. MR2707873

Jacobs, P. A. and Lewis, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis* **4**, 19–36. MR0711293

Jung, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association* **91**, 251–257. MR1394079

Karlsson, A. (2007). Nonlinear quantile regression estimation of longitudinal data. *Communications in Statistics—Simulation and Computation* **37**, 114–131. MR2422875

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**, 74–89. MR2083905

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman & Hall. MR0727836

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* **79**, 747–754.

Sim, C. H. (1986). Simulation of Weibull and gamma autoregressive stationary process. *Communications in Statistics—Simulation* **15**, 1141–1146. MR0876784

Sim, C. H. (1990). First-order autoregressive models for gamma and exponential processes. *Journal of Applied Probability* **27**, 325–332. MR1052304

Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science* **18**, 377–393. MR2056579

Sutradhar, B. C. (2010). Generalized Quasi-likelihood (GQL) Inference. StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies.

Sutradhar, B. C. (2011). *Dynamic Mixed Models for Familial Longitudinal Data*. *Springer Series in Statistics*. New York: Springer. MR2777359

V. Nagarajah
B. C. Sutradhar
Department of Mathematics and Statistics
Memorial University
St. John's, NL A1C 5S7
Canada
E-mail: varathan10@gmail.com
bsutradh@mun.ca

V. Jowaheer
Department of Mathematics
University of Mauritius
Reduit
Mauritius
E-mail: vandnaj@uom.ac.mu

A. Biswas
Applied Statistics Unit
Indian Statistical Institute
203, B.T. Road
Kolkata 700 108
India
E-mail: atanu@isical.ac.in